DATE
PAGE

ML-Assignment

Ishit Bajpai
2020380

1. To Prove: For Simple linear regression the least square fit passes through the $(\hat{x}, \hat{y})$.

Let $y = \vec{w} \cdot \vec{x} + b$ be the best fit line. (w are parameters and b is the bias) & $\vec{x}$ is input vector ~~consisting of features~~.

Alternatively; $y = w^T \vec{x} + b$

~~Since~~ Let $\hat{y}_i$ be the prediction for $\vec{x}_i$ and error in the prediction be ε_i.

$$\hat{y}_i = w^T \vec{x}_i + b$$

if y_i is the true value then

$$y_i = \hat{y}_i + \varepsilon_i \qquad i = 1, \cdots n.$$

$$\Rightarrow y_i = w^T \vec{x}_i + b + \varepsilon_i. \qquad i = 1 \cdots n.$$

Adding for $i = 1 \cdots n$.

$$\sum_{i=1}^{n} y_i = w^T \sum_{i=1}^{n} \vec{x}_i + \sum_{i=1}^{n} b + \sum_{i=1}^{n} \varepsilon_i.$$

$$\sum_{i=1}^{n} y_i = w^T \sum_{i=1}^{n} \vec{x_i} + nb + \sum_{i=1}^{n} \varepsilon_i \cdots (1)$$

On an informal note we can say that errors tend to cancel each other
$\therefore \sum_{i=1}^{n} \varepsilon_i = 0$. (as their $\mu = 0$)

Proving Mathematically.

$$J(w,b) = \frac{1}{2n} \sum_{i=1}^{n} (\hat{y_i} - y_i)^2 = \frac{1}{2n} \sum_{i=1}^{n} (w^T \vec{x_i} + b - y_i)^2$$

We have to minimize the cost function wrt to w and b.

$$\therefore \frac{\partial J(w,b)}{\partial b} = \frac{1}{2n} \cdot 2 \sum_{i=1}^{n} (w^T \vec{x_i} + b - y_i) \times 1$$

$$\frac{\partial J(w,b)}{\partial b} = 0$$ (Condition for finding ~~extr~~ critical point, ~~since this~~

$$\Rightarrow \sum_{i=1}^{n} (w^T \vec{x_i} + b - y_i) = 0$$

$$\Rightarrow -\sum_{i=1}^{n} \varepsilon_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} \varepsilon_i = 0.$$

Using (1)

$$\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = w^T \left(\frac{\sum_{i=1}^{n} x_i}{n}\right) + b + \bigcirc$$

$$\bar{Y} = w^T \bar{X} + b$$

$\therefore$ ~~$(\bar{X}, \bar{Y})$ passes through the~~

The least square fit line passes through $(\bar{X}, \bar{Y})$.

DATE
PAGE

b) Let's say two variables X and Y are highly correlated with a third variable Z.

Then it doesnot necessarily imply that X and Y are also highly correlated with each other.

for pearson's correlation

$$\gamma_{XY} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}$$

Using the result, ( )

$$\gamma_{XY} = \gamma_{YZ} \cdot \gamma_{ZX} - \sqrt{(1 - \gamma_{YZ}^2)(1 - \gamma_{ZX})^2}$$

- if we choose

ex 1 $\gamma_{YZ} = 0.75 \quad \gamma_{ZX} = 0.8$

We get $\gamma_{XY} = 0.2031$.

- if we choose

$\gamma_{YZ} = 0.7 \quad \gamma_{ZX} = 0.7$

We get $\gamma_{XY} = -0.02$.

The above two examples shows that even though r_{xy} & r_{xz} and r_{zx} are highly correlated, it doesnot imply r_{xy} also will also be highly correlated.

Informal example

Say we have X = Sugar, Y = Salt and Z = Water.

X, Z are highly correlated in sense that sugar dissolves easily in water.

Y, Z are also highly correlated as salt dissolves with ease in water However salt and sugar doesnot dissolve when mixed i.e that is they are not co-related.

Eg Using Partial Correlation.

$$\rho_{xy} = \rho_{xz}\rho_{zy} + \rho_{xy.z}\left(\sqrt{1-\rho_{xz}^2}\sqrt{1-\rho_{zy}^2}\right)$$

Since $\rho_{xy.z} \in [-1, +]$

$$\Rightarrow \rho_{xy} = \rho_{xz}\rho_{zy} \pm \sqrt{1-\rho_{xz}^2}\sqrt{1-\rho_{zy}^2}$$

if $\rho_{xz} = 0.7$ $\rho_{zy} = 0.8$

then $\rho_{xy} \in [0.131, 0.9884]$

Thus even though correlation b/w
ρ_{zy} & ρ_{xz} was high
ρ_{xy} can still take low value.

5) To Prove : weak law of large numbers

It states that as number of iid trials approaches ∞, the observed mean approaches theoretical Expected Mean.

OR

Sample Mean $\xrightarrow[n \to \infty]{}$ True Mean

where n is the # of IID Trials.

Let X_i ($i = 1 \dots n$) be n IID's with mean μ and variance $= \sigma^2$

True Mean $= E(X_i) = \mu \quad (i = 1 \dots n)$

$Var(X_i) = \sigma^2 \quad i = 1 \dots n$

∴ we have to prove

$$\text{lt}_{n \to \infty} \left(\frac{X_1 + X_2 + \dots X_n}{n} \right) = \mu$$

OR

$$P\left(\left| \frac{X_1 + X_2 \dots X_n}{n} - \mu \right| \geq \varepsilon \right) = 0$$

$$\text{Var}\left(\frac{X_1 + X_2 - - - + X_n}{n}\right)$$

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

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i \neq j} \text{Cov}(X_i, X_j)\right)$$

$$= \frac{1}{n^2}(n \times \sigma^2 + 0)$$ (as Covariance of IID Variables is equal to 0)

$$= \left(\frac{\sigma^2}{n}\right)$$

$\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$

$= E[X]E[Y] - E[X]E[Y]$

$= 0$

$$\therefore \text{Var}\left(\frac{X_1 + X_2 - - X_n}{n}\right) = \frac{\sigma^2}{n}$$

Using Chebyshev inequality

$$P\left(\left|\frac{X_1 + X_2 - - X_n}{n} - \mu\right| \geq \varepsilon\right) \leq$$

$$\frac{1}{\varepsilon^2}\text{Var}\left(\frac{X_1 + X_2 - - X_n}{n}\right)$$

$$P\left(\left|\frac{X_1 + X_2 - \cdots X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

as $n \to \infty$, $P\left(\left|\frac{X_1 + - - X_n}{n} - \mu\right| \geq \varepsilon\right) \leq 0$.

$$\Rightarrow \lim_{n\to\infty} P\left(\left| \frac{x_1 + x_2 \cdots x_n}{n} - \mu \right| \geq \varepsilon \right) = 0.$$

$$\Rightarrow \lim_{n\to\infty} \frac{x_1 + x_2 \cdots + x_n}{n} = \mu.$$

Hence Weak LLN is proved.

Pseudocode

Let's say x_i is ~ Bernoulli $(\frac{1}{2})$ i.e

~~probability of a coin~~

$P(x_i = 1)$

$x_i = 1$ if we get heads ~~$P(x_i=1)$~~ $= 1/2$

$x_i = 0$ if we get tails $P(x_i = 0) = 1/2$

~~Sum Expectation~~ Sum = 0

for $i = 1 \cdots n$

Sum += x_i

~~print~~ Expectation = Sum / i

As $n \to \infty$ Expectation $\to 1/2$

(True Expectation)

DