DSC 40A

Week 1 Tues

quantify faithfulness / closeness

Define loss function $\text{loss}(\bar{x}, x_i) = (\bar{x} - x_i)^2$ ← squared loss / error

($\bar{x}$: average; x_i: each individual data point)

Average loss function $R_{sq}(\bar{x}; D) = \frac{1}{n}\sum_{i=1}^{n} \text{loss}(\bar{x}; x_i) = \frac{1}{n}\sum_{i=1}^{n} (\bar{x} - x_i)^2$

↑ empirical risk

Find an estimator h to optimize empirical risk?

$\bar{x} = \arg\min_{h \in R} R_{sq}(h; D)$

$= \arg\min_{h \in R} \frac{1}{n}\sum_{i=1}^{n} (h - x_i)^2$ ← empirical risk minimization

↓ mean!

$h^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \text{Mean}(D)$

arg min : argument x that minimize $f(x)$

$f(h^*) \leq f(h), \forall h \in R$

The best prediction, in terms of mean squared error, is the mean (unique)

2 point example: $\frac{d}{dh}\frac{1}{2}[(h - x_1)^2 + (h - x_2)^2]$

$= \frac{1}{2}\frac{d}{dh}((h - x_1)^2 + (h - x_2)^2)$

$= \frac{1}{2}\left(\frac{d}{dh}(h - x_1)^2 + \frac{d}{dh}(h - x_2)^2\right)$

$= \frac{1}{2}(2(h - x_1) + 2(h - x_2))$

$= \frac{1}{2}(2(h - x_1 + h - x_2))$

$= 2h - x_1 - x_2$

$\Downarrow$

$2h - x_1 - x_2 = 0$

$h = \frac{x_1 + x_2}{2}$

more points example

$\frac{d}{dh}\left(\frac{1}{n}\sum_{i=1}^{n} (h - x_i)^2\right)$

$= \frac{1}{n}\left(\frac{d}{dh}\sum_{i=1}^{n} (h - x_i)^2\right)$

$= \frac{1}{n}\sum_{i=1}^{n} \frac{d}{dh}(h - x_i)^2$

$= \frac{1}{n}\sum_{i=1}^{n} 2(h - x_i)$

$= \frac{2}{n}\sum_{i=1}^{n} (h - x_i)$

$= \frac{2}{n}\left(nh - \sum_{i=1}^{n} x_i\right)$

$\Downarrow$

$\frac{2}{n}\left(nh - \sum_{i=1}^{n} x_i\right) = 0$

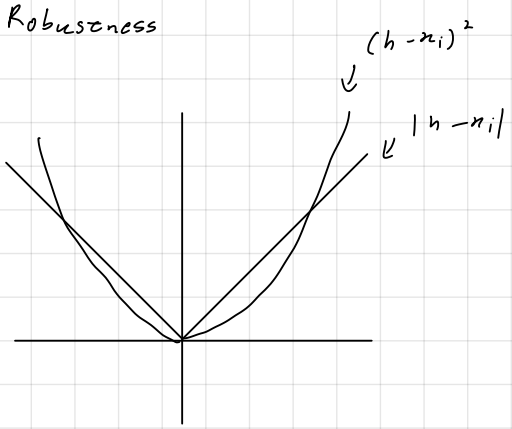
$nh - \sum_{i=1}^{n} x_i = 0$

$\bar{x} = h^* = \frac{1}{n}\sum_{i=1}^{n} x_i$

Outlier

Mean is sensitive to outliers.

Robustness



$R^{ab}(h;D) = \frac{1}{n}\sum_{i=1}^{n} |h - x_i|$ ← mean error (good for robustness)

Week 1 Thurs

For dataset $D = \{x_1, \dots x_n\}$

$loss(h, x_i)$

Empirical risk: $R(h, D) = \frac{1}{n}\sum_{i=1}^{n} loss(h, x)$.

Optimize: $\underset{h \in R}{\text{argmin}}\ R(h, D) = h^*$ ← $D_{training}$

Test: $R(h^*, D_{test})$.

Def. A loss function $L(h, x)$ takes in a prediction h and a true value, x, and outputs a number measuring how far h is from x

loss function:

- Squared loss — mean (D) — $Loss_{sq}(h, x_i) = (h - x_i)^2$
  - Sensitive to outliers.
- Absolute loss: $loss_{abs}(h, x_i) = |h - x_i|$
  - Median (D)

optimize: $\underset{h \in R}{\text{argmin}}\ R_{abs}(h, D) = \frac{1}{n}\sum_{i=1}^{n} loss_{abs}(h, x_i)$

$\lim_{\Delta \to 0} R_{abs}(h+\Delta, D) = R_{abs}(h, D)$

$\frac{\partial}{\partial h} R_{abs}(h, D) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial h} loss(h, x_i)$.

$\mathbb{1}\{h < x_i\} = \begin{cases} 1, & h < x_i \\ 0, & h \geq x_i \end{cases}$

$$\frac{\partial}{\partial h} \text{loss}(h, x_i) = \frac{\partial}{\partial h} |h - x_i| = \begin{cases} 1 & h > x_i \\ \text{NAN} & h = x_i \\ -1 & h < x_i \end{cases} \qquad h \neq x_i$$

$$\simeq -1 \cdot \mathbb{1}\{h < x_i\} + 1 \cdot \mathbb{1}\{h > x_i\}$$

$$\Rightarrow \frac{\partial}{\partial h} R_{abs}(h, D) = \frac{1}{n} \sum_{i=1}^{n} \left(-1 \cdot \mathbb{1}\{h < x_i\} + 1 \cdot \mathbb{1}\{h > x_i\} \right)$$

$$= \frac{1}{n} \left[-1 \sum_{i=1}^{n} \mathbb{1}\{h < x_i\} + (+1) \mathbb{1}\{h > x_i\} \right]$$

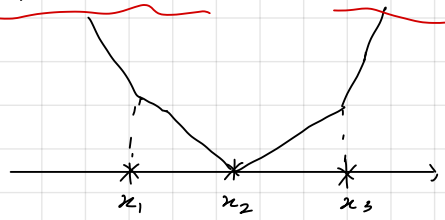
$\sum_{i=1}^{n} \mathbb{1}\{h < x_i\}$ = # of x greater than h.

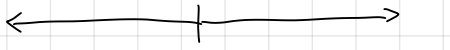
$\sum_{i=1}^{n} \mathbb{1}\{h > x_i\}$ = # of x less than h.

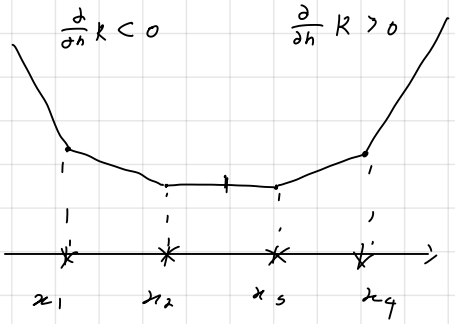
h^* = Median (D), best prediction for mean absolute error.

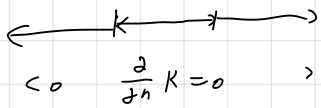
$$= \frac{1}{n} \left[\underbrace{\sum_{x_i < h} 1}_{\#\text{ of } x \text{ less than } h} - \underbrace{\sum_{x_i > h} 1}_{\#\text{ of } x \text{ greater than } h} \right]$$

odd: number is unique ←









< 0 $\quad \frac{\partial}{\partial h} R = 0 \quad > 0$

loss function (contin.)

$\text{loss}(h, x_i) = \mathbb{1}\{h \neq x_i\}$

ERM → Mode (D)

even: any number between the middle two data point minimizes mean absolute error. ←

Ex. $R_{abs}(h, D) = \frac{1}{n} \sum_{i=1}^{n} |h - x_i|$

$R_{sq}(h, D) = \frac{1}{n} \sum_{i=1}^{n} (h - x_i)^2$

$\sqrt{R_{sq}(h, D)}$

If $R_{ab}(h, D) \leq \varepsilon$, $\overset{?}{\Rightarrow} \sqrt{R_{sq}(h, D)} \leq \varepsilon$

Empirica risk – the average loss on the data sets:

$$R_L(h) = \frac{1}{n}\sum_{i=1}^{n} L(h, y_i)$$

R = "risk"

The goal of learning: find h that minimizes R_L.

This is called empirical risk minization (ERM).

- The choice of loss function determines the properties of the result.

↓

different loss function

different " minizer

different " prediction

ex). $R_{0,1}(h) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 0 & \text{if } h = y \\ 1 & \text{if } h \neq y \end{cases}$ $\Rightarrow h^* = \text{Mode}(P)$

loss	Minimizer	outliers	Differentiable
L_{abs}	median	insensitive	no
L_{sq}	mean	sensitive	yes
$L_{0,1}$	mode	insensitive	no

$L_{abs}(h, y) = |y - h|$

$L_{sq}(h, y) = (y - h)^2$

Emperical risk $R(h) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}(h, y)$.

$h^* = \arg\min_{h \in \mathbb{R}} R(h; D)$.

$R(h^*; D)$

$R_{sq}(h^*; D) = \frac{1}{n}\sum_{i=1}^{n} (h^* - x_i)^2$

$= \frac{1}{n}\sum_{i=1}^{n} (\bar{x} - x_i)^2$ ← Variance

Standard deviation $\sqrt{R_{sq}(h^*; D)}$

- $R_{abs}(h^*; D) = \frac{1}{n}\sum_{i=1}^{n} |h^* - x_i|$

$= \frac{1}{n}\sum_{i=1}^{n} |\text{Median}(D) - x_i|$

$D = \{1, 2, 2, 4\}$

$R_{ab}(h^*; D) = \frac{1}{2}$

- $R_{0,1}(h; D) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{h \neq x_i\}$

$= \frac{n-1}{n}$

$\sqrt{K_{sq}(h;D)}$ vs. $K_{abs}(h;D)$

$$K_{sq}(h;D) = \frac{1}{n}\sum_{i=1}^{n}(h-x_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\underbrace{|h-x_i|^2}_{|h-x_i|=a_i}$$

$$= \frac{1}{n}\sum_{i=1}^{n} a_i^2$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n} a_i^2\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i^2$$

$$= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{j=1}^{n} a_j^2$$

$$K_{abs}^2(h;D) = \left(\frac{1}{n}\sum_{i=1}^{n}|h-x_i|\right)^2$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} a_i\right)^2$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} a_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} a_j\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j$$

$$\frac{1}{n^2}\sum_i\sum_j \frac{a_i^2+a_j^2}{2}$$

Note: $\frac{a^2+b^2}{2} \geq ab$

Young's Ineq. s.(a)

$$K_{sq}(h;D) - K_{ab}^2(h;D)$$

$$= \frac{1}{n^2}\sum_i\sum_j\left(\frac{a_i^2+a_j^2}{2} - a_i a_j\right)$$

$$= \frac{1}{n^2}\sum_i\sum_j\left(\frac{1}{2}(a_i^2+a_j^2-2a_i a_j)\right)$$

$$= \frac{1}{n^2}\sum_i\sum_j\left(\frac{1}{2}(a_i-a_j)^2\right) \geq 0$$

$$\Downarrow$$

$$K_{sq}(h;D) \geq K_{ab}^2(h;D).$$

$$\Downarrow$$

$$\sqrt{K_{sq}(h;D)} \geq K_{ab}(h;D), \quad \forall h.$$

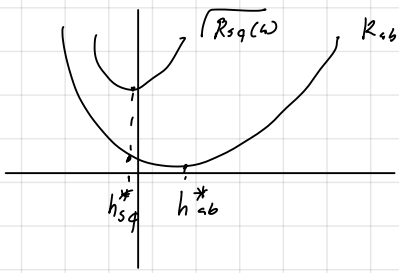
want:

$\sqrt{K_{sq}(h^*;D)}$ vs. $K_{ab}(h^*;D)$.

$\min\sqrt{K_{sq}(h;D)} \geq \min K_{ab}(h;D)$

- $K_{abs}(h^*_{ab}, D) \leq K_{ab}(h^*_{sq}, D) \leq \sqrt{K_{sq}(h^*_{sq}, D)}$

$\Rightarrow$



Center and Spread

- The input h^* that minimizes $R(h)$ is some measure of the center of the data set.
  e.g. median, mean, mode.

- The minimum output $R(h^*)$ represent some measure of the spread, or variance, in the data set.

$\Downarrow$

For absolute loss: $R_{abs}(h^*) = R_{abs}(\text{Median}(y_1, y_2, \cdots, y_n))$

$= \frac{1}{n}\sum_{i=1}^{n} |y_i - \text{median}(y_1, y_2, \cdots, y_n)|$.

this is the mean absolute deviation from the median.

For squared loss: $R_{sq}(h^*) = R_{sq}(\text{Mean}(y_1 \cdots y_n))$

$= \frac{1}{n}\sum_{i=1}^{n} (y_i - \text{Mean}(y_1 \cdots y_n))^2$

this is called the variance

the square root of variance is standard deviation.

Huber loss / Gradient Descent

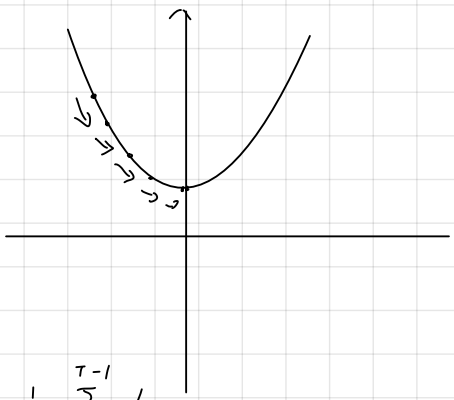
$$\text{loss}(h, x_i) = \begin{cases} |h - x_i| - \frac{1}{4}, & |h - x_i| > \frac{1}{2} \\ (h - x_i)^2, & |h - x_i| \le \frac{1}{2} \end{cases}$$

$$R_{Huber} = \frac{1}{n}\sum_{i=1}^{n} \text{loss}(h, x_i)$$

Initialize $h_0 = 0$

$$h_{t+1} = h_t - \Delta_t \cdot \frac{d}{dh} R_{Huber}(h, 0)$$

$$\frac{dh(t)}{dt} = \frac{\partial}{\partial h} R_{Huber}(h_t, 0)$$

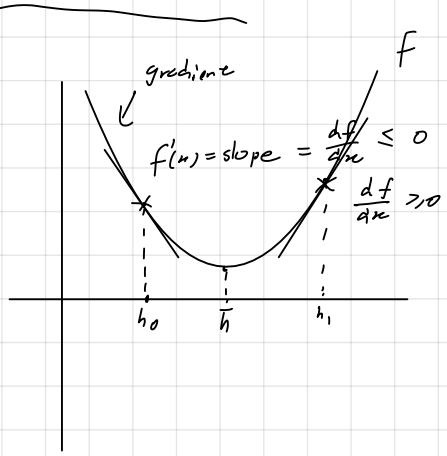


Output $h_t = \frac{1}{t}\sum_{t=1}^{T-1} h_t$

Gradient Descent

From Mach 20 c:

gradient -> direction increase the most

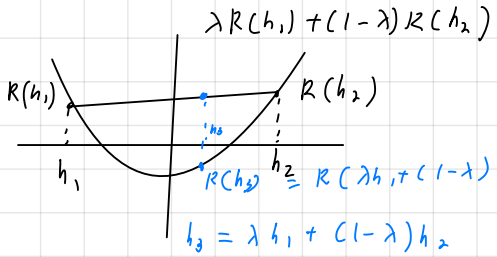


- Pick α to be a positive number, the learning rate.
- Pick a starting prediction, h_0.
- On step i, perform update $h_i = h_{i-1} - \alpha \cdot \frac{dR}{dh}(h_{i-1})$
- Repeat until convergence.

$$x_{t+1} = x_t - \Delta \frac{df(x_t)}{dx}$$

$t = 0, 1, 2, \ldots$

Convex function



$h_3 = \lambda h_1 + (1-\lambda) h_2 \quad \lambda \in [0,1]$

$$R(h_3) = R(\lambda h_1 + (1-\lambda) h_2)$$

$$R(\lambda h_1 + (1-\lambda) h_2) \leq \lambda R(h_1) + (1-\lambda) R(h_2)$$

$$\forall \lambda \in [0,1], \forall h_1, h_2.$$

(5) b. loss convex $\Rightarrow R(h, D)$ convex.

Thm: For Huber loss, with suitable $\{\Delta_0, \ldots, \Delta_t\}$ (step size)

$$R(\bar{h}_T, D) - R(h^*, D) \leq \frac{4(h_0 - h^*)^2}{T}$$

this means hubber loss will converge

$\bar{h} = \frac{1}{T} \sum_{t=0}^{T-1} h_t$, all h_t from gradient descent algorithm

T: number of iteration that G.D. converges.

If you want error ε: (range of data point)

Need $T \geq \frac{4(h_0 - h^*)^2}{\varepsilon} = \frac{4(h^*)^2}{\varepsilon} \leq \frac{4K^2}{\varepsilon}$

$\Delta t = \frac{1}{4}$. $\forall t$. for Hubber

proof: $\frac{\partial^2}{\partial h}\, loss(h, h_i) = \begin{cases} 0, & |h - h_i| > \frac{1}{2} \\ 2, & |h - h_i| \leq \frac{1}{2} \end{cases} \geq 0$

Hubber loss only has minimum, no maximum

$0 \leq \frac{\partial^2}{\partial h^2} loss(h, h_i) \leq 2$

↑ convex characteristic ↑ smooth function

$\Rightarrow 0 \leq \frac{\Delta^2 K(h, 0)}{\partial h^2} \leq 2$ (risk)

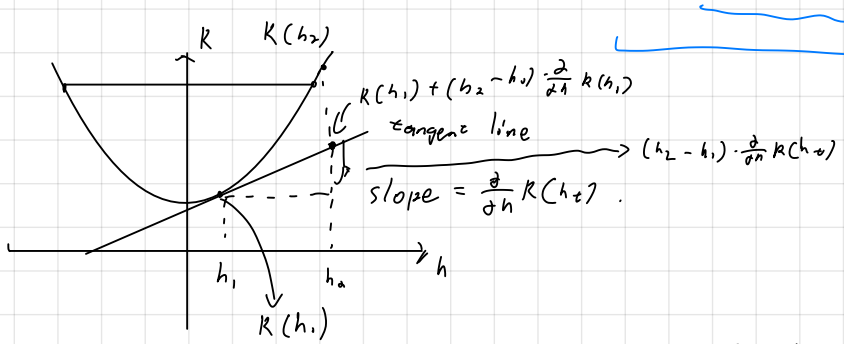
$\Rightarrow \left(\frac{\partial}{\partial h} K(h)\right)^2 \leq 4\left(K(h) - K(h^*)\right)$ for Hw2.

$(h_{t+1} - h^*)^2 = (h_t - \Delta t \cdot \frac{\partial}{\partial h} K(h_t) - h^*)^2$

$= \left[(h_t - h^*) - \Delta t \cdot \frac{\partial}{\partial h} K(h_t)\right]^2$

$= (h_t - h^*)^2 - 2\Delta t (h_t - h^*) \cdot \frac{\partial}{\partial h} K(h_t) + \Delta t^2 \left(\frac{\partial}{\partial h} K(h_t)\right)^2$

$\geq K(h_t) - K(h^*)$



So, $K(h_2) \geq K(h_1) + (h_2 - h_1)\frac{\partial}{\partial h} K(h_1)$

$\Rightarrow (h_2 - h_1) \cdot \frac{\partial}{\partial h} K(h_1) \geq K(h_1) - K(h_2)$

$(h_t - h^*) \cdot \frac{\partial}{\partial h} K(h^*) \geq K(h_t) - K(h^*)$

$\leq (h_t - h^*)^2 - 2\Delta t \left(K(h_t) - K(h^*)\right) + \Delta t^2 \left(K(h_t) - K(h^*)\right)$

$=$

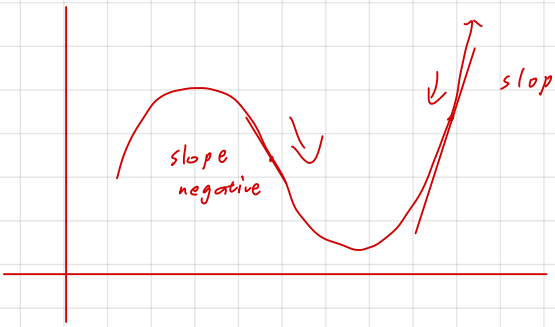
$$\leq (h_t - h^*)^2 - 2\Delta t(1-2\delta t)\,(R(h_t) - R(h^*)).$$

maximize

$$\Downarrow$$

$$\delta t = \frac{1}{4}$$

Gradient Descent



slope is positive

initial guess.

$$h_1 = h_0 - \frac{dR}{dh} / \nabla R(h_0)$$

opposite the direction of the derivative / gradient.

Algorithm:

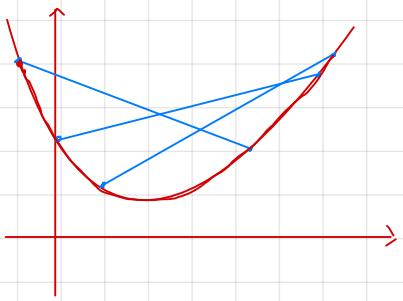
- Pick α to be positive number, the learning rate, also known as step size.
- Pick a starting prediction, h_0.
- On step i, perform update $h_i = h_{i-1} - \alpha \cdot \frac{dR}{dh}(h_{i-1})$
- Repeat until convergence.

Convex function

- f is convex if for every a, b in the domain of f, the line segment between

  $(a, f(a))$ & $(b, f(b))$.

  does not go below the plot of f.



function $f: R \to R$ is convex if $\forall a, b \in R$ and $t \in [0, 1]$,

$$(1-t)f(a) + tf(b) \geq f((1-t)a + tb).$$

Thm: If $R(h)$ is convex and differentiable, then gradient descent converges to a global minimum of R provided that the step size is small enough.

↑

because f convex and has local min $\Rightarrow$ local min is global min.

Test for convexity.

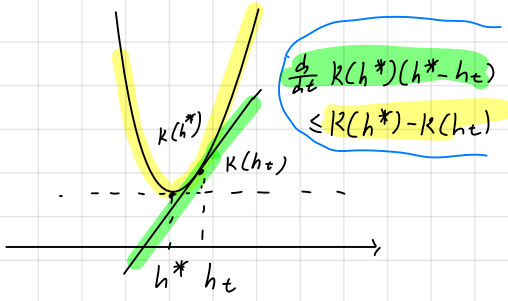
f is convex $\Leftrightarrow \frac{d^2 f}{dx^2}(x) \geq 0. \quad \forall x.$

- Sum of convex functions is convex.
  - ↳ if $L(h, y)$ is convex, then $R(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, y_i)$ is convex.

- $L_{sq}(h, y) = (y - h)^2$ is convex
- $L_{abs}(h, y) = |y - h|$ is convex

Gradient Descent

$$h_{t+1} = h_t - \alpha \frac{d}{dh} R(h_t).$$



$$0 \leq \frac{d^2}{dh^2} R(h) \leq L$$

$L = 2$ for Hubber Risk

L: constant

$0 \leq$: convex

$\leq L$: smooth

$$(h_{t+1} - h^*)^2 = (h_t - \Delta t \frac{d}{dh} R(h_t) - h^*)^2$$

$$= (h_t - h^*)^2 - 2\Delta t \frac{d}{dh} R(h_t)(h_t - h^*) + \Delta t^2 \left(\frac{d}{dh} R(h_t)\right)^2.$$

$\frac{d}{dh} R(h_t)(h_t - h^*) \geq R(h^*) - R(h_t)$

$\left(\frac{d}{dh} R(h_t)\right)^2 \leq 2L(R(h_t) - R(h^*))$. (HW 2 Q5)

$$\leq (h_t - h^*)^2 - 2\Delta t(1 - L \cdot \Delta t)(R(h_t) - R(h^*)).$$

$\Delta t = \frac{1}{2L}$ for optimize

for maximum $\Rightarrow$ less iteration for gradient descent.

$\frac{1}{2L}$

So, let $\Delta t = \frac{1}{2L}$,

$$\Rightarrow (h_{t+1} - h^*)^2 \leq (h_t - h^*)^2 - \frac{1}{2L}(R(h_t) - R(h^*))$$

this gets 0 when $R(h_t) = R(h^*)$, in which case, converges.

$$\Rightarrow \frac{1}{2L}(R(h_t) - R(h^*)) \leq (h_t - h^*)^2 - (h_{t+1} - h^*)^2.$$

Sum:

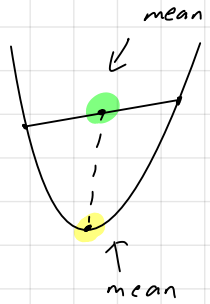
$$\frac{1}{2L}(R(h_0) - R(h^*)) \leq (h_0 - h^*)^2 - (h_1 - h^*)^2.$$
$$\frac{1}{2L}(R(h_1) - R(h^*)) \leq (h_1 - h^*)^2 - (h_2 - h^*)^2.$$
$$\frac{1}{2L}(R(h_{t-1}) - R(h^*)) \leq (h_{t-1} - h^*)^2 - (h_t - h^*)^2.$$

telescoping sum (canceling out terms).

$$\Rightarrow \frac{1}{2L} \sum_{t=1}^{T-1} (R(h_t) - R(h^*)) \leq (h_0 - h^*)^2 - (h_t - h^*)^2 \leq (h_0 - h^*)^2$$

$$\Rightarrow \frac{1}{2L} \cdot \frac{1}{T} \sum_{t=0}^{T-1} (R(h_t) - R(h^*)) \leq \frac{1}{T}(h_0 - h^*)^2$$

let: $\bar{h}_t = \frac{1}{T}\sum_{t=0}^{T-1} h_t$, then $K(\bar{h}_t) - K(h^*) \le \frac{1}{T}\sum_{t=0}^{T-1} K(h_t) - K(h^*)$.



b/c K convex

smooth constant

$$K(\bar{h}_t) - K(h^*) \le \frac{2L}{T}(h_0 - h^*)^2$$

If we want

$K(\bar{h}) - K(h^*) \le \varepsilon$, $\forall \varepsilon > 0$.

take $T = \frac{4}{\varepsilon}(h_0 - h^*)^2$.

Bound $|\bar{x}(D_S^{bs}) - \bar{x}(D_\infty^{bs})|$

$$(\bar{x}(D_n) - \bar{x}(D_\infty)^2) = O(1/n)$$

how large constant

$$(\bar{x}(D_1) - \bar{x}(D_\infty))^2$$

Use Var to estimate: $\le \frac{C}{n}\sum_{i=1}^{n}(x_i - \bar{x}(D_\infty))^2 \le \frac{C}{n}\sum_{i=1}^{n}(x_i - \bar{x}(D_n))^2$

true mean — empirical mean

$$\sum_{i=1}^{n}(x_i - \bar{x}(D_\infty))^2 \approx \sum_{i=1}^{n}(x_i - \bar{x}(D_n))^2 = C \cdot Var(D_n)$$

$$= (x_i - \bar{x}(D_\infty) + \bar{x}(D_\infty) - \bar{x}(D_n))^2$$

$$= (x_i - \bar{x}(D_\infty))^2 + (\bar{x}(D_\infty) - \bar{x}(D_n))^2 + 2(\cdots)(\cdots)$$

$$\frac{1}{n}\cdot(x_i - \bar{x}(D_\infty))^2$$

Put together: $(\bar{x}(D_n) - \bar{x}(D_\infty))^2 \le \frac{C}{n}\cdot Var(D_n)$

$$|\bar{x}(D_n) - \bar{x}(D_\infty)| \le \frac{C}{\sqrt{n}}\cdot STD(D_n)$$

$$|\bar{x}(D_n) - \bar{x}(D_\infty)| \le \frac{\log(1/\delta)}{\sqrt{n}} \cdot STD(D_n)$$

confidence level

the larger the confidence, the larger range of error from true mean

Sequence Decision Making (Reinforcement Learning)

- Bandits Problem.

↓

Explore then commit

- spend the first N rounds uniformly on the two options.
- Then commit to the one with higher average reward: $\operatorname{argmin}_{a=1,2} \bar{R}^a$

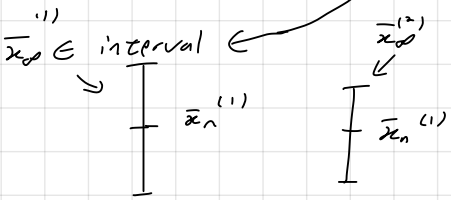
Want: $\Delta\text{reward} = O\left(\frac{STD(D_N)}{\sqrt{n}}\right)$

↑ confidence level constant

$\Leftarrow$ $|\bar{x}_n - \bar{x}_\infty| \le \frac{STD(D_n)}{\sqrt{n}} \cdot \log\left(\frac{1}{\delta}\right)$

↑ $(1-\delta)$ probability

$\bar{x}_\infty^{(1)} \in$ interval



How to compare => when two interval do not overlap

⇓

one is definitely greater than the other.

⇑

by let $\log\left(\frac{1}{\delta}\right) = |\bar{x}_\infty^{(1)} - \bar{x}_\infty^{(2)}|$

$= \Delta\text{reward}$

↑ difference in true mean

↑ How to figure out?

Algorithm:

- keep track N^a
- optimistic estimate: $\hat{R}^a = \bar{R}^a + C \cdot \frac{STD(D_{N^a})}{N^a}$
- choose action $\operatorname{argmin}_{a=1,2} \hat{R}^a$

↑ # of times chosen action a

Linear Regression

- Feature – an attribute, information
  - predictor variable – the feature – x
  - response variable – the quantity – y – we are trying to predict.
  - hypothesis function / prediction rule.

- Use loss function to quantify the quality of prediction.
  - Squared loss : $(\text{actual} - \text{predicted})^2$

↓

Empirical risk : $R_{sq}(H) = \frac{1}{n} \sum_{i=1}^{n} (y_i - H(x_i))^2$

↓

Smallest mean squared error → good prediction

↑

Not overfitting!

↓

Choose H to be certain function!

↓

linear function : $H(x) = w_0 + w_1 x$

(w_0: intercept, w_1: slope)

↓

H^* be the linear function that minimizes $R_{sq}(H) = \frac{1}{n} \sum_{i=1}^{n} (y_i - H(x_i))^2$ (least squares regression).

↓ plug in H (parameter)

$$R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

Find slope and intercept that minimizes.

Solve for gradient

↓

$$\frac{\partial R_{sq}}{\partial w_0} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))$$

$$\frac{\partial R_{sq}}{\partial w_1} = \frac{-2}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i)) x_i$$

↓

Solve for $\frac{\partial RSq}{\partial w_0} = 0$ & $\frac{\partial RSp}{\partial w_1} = 0$

$$w_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - w_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$w_0^* = \bar{y} - w_1^* \bar{x}$$

↑ best intercept

$$w_1^* = \frac{\sum_{i=1}^{n} (y_i - \bar{y}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}$$

, where

↑ best slope

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

the process is called "fitting to the data".

lease square solution (optimal parameter).

$$H^*(w) = w_0^* + w_1^* x$$

- Correlation coefficient
  - a measure of the strength of the linear association of two variables, x & y.
  - coefficient, r, the average of the product of x and y, in standard unit

$$r = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_x}\right)$$

→ mean

→ standard deviation.

↓

- r has no unit
- $r = 1$ → perfect positive linear association
- $r = -1$ → perfect negative linear association
- the closer r is to 0 → the weaker the linear association

↓ optimal slope for the linear prediction rule w_1^* in r

$$w_1^* = r\frac{\sigma_y}{\sigma_x}$$ →

- sign (r) = sign (w_1^*)
- y value get spread out → $\sigma_y \uparrow$ → slope ↑
- x value get spread out → $\sigma_x \uparrow$ → slope ↓

- intercept

$$H^*(\bar{x}) = \bar{y}$$

Regression

ex. Look at salary (y) as function of years of experience (x).

↑ response variable

↑ predictor variable

$D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

$y \approx H(x)$

↑

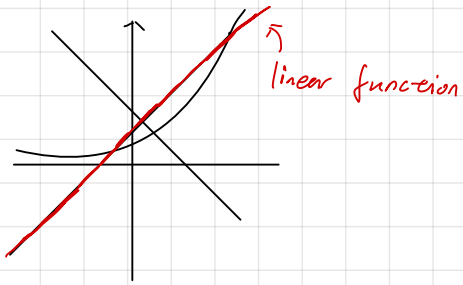
Hypothesis function / prediction rule.

Ex. $y = ax + b$

$y = e^{ax}$

$y = b - x$

... ...



which $H(x)$ is good?

loss function

$\text{loss}(H; (x_i, y_i)) = (H(x_i) - y_i)^2$ squared loss

Empirical risk: $R_{sq}(H, D) = \frac{1}{n} \sum_{i=1}^{n} \text{loss}(H; (x_i, y_i))$

$= \frac{1}{n} \sum_{i=1}^{n} (H(x_i) - y_i)^2$

Optimization: choose prediction rule by $\min_{H \in \mathcal{H}} R_{sq}(H; D)$

Out of all loss function: $H(x) = w_1 x + w_0$. Find H^* that minimizes $R_{sq}(H; D)$.

↓

optimize w_1 & w_0.

squared loss ↓

$$\min_{w_1, w_0 \in \mathbb{R}} R_{sq}(H; D)$$

$$= \min \frac{1}{n} \sum_{i=1}^{n} (H(x_i) - y_i)^2$$

$$= \min \frac{1}{n} \sum_{i=1}^{n} (w_1 x_i + w_0 - y_i)^2$$

optimal prediction

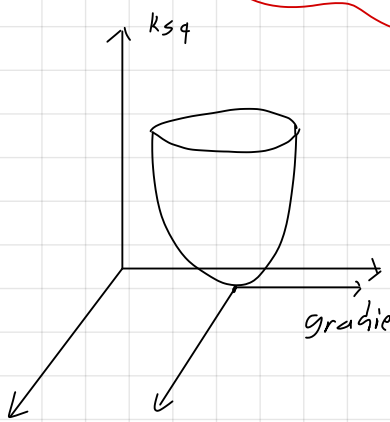
$$w_1^*, w_0^* = \underset{w_1, w_0}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} (w_1 x_i + w_0 - y_i)^2.$$

let $\nabla_{(w_1, w_0)} \left(\frac{1}{n} \sum_{i=1}^{n} (w_1 x_i + w_0 - y_i)^2 \right) = 0$

⇓

$$\nabla_{(w_1,w_0)}\left(f(w_1,w_0)\right) = \begin{pmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_0} \end{pmatrix} = 0$$

solve for critical point



gradient / partial derivative = 0.

$$\Downarrow$$

$$\begin{cases} \frac{2\,Ksq((w_1,w_0),D)}{2w_0} = 0 & ① \\ \frac{2\,Ksq((w_1,w_0),D)}{2w_1} = 0 & ② \end{cases}$$

① $\frac{2}{2w_0}\left(\frac{1}{n}\sum_{i=1}^{n}((w_1x_i+w_0)-y_i)^2\right)=0$

$= \frac{1}{n}\sum_{i=1}^{n}\frac{2}{2w_0}\left[(w_1x_i+w_0-y_i)^2\right]$

$= \frac{1}{n}\sum_{i=1}^{n}2\cdot(w_1x_i+w_0-y_i)$

$= 2\left(\frac{1}{n}\sum_{i=1}^{n}w_0+\frac{1}{n}\sum_{i=1}^{n}w_1x_i-\frac{1}{n}\sum_{i=1}^{n}y_i\right)$

$= 2\left(w_0+w_1\bar{x}-\bar{y}\right).$

$\bar{x}$ = x mean

$\bar{y}$ = y mean

$$\Downarrow$$

$2(w_0+w_1\bar{x}-\bar{y})=0$

$w_1\bar{x}+w_0=\bar{y}$ optimal solution pass through $(\bar{x},\bar{y})$.

$\boxed{w_0=\bar{y}-w_1\bar{x}}$

② $\frac{2}{2w_1}\left(\frac{1}{n}\sum_{i=1}^{n}(w_1x_i+w_0-y_i)^2\right)=0.$

$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2}{2w_1}(w_1x_i+w_0-y_i)^2\right)$

$= \frac{1}{n}\sum_{i=1}^{n}(2\cdot(w_1x_i+w_0-y_i)\cdot x_i)$

$= 2\frac{1}{n}\sum_{i=1}^{n}(w_1x_i^2+w_0x_i-x_iy_i)$

② + ①:

$0 = \frac{1}{n}\sum_{i=1}^{n}(w_1x_i^2+x_i(\bar{y}-w_1\bar{x})-x_iy_i)$

$0 = \frac{1}{n}\sum_{i=1}^{n}(w_1x_i^2+x_i\bar{y}-x_iw_1\bar{x}-x_iy_i).$

$0 = w_1\frac{1}{n}\sum_{i=1}^{n}(x_i^2-x_i\bar{x})+\frac{1}{n}\sum_{i=1}^{n}(x_i\bar{y}-x_iy_i)$

$w_1\frac{1}{n}\sum_{i=1}^{n}(x_i^2-x_i\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_iy_i-x_i\bar{y})$

$$\Downarrow$$

$\frac{1}{n}\sum_{i=1}^{n}x_i(y_i-\bar{y}).$

$$\Downarrow$$

$\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})(y_i-\bar{y})$

since

$\frac{1}{n}\sum_{i=1}^{n}\bar{x}(y_i-\bar{y})=0$

since $\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})=0.$

$w_1\frac{1}{n}\sum_{i=1}^{n}x_i(x_i-\bar{x})$

$$\Downarrow$$

$w_1\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})$

$= w_1\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

since $\frac{1}{n}\sum_{i=1}^{n}\bar{x}(x_i-\bar{x})=0.$

$$\Downarrow$$

sum of product

Thus, $w_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2}$

correlation

Variances in x.

Determine the direction of the regression line.

Positive or Negative

since

$\sum \bar{x}(y_i-\bar{y}) = 0$

$\sum \bar{y}(y_i-\bar{y}) = 0$

$\sum (x_i-\bar{x})(y_i-\bar{y})$

$= \sum x_i(y_i-\bar{y})$ $\quad x_i y_i - x_i\bar{y}$

$= \sum y_i(x_i-\bar{x})$ $\quad x_i y_i - y_i\bar{x}$

$= \sum x_i y_i - \bar{x}\bar{y}$.

Midterm Review

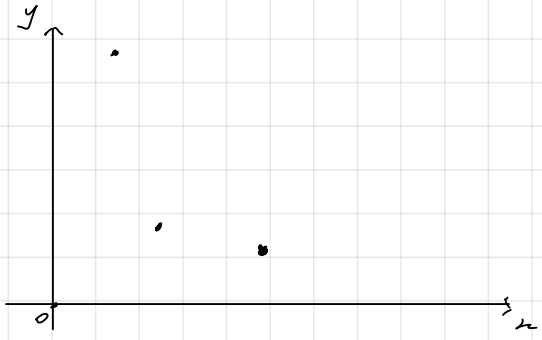
HW1

some question in

$\bar{x} = 5$

$\bar{y} = 4$

ex)	x_i	y_i
(x_1, y_1)	3	7
(x_2, y_2)	4	3
(x_3, y_3)	8	2



$y = H(x) = w_1 x + w_0$

What is the bese prediction rule ?

that describe the data set?

$w_1^* = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$

$= \frac{(-2)(3)+(-1)(-1)+(3)(-2)}{(-2)^2+(-1)^2+(3)^2}$

$= \frac{-6+1-6}{4+1+9}$

$= \frac{-11}{14}$

$w_0^* = \bar{y} - w_1 \cdot \bar{x}$

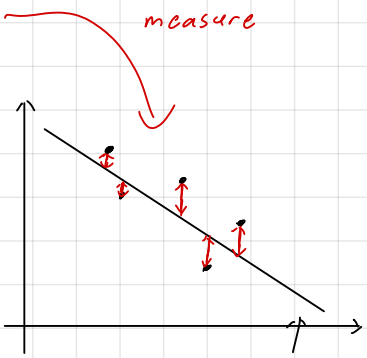
$= 4 - \frac{-11}{14} \cdot 5$

$= \frac{101}{14}$

So, $H(x) = -\frac{11}{14}x + \frac{101}{14}$.

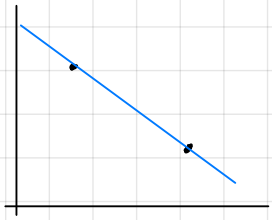
$R_{sq}(H^*; D) = R_{sq}((w_1^* x + w_0^*); D)$. Mean squared error of a data set

$$= \frac{1}{n}\sum_{i=1}^{n}\left((w_1^* x_i + w_0^*) - y_i\right)^2.$$



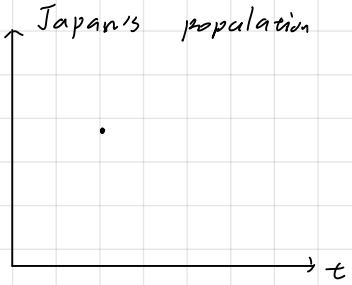
linear prediction rule.

D has two data point



$R_{sq} = 0$.

ex). Japan's population decrease by half by 2060, will vanish by 2100.



Choices of function close of H.

1. Linear
2. Polynomial
3. Exponential.

next year population ↓ X_{t+1} = Current population ↓ X_t + birth rate $\alpha \cdot X_t$

$$X_{t+1} = X_t + \alpha \cdot X_t - \beta X_t$$

$$= (1 + \alpha - \beta) X_t$$

$$= \delta X_t$$

(β: death rate)

2. $H(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_n x^n$.

3. $H(x) = c_0 \cdot e^{c_1 x}$

$$X_t = \delta X_{t-1} = \delta^2 \cdot X_{t-2} = \cdots = \delta^t \cdot X_0$$

(X_0: initial population)

$$X(t) = \delta^t \cdot w$$

Can we learn (δ, w) from data $\{(t_1, X_1) \cdots (t_n, X_n)\}$?

$$\ln(X(t)) = \ln(\delta^t \cdot w)$$

$$= \ln(\delta^t) + \ln(w)$$

$$= t \cdot \ln(\delta) + \ln(w)$$

($\ln(\delta)$: w_1, $\ln(w)$: w_0)

$$\ln(x(t)) = w_1 t + w_0$$

So,

t_i	$\ln(x(t))$
⋮	⋮

do least square solution for w_1 and w_0.

Conclusion:

$$g(y) = w_0 + w_1 f(x)$$

idea: transform non-linear relationship to linear one. Then do lesst square solution.

$$y = \delta^x \cdot w$$

Multi-linear regression

$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

Want to fit:

$$y = H(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \cdots + w_m x^n$$

$$y = H(x^{(1)}, x^{(2)}, x^{(3)}, \cdots, x^{(n)}) \leftarrow \text{multiple feature}$$

↑ Salary ↑ YOE ↑ location ↑ job family, ...

$$= w_0 + w_1 x^{(1)} + \cdots + w_n x^{(n)}$$

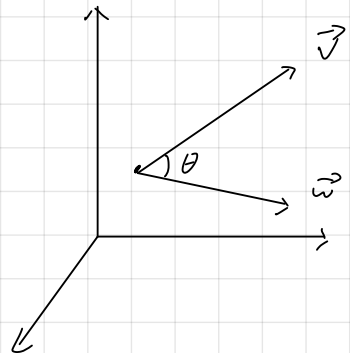
$$= w_0 + \sum_{i=1}^{n} w_i \cdot x^{(i)}$$

Linear Algebra (Review):

Vector:

$$\vec{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix} \qquad \vec{v} \in \mathbb{R}^d.$$

Geometrically



$\|\vec{v}\|$, $\frac{v^T w}{\|v\| \|w\|} = \cos\theta$.

$v^T w = \sum_{i=1}^{d} v_i \cdot w_i = w^T v$ ⇐ inner product.

$$[v_1, \cdots, v_d] \cdot \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

Matrix :

$$A = \begin{pmatrix} \vec{u}_1^T \\ \vdots \\ \vec{u}_n^T \end{pmatrix} = \begin{pmatrix} u_{11} & \cdots & u_{1m} \\ \vdots & & \vdots \\ u_{n1} & \cdots & u_{nm} \end{pmatrix} \qquad A \in \mathbb{R}^{n \times m}$$

$$= (\vec{v}_1 \cdots \vec{v}_m).$$

Matrix Vector :

$$\vec{y} = A \cdot w = \begin{bmatrix} u_{11} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & u_{nm} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}$$

$A \in \mathbb{R}^{n \times m}$, $w \in \mathbb{R}^m$

$$= \begin{bmatrix} u_1^T \cdot w \\ \vdots \\ u_n^T \cdot w \end{bmatrix} \in \mathbb{R}^n$$

$$= \sum_{i=1}^{m} w_i \cdot \vec{v}_i$$

$\vec{y} \in \text{span}\{\vec{v}_1, \cdots, \vec{v}_m\}$

Using L.A. for Multiple linear regression :

So, $w_0 + \sum_{i=1}^{d} w_i \cdot x^{(i)}$ ↓ $\vec{w}^T \cdot \vec{x}$

$$\vec{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$= w_0 x^{(0)} + \sum_{i=1}^{d} w_i \cdot x^{(i)}, \quad \text{where} \quad x^{(0)} = 1.$$

$$= \sum_{i=0}^{d} w_i \cdot x^{(i)}$$

$$= \vec{x}^T \cdot \vec{w} \Leftarrow \text{Thus, } \vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}, \vec{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}.$$

$$\vec{w}, \vec{x} \in K^{(d+1)}$$

$$H(\vec{x}) = \vec{x}^T \vec{w}$$

$$\vec{x} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} \in K^{d+1}$$

$$\vec{w} = \begin{pmatrix} w_0 \\ \vdots \\ w_d \end{pmatrix} \in K^{d+1}$$

Then, $y = H(\vec{x}) = \vec{x}^T \cdot \vec{w}$ ← parameter to learn

A data set will be like

	Salary	YOE	location	...
	y_1	$x_1^{(1)}$	$x_1^{(2)}$	...
	y_2	$x_2^{(1)}$	$x_3^{(2)}$	
	⋮	⋮	⋮	
	y_n	$x_n^{(1)}$	$x_n^{(2)}$	

Data point: $(\vec{x}, y)$.

$$D = \{(\vec{x}_1, y_1) \cdots (\vec{x}_n, y_n)\}.$$

Least Square:

$$\text{Loss}(H(\vec{x}, y)) = (y_i - H(\vec{x}_i))^2.$$

$$\text{Loss}(\vec{w}, (\vec{x}_i, \vec{y})) = (y_i - \vec{x}_i^T \vec{w})^2.$$

$$\text{So, } R(\vec{w}, D) = \frac{1}{n} \sum_{i=1}^{n} \text{loss}(\vec{w}, (\vec{x}_i, y_i))$$

$$= \frac{1}{n} \sum_{i=1}^{n} (y_i - \vec{x}_i^T \vec{w})^2$$

$$= \frac{1}{n} \| y_i - \vec{x}^T \vec{w} \|^2$$

$$= \frac{1}{n} \| \vec{V} \|^2$$

$$\| \vec{v} \|^2 = \vec{v}^T \vec{v} = \sum_{i=1}^{n} v_i^2$$

let $V_i = y_i - \vec{x}_i^T \vec{w}$

$$\text{So, } \vec{V} = \begin{bmatrix} y_1 - \vec{x}_1^T \vec{w} \\ y_2 - \vec{x}_2^T \vec{w} \\ \vdots \\ y_n - \vec{x}_n^T \vec{w} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \vec{x}_1^T \vec{w} \\ \vdots \\ \vec{x}_n^T \vec{w} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \vec{w} \begin{bmatrix} \vec{x}_1 \\ \vdots \\ \vec{x}_n \end{bmatrix}$$

$\vec{w} \in K^{d+1}$

$$= \vec{y} - X \cdot \vec{w}$$

$\vec{y} \in K^n$, $X \in K^{n \times (d+1)}$, $\vec{w} \in K^{d+1}$

$X \in K^{n \times (d+1)}$.

Design Matrix

$y = H(\vec{x}) = \vec{x}^T \cdot \vec{w}$

Thus,

$X \in R^{n \times (d+1)}$

$\vec{y} \in R^n$

$\vec{w} \in R^{d+1}$

Empirical Risk

$\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{1(1)} & \cdots & x_{1(d)} \\ \vdots & & \vdots \\ x_{n(1)} & \cdots & x_{n(d)} \end{pmatrix}$

Design Matrix

$R(\vec{w}; D) = \frac{1}{n} \| \vec{y} - X \cdot \vec{w} \|^2$

$= \frac{1}{n} (\vec{y} - X \cdot \vec{w})^T (\vec{y} - X \cdot \vec{w})$

$= \frac{1}{n} (\vec{y}^T - \vec{w}^T X^T)(\vec{y} - X \cdot \vec{w})$

$= \frac{1}{n} (\vec{y}^T \vec{y} - \vec{y}^T X \vec{w} - \vec{w}^T X^T \vec{y} + \vec{w}^T X^T X \vec{w})$

$\vec{y}^T X \vec{w}$: scalar $\Rightarrow (\vec{y}^T X \vec{w})^T = \vec{w}^T X^T y$

$\vec{w} = \underset{\vec{w} \in R^{d+1}}{\operatorname{argmin}} R_{sq}(\vec{w}, D)$

$= \frac{1}{n} (\vec{y}^T \vec{y} - 2 \vec{w}^T X^T \vec{y} + \vec{w}^T X^T X \vec{w})$

$\nabla_{\vec{w}} R_{sq}(\vec{w}, D)$: R^1 R^1 R^1

$= \frac{1}{n} (\nabla_{\vec{w}} y^T y - 2 \nabla_{\vec{w}} \vec{w}^T X^T \vec{y} + \nabla_{\vec{w}} \vec{w}^T X^T X \vec{w})$

$= 0 \quad = X^T \vec{y} \in R^{d+1} \quad = 2 X^T X \vec{w}$ (from Hw4)

Since $\nabla_{\vec{w}} \vec{w}^T \vec{b}$

$= \begin{pmatrix} \frac{\partial}{\partial w_0} \\ \vdots \\ \frac{\partial}{\partial w_1} \end{pmatrix} (w_1 b_1 + \cdots + w_n b_n)$

$= \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \vec{b}$

Thus, $0 = \frac{1}{n} (-2 X^T \vec{y} + 2 X^T X \vec{w})$

$0 = -X^T \vec{y} + X^T X \vec{w}$

$X^T \vec{y} = X^T X \vec{w}$

Normal Equation

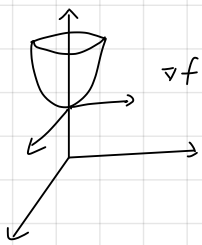
$X^T X$: $R^{(d+1) \times n}$ $R^{n \times (d+1)}$ $\Rightarrow R^{(d+1) \times (d+1)}$

invertible : square matrix & full rank (column of $X^T X$ linearly independent)

From one-dimensional LS

$\begin{cases} \frac{\partial}{\partial w_0} R((w_0, w_1), D) = 0 \\ \frac{\partial}{\partial w_1} R((w_0, w_1), D) = 0 \end{cases}$

multi-dimensional w.r.t vector



$\begin{pmatrix} \frac{\partial}{\partial w_0} R(\vec{w}, D) \\ \vdots \\ \frac{\partial}{\partial w_d} R(\vec{w}, D) \end{pmatrix} = 0$

$\begin{pmatrix} \frac{\partial}{\partial w_0} \\ \vdots \\ \frac{\partial}{\partial w_d} \end{pmatrix} R(\vec{w}, D) = 0$

$\nabla_{\vec{w}} R(\vec{w}, D) = 0$

$\in R^{d+1}$ R

Thus, $\exists (X^TX)^{-1}$,

$(X^TX)^{-1} X^TX = I$

Our normal equation becomes

$$\vec{w} = (X^TX)^{-1} X^T \vec{y}$$

why matter?

$(X^TX)\vec{w} = \sum_{i=0}^{d} w_i \cdot \vec{b_i}$

if dependent, $\exists \vec{w}' \neq 0$, s.t. $(X^TX)\vec{w}' = 0$.

→ linear combination of X^TX s.t. $= 0$.

Q: When is X^TX full rank?

① X is square matrix, $n = d+1$

$$X^TX = \begin{pmatrix} 1 & \cdots & 1 \\ x_{1(1)} & \cdots & x_{n(1)} \\ \vdots & & \vdots \\ x_{1(d)} & \cdots & x_{n(d)} \end{pmatrix} \begin{pmatrix} 1 & x_{1(1)} & \cdots & x_{n(1)} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1(d)} & \cdots & x_{n(d)} \end{pmatrix}$$

$X^T(X\vec{w}) \rightarrow X^T$ full rank $\rightarrow$ X full rank

(data point not repeating or not linearly combination of other data point)

② data point linealy dependent

⇓

$n < d+1$ over-parametrice

⇓

infinity many solutions.

always solution of normal equation

Meaning of optimal solution

$\vec{y} \simeq H(\vec{x}) = \vec{X}^T \vec{w}^*$

$= w_0^* + w_1^* x_{(1)} + w_2^* x_{(2)} + w_3^* x_{(3)} + \cdots + w_d^* x_{(n)}$

Salary ↑

Y.O.E. ↑ (average salary of industry)

years of between job ↑

w_3^* : 100,000 -> 100,001

w_1^* : 1 year -> 2 years

⇓

$w_1^* \gg w_3^*$

w_3^* : K$100 -> K$101

Scalar

Standarlize (z-score)

$$\frac{X_i(1) - \bar{X}(1)}{STD(D(1))}$$

↓

$$STD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

If we standalize ⇓

w^* reflect correlation.

ex) $H(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \dots + w_d x^d$ ← linear regression okay

limitation of linear regression?

No, b/c Taylor series idea -> polynomial approximate any function.

$H(x) = w_0 + w_1 \sin(x) + d_1 \cos(x) + w_2 \sin(2x) + d_2 \cos(2x) + \dots$

↑

can be used to model periodical function.

Problem: • Overfitting

• efficience

ex). Free falling obj.

$H(x) = w_0 + w_1 x + w_2 x^2 + \dots$ (not necessary)

$$H(\vec{x}) = w_0 + w_1^T \vec{x} + \vec{x}^T W_2^{d\times d} \vec{x} \quad \cdots$$

hard to solve (high dimensional)

Hierarchial model (deep learning model).

classification problem

0

1

$W^1 \in \mathbb{R}^{\square \times d}$ ← depends on how large the next layer you want

$$h^1(\vec{x}) = \sigma(W^1 \vec{x})$$

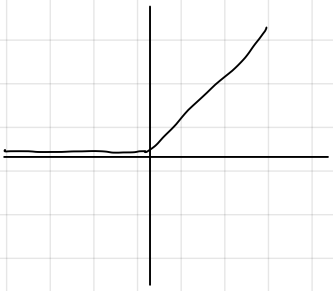
$$h^2(\vec{x}) = \sigma(W^2 h^1(\vec{x}))$$

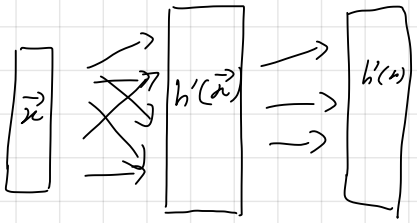
$$\vdots$$

$$h^t(\vec{x}) = \sigma(W^t h^{t-1}(\vec{x}))$$

$$\vdots$$

$$h^T(\vec{x}) = \vec{w}^T h^{T-1}(\vec{x})$$



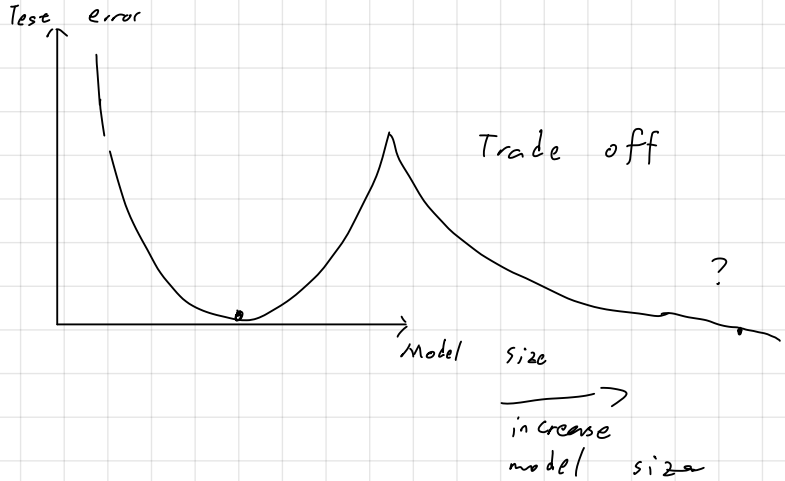


$$\sigma_i(\vec{z}) = \begin{cases} 0, & z_i < 0 \\ z_i, & z_i \geq 0 \end{cases}$$

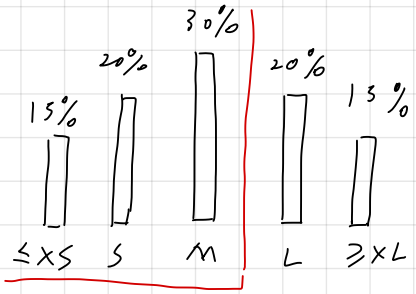
Rectified Linear Unit

On Final

Vector Calculus



Probability theory



distribution are similar

Probability Statement :

Outcome C "≤XS", "S", "M", "L", "≥XL")

Sample space $S = \{$ ↓ $\}$

Event E is a subset of sample space

Probability $\mathbb{P}$: a function that maps event to a real number.

① $0 \le \mathbb{P}(E) \le 1$, $\forall E \subseteq S$

② $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$, $\forall E_1$ and E_2 not intersecting each other.

$\forall E_1 \cap E_2 = \emptyset$

$E_1 = \{$ "≤XS", "S," ,"M" $\}$

$E_2 = \{$ "M", "L" $\}$

$E_1 \cup E_2 = \{$ "≤XS", "S", "M", "L" $\}$

$E_1 \cap E_2 = \{$ "M" $\}$

Discrete Sample Space :

$$\mathbb{P}\left(\bigcup_{i=1}^{n} \{s_i\}\right) = \mathbb{P}\left(\bigcup_{i=1}^{n-1} (s_i) \cup s_n\right)$$

$$= \mathbb{P}\left(\bigcup_{i=1}^{n-1} (s_i)\right) + \mathbb{P}(s_n)$$

$$\cdots$$

$$= \sum_{i=1}^{n} \mathbb{P}(\{s_i\})$$

$$= 1$$

point :

interval :

⇒ Probability P : a fuction maps outcomes to real number

① $0 \le \mathbb{P}(\{s_i\}) \le 1$, $\forall s_i \in S$.

② $\sum_{i=1}^{n} \mathbb{P}(\{s_i\}) = 1$

ex). $E_1 = \{1, 2, 3\}$, $E_2 = \{1, 5, 6\}$

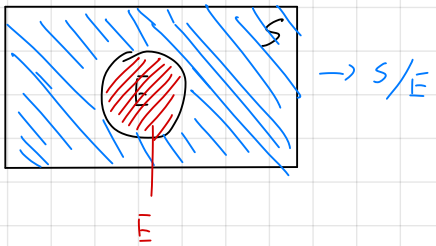
$\mathbb{P}(E_1) = \frac{1}{2} = \mathbb{P}(E_2)$ → Uniform Probability

Venn Diagram

Event E. $\widetilde{E} = S \setminus E$, $\mathbb{P}(\widetilde{E}) = 1 - \mathbb{P}(E)$.

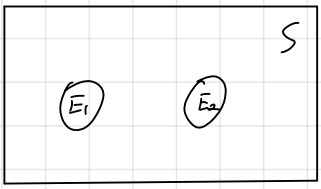
↑ complement of E

$\mathbb{P}(E) + \mathbb{P}(\widetilde{E}) = \mathbb{P}(E \cup \widetilde{E}) = \mathbb{P}(S) = 1$.



① $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$, if $E_1 \cap E_2 = \emptyset$

↑ mutually exclusive.



② $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$.

$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1 \setminus E_1 \cap E_2) + \mathbb{P}(E_2 \setminus E_1 \cap E_2) + \mathbb{P}(E_1 \cap E_2)$.

$= \mathbb{P}(E_1) - \mathbb{P}(E_1 \cap E_2) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2) + \mathbb{P}(E_1 \cap E_2)$

$= \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$.

ex). 0.3 prob. roommates home → A

0.2 prob. roommates partner home → B

0.1 prob. both of them home → $A \cap B$

Prob no one is at home ← let C : no one is at home

$\mathbb{P}(C) = 1 - \mathbb{P}(\bar{C}) = 1 - \mathbb{P}(A \cup B)$

$= 1 - (\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B))$

$= 1 - 0.3 - 0.2 + 0.1$

$= 0.6$.

Sample space $= \{$(R home, P home),
(R not h, P h),
(R h, P not h),
(R not h, P not h)$\}$

③ $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\,\mathbb{P}(E_2|E_1) = \mathbb{P}(E_2)\cdot\mathbb{P}(E_2|E_1)$

↑ prob of E_2, given condition on E_1.

$E_1, E_2 \rightarrow \mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\,\mathbb{P}(E_2)$

↑ independent ⇕

$\mathbb{P}(E_2) = \mathbb{P}(E_2|E_1)$.

④ Are mutually exclusive event independent?

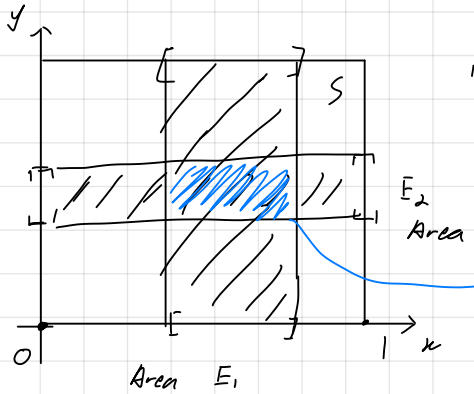
$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(\emptyset) = 0$

$\mathbb{P}(E_1|E_2) = 0$ if $\mathbb{P}(E_2) \neq 0$.

$= \dfrac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)}$ ← $=0$ / ← $\neq 0$

a. if $\mathbb{P}(E_1) \neq 0$ & $\mathbb{P}(E_2) \neq 0$, $E_1 \cap E_2 = \emptyset$ ⇒ E_1, E_2 not independent

b.



independent

venn

diagram

$\mathbb{P}(E_1 E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$

ex) Roll a die 3 times

prob {face 1 never appear}

prob {face 1 appear at least one}. ⇐ B_1

sample space for each roll : $\{1, \cdots, 6\}$

$A_1 = \{$ roll 1 = face 1 $\}$. $\mathbb{P}(A_1) = \frac{1}{6}$

$\bar{A}_1 = \{$ roll 1 ≠ face 1 $\}$. $\mathbb{P}(\bar{A}_1) = \frac{5}{6}$

$\mathbb{P}(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3) = \mathbb{P}(\bar{A}_1) \cdot \mathbb{P}(\bar{A}_2) \cdot \mathbb{P}(\bar{A}_3)$

$= (\frac{5}{6})^3$.

$\mathbb{P}(B_1) = 1 - \mathbb{P}(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3)$

$= 1 - (\frac{5}{6})^3$

ex) Two pets Equal prob. {cat, dog}.

Q1 Prob both are dogs, given oldest is dog ? (D, D)

Q2 -----, at least one of them is dog ? (D, C), (C, D)
weaker condition (D, D).

Q1, Event $B = \{$ both are dogs. $\}$ $\mathbb{P}(B \cap A_1) = \mathbb{P}(B) = (\frac{1}{2})^2 = \frac{1}{4}$

Event $A_1 = \{$ oldest is dog $\}$ $\mathbb{P}(A_1) = \frac{1}{2}$

$$\mathbb{P}(B|A_1) = \frac{\mathbb{P}(B \cap A_1)}{\mathbb{P}(A_1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

Q2 Event $A_2 = \{$ at least one is dog $\}$ $\mathbb{P}(A) = 1 - \mathbb{P}(\bar{A}_2) = \frac{3}{4}$

$\bar{A}_2 = \{$ none of them are dogs $\}$ $\mathbb{P}(\bar{A}_2) = \frac{1}{4}$

$$\mathbb{P}(B|A_2) = \frac{\mathbb{P}(B \cap A_2)}{\mathbb{P}(A_2)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3} \qquad \mathbb{P}(B \cap A_2) = \mathbb{P}(B) = \frac{1}{4}$$

Condition Probability

$$\mathbb{P}(B \mid A)$$

A: condition

Law of Total Probability

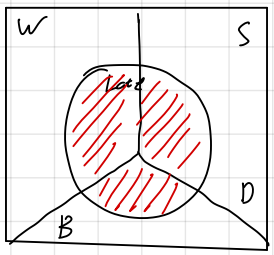
ex)

		Late	on time
Getting to campus	walk :	0.06	0.24
	bike :	0.03	0.07
	drive :	0.36	0.24

Q1 Prob a random is late?

$$\mathbb{P}(\text{Late}) = \mathbb{P}(\text{Late} \cap \text{walk}) + \mathbb{P}(\text{Late} \cap \text{bike}) + \mathbb{P}(\text{Late} \cap \text{drive}).$$

$$= \mathbb{P}(\text{Late} \mid \text{walk})\,\mathbb{P}(\text{walk}) + \mathbb{P}(\text{Late} \mid \text{bike})\,\mathbb{P}(\text{bike}) + \mathbb{P}(\text{Late} \mid \text{drive})\,\mathbb{P}(\text{drive}).$$



$(L \cap D) \cup (L \cap B) \cup (L \cap W) = L$ ← disjoint

Property $(A \cap B) \cup (A \cap C) = A \cap (B \cup C)$

$$\mathbb{P}(L \cap D) + \mathbb{P}(L \cap B) + \mathbb{P}(L \cap W) = \mathbb{P}(L)$$

Partition $\{E_1, \cdots, E_k\}$ is partition of S if

① $\mathbb{P}(E_i \cap E_j) = 0$, $\forall i \neq j$. $\quad (E_i \cap E_j = \emptyset,\ \forall i \neq j)$.

② $\sum_{i=1}^{k} \mathbb{P}(E_i) = 1$ $\quad \left(\bigcup_{i=1}^{k} E_i = S\right)$.

③ $\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A \cap E_i)$.

$$= \sum_{i=1}^{k} \mathbb{P}(A \mid E_i)\,\mathbb{P}(E_i)$$

Q2 $\mathbb{P}(\text{Walk} \mid \text{Late}) = \frac{\mathbb{P}(\text{Late} \wedge \text{walk})}{\mathbb{P}(\text{Late})} = \frac{0.06}{0.06 + 0.03 + 0.36}$

Q3 $\mathbb{P}(\text{Late}) = 0.45$

$\mathbb{P}(\text{walk}) = 0.3$

$\mathbb{P}(\text{Late} \mid \text{walk}) = 0.2$

$\mathbb{P}(\text{Walk} \mid \text{Late}) = \frac{\mathbb{P}(\text{walk} \wedge \text{Late})}{\mathbb{P}(\text{Late})} = \frac{\mathbb{P}(\text{Late} \mid \text{walk})\,\mathbb{P}(\text{walk})}{\mathbb{P}(\text{Late})}$

↓

Bayes Theorem

$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$

Statistical Model

↓

$$\frac{\mathbb{P}(\text{Data} \mid \text{Parameter})\,\mathbb{P}(\text{Parameter})}{\mathbb{P}(\text{Data})} = \mathbb{P}(\text{Parameter} \mid \text{Data})$$

↑ prior belief

$\mathbb{P}(\text{Rain} \mid \text{Data})$

ex) Product detects Covid-19 w/ 95% prob.

15% prob. turning positive for people w/o covid-19.

10% have Covid-19

Q: Prob. people get covid-19, if they test positive?

T: {test positive}

H: {have covid-19}

$\mathbb{P}(T) = \mathbb{P}(T \mid H)\,\mathbb{P}(H) + \mathbb{P}(T \mid H^c)\,\mathbb{P}(H^c)$

$= 0.95 \cdot 0.1 + 0.15 \cdot 0.9$

$\mathbb{P}(T \mid H) = 0.95$

$\mathbb{P}(T \mid H^c) = 0.15$

$\mathbb{P}(H) = 0.1$

$$\mathbb{P}(H \mid T) = \frac{\mathbb{P}(T \mid H)\,\mathbb{P}(H)}{\mathbb{P}(T)} = \frac{0.95 \cdot 0.1}{\mathbb{P}(T)} \approx 0.4$$

Differential Privacy

538 website takes a poll

Republic.	Dem.
49%	51%

- Differential attack.
- Privacy - accuracy trade off
- Guard against privacy leakage

Flip a coin — "head" -> answer faith fully

"tail" -> answer 50% - 50%

Rep. Dem.

Q: How safe is this ?

Denote "Rep" - 0

"Dem" - 1

Private preference - $X_n \in \{0, 1\}$

Observation - $Y_n \in \{0, 1\}$

Question: $\mathbb{P}(X_n = 1 \mid Y_n = 0) = ?$ ——> $\mathbb{P}$(prefer Dem. | Answered "Rep").

(private ↓, observed ↑)

$$= \frac{\mathbb{P}(Y_n = 0 \mid X_n = 1)\,\mathbb{P}(X_n = 1)}{\mathbb{P}(Y_n = 0)}$$

$\mathbb{P}(Y_n = 0 \mid X_n = 1) = \mathbb{P}(\text{faithful})\,\mathbb{P}(Y_n = 0 \mid X_n = 1, \text{faithful}) + \mathbb{P}(\text{Not faithful})\,\mathbb{P}(Y_n = 0 \mid X_n = 1, \text{not faithful})$

$= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2}$

$= \frac{1}{4}$

$\mathbb{P}(Y_n = 0 \mid X_n = 0) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$

$\mathbb{P}(X_n = 1) = 0.51$
$\mathbb{P}(X_n = 0) = 0.49$
) Assume Distribution of vote

Law of Total Probability

$\mathbb{P}(Y_n = 0) = \mathbb{P}(Y_n = 0 \mid X_n = 0)\,\mathbb{P}(X_n = 0) + \mathbb{P}(Y_n = 0 \mid X_n = 1)\,\mathbb{P}(X_n = 1)$

$= \frac{3}{4}(0.49) + \frac{1}{4}(0.51)$

$= 0.495$

$$\frac{\mathbb{P}(Y_n = 0 \mid X_n = 1)\,\mathbb{P}(X_n = 1)}{\mathbb{P}(Y_n = 0)} = \frac{\frac{1}{4} - 0.51}{0.495} \approx \frac{1}{4}.$$

$$\mathbb{P}(X_n = 0 \mid Y_n = 0) = 1 - \mathbb{P}(X_n = 1 \mid Y_n = 0) \approx \frac{3}{4}.$$

↑ prob that attacker success.

Q: Can we recover the poll result?

Don't know $\mathbb{P}(X=1)$ but know $\mathbb{P}(Y=1)$.

Expectation:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\left(\sum_{x_i \in D_0} x_i + \sum_{x_j \in D_1} x_j\right)$$

$n = n_0 + n_1$ (↑ Rep. D_0, ↑ Dem. D_1)

$$= \frac{1}{n}(n_0 \cdot 0 + n_1 \cdot 1)$$

$$= \frac{n_0}{n} \cdot 0 + \frac{n_1}{n} \cdot 1$$

As $n \to \infty$ $= \mathbb{P}(X=0)\cdot 0 + \mathbb{P}(X=1)\cdot 1$

$= \mathbb{P}(X=1)$

So, $\frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{n\to\infty} \sum_{k=1}^{k} \mathbb{P}(z = a_k)\, a_k$ ← PMF

↑ value pmf take

Expectation.

Similarly, $\frac{1}{n}\sum_{i=1}^{n} Y_i = \mathbb{P}(Y=0)\cdot 0 + \mathbb{P}(Y=1)\cdot 1$

$= \mathbb{P}(Y=1)$ ← law of total probability

↑ Expectation of Y

$= \mathbb{P}(Y=1 \mid X=0)\,\mathbb{P}(X=0) + \mathbb{P}(Y=1 \mid X=1)\,\mathbb{P}(X=1)$

$= \frac{1}{4}\mathbb{P}(X=0) + \frac{3}{4}\mathbb{P}(X=1)$.

$= \frac{1}{4}(1 - \mathbb{P}(X=1)) + \frac{3}{4}\mathbb{P}(X=1)$

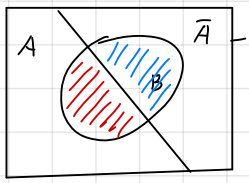
$= \frac{1}{4} - \frac{1}{4}\mathbb{P}(X=1) + \frac{3}{4}\mathbb{P}(X=1)$

proof:

$\mathbb{P}(A \mid B) \overset{?}{=} \mathbb{P}(\bar{A} \mid B)$

$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \overset{?}{=} \frac{\mathbb{P}(B)}{\mathbb{P}(B)} - \frac{\mathbb{P}(\bar{A} \cap B)}{\mathbb{P}(B)}$

$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B) - \mathbb{P}(\bar{A} \cap B)}{\mathbb{P}(B)}$



$(A \cap B) \cup (\bar{A} \cap B)$

$= B$

$\mathbb{P}(A \cap B) + \mathbb{P}(\bar{A} \cap B)$

$= \mathbb{P}(B)$

$= \frac{1}{4} + \frac{1}{2}\mathbb{P}(X=1)$.

↑

Expectation of X

$\mathbb{P}(Y=1) = \frac{1}{4} + \frac{1}{2}\mathbb{P}(X=1)$ This is how to recover the result!

$\mathbb{P}(X=1) = 2\left(\mathbb{P}(Y=1) - \frac{1}{4}\right)$.

Probability on continuous space

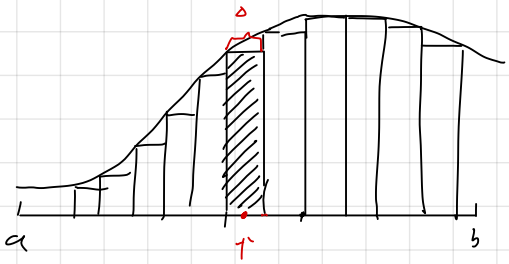
Discrete: $s \in S$, ① $0 \le \mathbb{P}(s) \le 1$.

② $\sum_{s \in S} \mathbb{P}(s) = 1$.

$\lim_{\substack{k\to\infty \\ \Delta\to 0}} \sum_{k=1}^{k} \mathbb{P}\left(X \in \left[s_k - \frac{\Delta}{2}, s_k + \frac{\Delta}{2}\right]\right) = 1$.

$= \lim_{\Delta\to 0} \sum_{k=1}^{k} p(s_k)\cdot\Delta$

$= \int_a^b p(s)\,ds$.



$\lim_{\Delta\to 0} \frac{\mathbb{P}\left(X \in \left[s - \frac{\Delta}{2}, s + \frac{\Delta}{2}\right]\right)}{\Delta} = p(s)$

↑

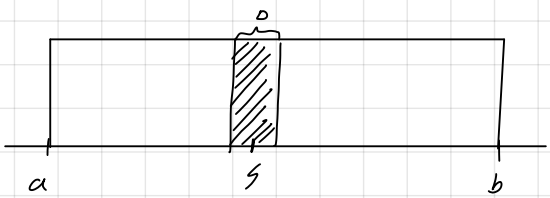
Probability Density Function

Continuous:

① $p(s) \ge 0$

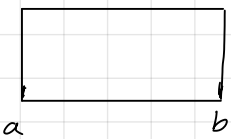
② $\int_{s\in S} p(s)\,ds = 1$.

Uniform probability case



$\mathbb{P}\left(X \in \left[s - \frac{\Delta}{2}, s + \frac{\Delta}{2}\right]\right) = \frac{\Delta}{b-a}$

- Uniform Distribution.



p.d.f. $p(s) = \frac{1}{b-a}$

- Normal Distribution.

p.d.f. defined on $\mathbb{R}$.

p.d.f $p(s) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left(-(s-\mu)^2/2\sigma^2\right)$

$\int p(s)\,ds = 1$

$\Leftarrow \int_{\mathbb{R}} \exp\left((s-\mu)^2/2\sigma^2\right) ds$

$= \sqrt{2\pi}\cdot\sigma$

μ: mean / Expectation of distribution

↓

① Discrete Expectation : $\frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow[n\to+\infty]{} E[X]$

$$E[X] = \sum_{k=1}^{k} \mathbb{P}(X = s_k)\cdot s_k$$

② Continuous Expectation:

$$E[X] = \lim_{\Delta\to 0} \sum_{k=1}^{k} \mathbb{P}\left(X \in [s_k - \tfrac{\Delta}{2}, s_k + \tfrac{\Delta}{2}]\right)\cdot s_k$$

$$= \lim_{\Delta\to 0} \sum_{k=1}^{k} p(s_k)\,\Delta \cdot s_k$$

$$= \int_{s\in S} p(s)\cdot s \, ds.$$

For normal distribution: $E[X] = \int_k P_N(s)\cdot s\, ds = \mu$

σ: STD σ^2: Variance

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

↓ $n\to\infty$

$$\int_{s\in S} p(s)\cdot(s-\mu)^2\, ds = E[(X-\mu)^2].$$

↑

$$\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \xrightarrow[n\to\infty]{} \sum_{k=1}^{k}\mathbb{P}(x = s_k)\cdot(s_k-\mu)^2.$$

$$E[f(x)] = \sum_{k=1}^{k}\mathbb{P}(X = s_k)\cdot f(s_k).$$

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\int_{-\infty}^{\infty} \exp\left(-(s-\mu)^2/2\sigma\right)(s-\mu)^2\, ds.$$

let $y = s-\mu$

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\int_{-\infty}^{\infty} \exp\left(\frac{-y^2}{2\sigma^2}\right)(y^2)\, dy.$$

do integral by parts twice

$$\exp'(-y^2/2\sigma) = \exp(-y^2/2\sigma^2)\cdot\left(\frac{-2y}{2\sigma^2}\right) dy$$

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\int_{-\infty}^{\infty} \exp\left(\frac{-y^2}{2\sigma^2}\right)(y^2)\,dy.$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\left(-\frac{\sigma^2}{y}\right)\cdot d\exp\left(-y^2/2\sigma^2\right).$$

$$= -\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left(\frac{-y^2}{\sigma^2}\right)\cdot(-\sigma^2)\,dy$$

$$= \sigma^2\cdot\frac{1}{\sqrt{2\pi}\sigma}\underbrace{\int_{-\infty}^{\infty}\exp(-y^2/\sigma^2)\,dy}_{\sqrt{2\pi}\cdot\sigma}$$

$$= \sigma^2$$

- $X \sim N(\mu, \sigma^2)$ $\quad p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-(x-\mu)^2/2\sigma^2)$

$$E[X] = \int_{-\infty}^{\infty} p(x)\cdot x\,dx = \mu$$

$$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} p(x)(x-\mu)^2\,dx = \sigma^2.$$

$$E[f(x)] = \int_{-\infty}^{\infty} p(x) f(x)\,dx$$

$$E[X+Y] = \iint_{K^2} p_{(X,Y)}(x,y)\cdot(x+y)\,dx\,dy$$

$$\int_{-\infty}^{\infty} p_{(X,Y)}(x,y)\,dy = p(x)$$

$$p_{X,Y}(x,y) = \lim_{\Delta\to 0}\frac{\mathbb{P}(X\in[x-\frac{\Delta}{2}, x+\frac{\Delta}{2}]\wedge Y\in[y-\frac{\Delta}{2}, y+\frac{\Delta}{2}])}{\Delta^2}$$

$$= \iint_{K^2} p_{(X,Y)}(x,y)\cdot x\,dx\,dy + \iint_{K^2} p_{(X,Y)}(x,y)\,y\,dx\,dy$$

$$\int_{-\infty}^{\infty} p_{X,Y}(x,y)\,dy = \sum_{\Delta=1}^{\infty}\frac{\mathbb{P}(X\in[x-\frac{\Delta}{2}, x+\frac{\Delta}{2}]\wedge Y\in[y-\frac{\Delta}{2}, y+\frac{\Delta}{2}])}{\Delta^2}\cdot\Delta$$

$$= \frac{\mathbb{P}(X\in[x-\frac{\Delta}{2}, x+\frac{\Delta}{2}]\wedge Y\in K)}{\Delta}$$

$$= \frac{\mathbb{P}(X\in[x-\frac{\Delta}{2}, x+\frac{\Delta}{2}])}{\Delta} = p_X(x)$$

$$= \int_K x\int_K p_{X,Y}(x,y)\,dy\,dx + \int_K y\int_K p_{X,Y}(x,y)\,dx\,dy$$

$$= \int_K x\,p_X(x)\,dx + \int_K y\,p_Y(y)\,dy$$

$$= E[X] + E[Y]$$

Independence & Expectation

$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

$$p_{(X,Y)}(x,y) = p_X(x)\, p_Y(y)$$

b/c

$$p_{X,Y}(x,y) = \lim_{\Delta \to 0} \frac{\mathbb{P}(X \in [x-\frac{\Delta}{2}, x+\frac{\Delta}{2}]) \cdot \mathbb{P}(Y \in [y-\frac{\Delta}{2}, y+\frac{\Delta}{2}])}{\Delta^2} \quad \text{when independence}$$

$$= \lim_{\Delta \to 0} \underbrace{\frac{1}{\Delta}\mathbb{P}(X \in [x-\frac{\Delta}{2}, x+\frac{\Delta}{2})}_{\downarrow\ p_X(x)} \underbrace{\frac{1}{\Delta}\mathbb{P}(Y \in [y-\frac{\Delta}{2}, y+\frac{\Delta}{2}])}_{\downarrow\ p_Y(y)}$$

$$E[XY] = \iint_R xy\, p_X(x) p_Y(y)\, dx\, dy \quad \text{for independence.}$$

$$= \int x\, p_X(x) \int y\, p_Y(y)\, dx\, dy$$

$$= \int x\, p_X(x)\, E[Y]$$

$$= E[X]\,E[Y]$$

$X \sim N(\mu_1, \sigma_1^2) \Rightarrow X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ when independence.
$Y \sim N(\mu_2, \sigma_2^2)$

b/c

$$E[X+Y] = E[X] + E[Y]$$

$$Var(X+Y) = E[(X+Y - E[X+Y])^2]$$

$$= E[((X-E[X]) + (Y-E[Y]))^2]$$

$$= E[(X-E[X])^2 + (Y-E[Y])^2 + 2(X-E[X])(Y-E[Y])]$$

$$= E[(X-E[X])^2] + E[(Y-E[Y])^2] + E[2(X-E[X])(Y-E[Y])].$$

$$= Var(X) + Var(Y) + 2\,E[(X-E[X])(Y-E[Y])].$$

$\downarrow \sigma^2$ $\downarrow \sigma^2$

$$= E[X-E[X]]\,E[Y-E[Y]].$$

$$= E[X] - E[E[X]] \cdots$$

$$= E[X] - E[X] \cdots$$

$$= 0$$

- $X + X = 2X \sim N(2\mu, 4\sigma^2)$, $\quad X + Y \sim N(2\mu, 2\sigma^2)$ (due to cancellation)

  Y independent of X

  $E[2X] = 2E[X] = 2\mu$

  $E[(2x - E[2x])^2] = 4E[(x - E[X])^2]$

  $= 4\sigma^2$

Empirical Risk Minimization

- $\text{loss}(h, x_i)$
- $R(H; D) = \frac{1}{n} \sum_{i=1}^{n} \text{loss}(h, x_i)$

Go back to differential privacy:

Mechanism: w/ ε probability faithfully

w/ $1-\varepsilon$ probability unfaithfully random

Can attacker deduce x, based on data?

Assume they know: $\mathbb{P}(\text{"Dem"}) = \frac{1}{4}$.

Observe $\{\text{"D"}, \text{"R"}, \text{"R"}, \text{"D"}, \text{"R"}, \text{"R"}\} = S$.

If we know $\mathbb{P}(X \in [a - \frac{\Delta}{2}, a + \frac{\Delta}{2}] \mid \text{Data} = s)$.

$$a^* = \arg\min_a \mathbb{P}(X \in [a - \tfrac{\Delta}{2}, a + \tfrac{\Delta}{2}] \mid \text{Data} = s).$$

$$a^* = \arg\min_a P_X(a \mid \text{Data} = s)$$

baye's

$$= \frac{\mathbb{P}(\text{Data} = s \mid X \in \cdots)\,\mathbb{P}(X \in \cdots)}{\mathbb{P}(\text{Data} = s)}$$

For small Δ: $= \frac{\mathbb{P}(\text{Data} = s \mid X = a)\, P_X(a)}{\mathbb{P}(\text{Data} = s)}$.

$$\propto \mathbb{P}(\text{Data} = s \mid X = a)\, P_X(a)$$

likehood $\quad$ prior probability

Probability Theory for ERM

DP problem with x prob, reveal true preference

$(1-x)$ prob, uniform random.

$\mathbb{P}(\text{"Dem"}) = \frac{1}{4}$, $\quad \mathbb{P}(\text{"Rep"}) = \frac{3}{4}$. Observe $\{\text{"D"}, \text{"R"}, \text{"R"}, \text{"D"}, \text{"R"}, \text{"R"}\} = S$.

$$\mathbb{P}\left(x \in [a-\tfrac{\Delta}{2}, a+\tfrac{\Delta}{2}] \mid Data = s\right) = P_X(x \mid Data = s)$$ ↑ Posterior

pick $x^* = \underset{a}{\operatorname{argmax}}\ \mathbb{P}\left(x \in [a-\tfrac{\Delta}{2}, a+\tfrac{\Delta}{2}] \mid Data = s\right)$.

Bayes Thm: $\mathbb{P}\left(x \in [a-\tfrac{\Delta}{2}, a+\tfrac{\Delta}{2}] \mid Data = s\right)$.

$$= \mathbb{P}\left(Data = s \mid x \in [a-\tfrac{\Delta}{2}, a+\tfrac{\Delta}{2}]\right) \mathbb{P}\left(x \in [a-\tfrac{\Delta}{2}, a+\tfrac{\Delta}{2}]\right) \Big/ \mathbb{P}(Data = s).$$

$$\propto \mathbb{P}\left(Data = s \mid x \in [a-\tfrac{\Delta}{2}, a+\tfrac{\Delta}{2}]\right) \mathbb{P}\left(x \in [a-\tfrac{\Delta}{2}, a+\tfrac{\Delta}{2}]\right).$$

small Δ

$$\propto \mathbb{P}(Data = s \mid x = a)\, P_X(a)$$

$\mathbb{P}(Data = s \mid x = a)$: ↑ Likelihood

$P_X(a)$: ↑ prior probability / belief

Maximum Likelihood Estimate

$$\mathbb{P}(Data = s \mid x = a) = \prod_{i=1}^{6} \mathbb{P}(Data(i) = s(i) \mid x = a).$$

⇕

$Data(1) = s(1) \wedge Data(2) = s(2) \cdots$ ↑ independence

$$\mathbb{P}(Data = \text{"D"} \mid x = a) = a \cdot \tfrac{1}{4} + (1-a) \cdot \tfrac{1}{2}$$

$$= \tfrac{1}{2} - \tfrac{1}{4}a$$

$$\mathbb{P}(Data = \text{"K"} \mid x = a) = 1 - \mathbb{P}(Data = \text{"D"} \mid x = a).$$

$$= \tfrac{1}{2} + \tfrac{1}{4}a$$

Empirical Risk

$$= \left(\tfrac{1}{2} - \tfrac{1}{4}a\right)^2 \left(\tfrac{1}{2} + \tfrac{1}{4}a\right)^4$$

↓

$$Risk(a \mid s) = -\ln \mathbb{P}(Data = s \mid x = a)$$

$$= -2\ln\left(\tfrac{1}{2} - \tfrac{1}{4}a\right) - 4\ln\left(\tfrac{1}{2} + \tfrac{1}{4}a\right).$$

$\mathbb{P}(Data(i+1) \mid Data(i), x = a)$

Not independent case

So $\frac{d}{da} R(a \mid s) = -2\,\frac{-\frac{1}{4}}{\frac{1}{2} - \frac{1}{4}a} - 4\,\frac{\frac{1}{4}}{\frac{1}{2} + \frac{1}{4}a} = 0$

⇓

$$\frac{1}{2} \cdot \frac{1}{\frac{1}{2} - \frac{1}{4}\alpha} = \frac{1}{\frac{1}{2} + \frac{1}{4}\alpha}$$

$$\frac{1}{2}\left(\frac{1}{2} + \frac{1}{4}\alpha\right) = \frac{1}{2} - \frac{1}{4}\alpha$$

$$\frac{1}{4} + \frac{1}{8}\alpha = \frac{1}{2} - \frac{1}{4}\alpha$$

$$\frac{3}{8}\alpha = \frac{1}{4}$$

$$\alpha = \frac{2}{3}$$

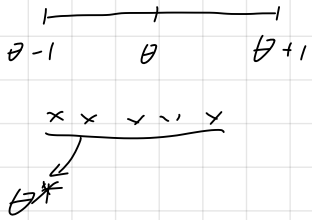
Q: What if the calculated α is negative?

⇓

prior belief

$$P_X(\alpha) = \begin{cases} 1, & 0 \le \alpha \le 1 \\ 0, & \text{else} \end{cases}$$

Relating to groupwork



$P(\theta|D)$: Posterio

$P(D|\theta)$: likelihood. $= \prod_{i=1}^{n} P(x_i|\theta)$

$x_i \sim Unif[\theta-1, \theta+1]$

$$P_{x_i}(x) = \begin{cases} \frac{1}{2}, & x_i \in [\theta-1, \theta+1] \\ 0, & \text{else} \end{cases}$$

if $x_1 \cdots x_n \in [\theta^*-1, \theta^*+1]$

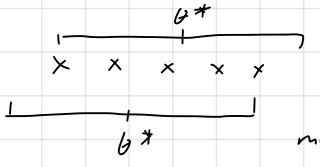
⇓

$P(D|\theta) > 0$

if $\exists x_i \notin [\theta^*-1, \theta^*+1]$,

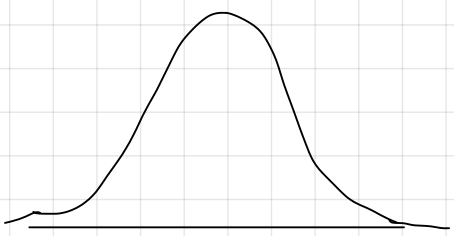
⇓

$P(D|\theta) = 0$



multiple θ^*

Central Limit Theorem:



$$P(\theta|D) \propto \prod_{i=1}^{n} P(x_i|\theta) \cdot P(\theta) \longrightarrow \text{Normal Distribution}$$

Bernstein von Mises

Q: Why not Model data as normal R.V. ?

$P(x_i|\theta)$: $x_i \sim N(\theta, \sigma^2)$.

Then $P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$ likelihood

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{(-(x_i-\theta)^2/2\sigma^2)}$$

$$\propto \prod_{i=1}^{n} e^{(-(x_i-\theta)^2/2\sigma^2)}$$

$$= e^{\left(-\sum_{i=1}^{n}(x_i-\theta)^2/2\sigma^2\right)}$$

$$= e^{\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)}$$

Empirical risk

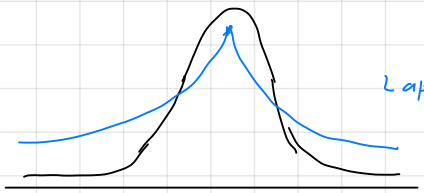
$$R(\theta, D) = -\ln P(D|\theta)$$

$$= \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2$$

→ Empirical risk squared loss

$$\theta^* = \arg\max_{\theta} P(D|\theta) = \arg\min_{\theta} R(\theta, D) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

With finite data, how to model outliers? Normal distribution



Laplace distribution

$P(x_i|\theta) \propto \exp(-(x_i-\theta)^2/2\sigma^2)$ value decrease faster

$$P(x_i|\theta) = \frac{1}{2}\exp(-|x_i-\theta|)$$

$$P(D|\theta) = \prod_{i=1}^{n}\exp(-|x_i-\theta|) \propto \exp\left(-\sum_{i=1}^{n}|x_i-\theta|\right)$$

$$R(\theta|D) = \sum_{i=1}^{n}|x_i-\theta|$$ Absolute loss

Regression

$$y_i \approx H(x_i, \theta)$$

Linear regression

$$y_i \simeq \langle \vec{x_i}, \vec{w} \rangle$$

$$y_i \sim N(\langle \vec{x_i}, \vec{w} \rangle, \sigma^2) \Rightarrow P(y_i \mid \vec{x_i}, \vec{w})$$

$$P(D \mid \vec{w}) = \prod_{i=1}^{\ell} P((x, y) \mid \vec{w})$$

$$P((x_i, y_i) \mid \vec{w}) = P(y_i \mid \vec{x_i}, \vec{w}) \underbrace{P(\vec{x_i} \mid \vec{w})}_{P(\vec{x_i})}$$

$$\propto P(y_i \mid \vec{x_i}, \vec{w}) \quad P(\vec{x_i})$$

$$\propto \exp\left(-(y_i - \langle \vec{x_i}, \vec{w} \rangle)^2 / 2\sigma^2\right).$$

Risk = least square

P.3

A : J

B : J or K or A

C : Heart

$P(A \cap B \mid C) = \frac{1}{12}$

$P(A \mid C)\, P(B \mid C)$

$= \frac{1}{12} \cdot \frac{4}{12} = \frac{4}{24} = \frac{2}{12}$

$P(A \mid B \cap C) = \frac{4}{8} = \frac{1}{2}$

$P(A \mid C) = \frac{1}{2}$

$P(A \mid B \cap C) = \frac{1}{4}$

$P(A \mid C) = \frac{1}{12}$

A : heart

B : face

C : red

condition

$P(A \cap B \mid C) = \frac{4}{24} = \frac{1}{6}$

$P(A \mid C)\, P(B \mid C) = \frac{12}{24} \cdot \frac{8}{24} = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$

$P(B \mid C)$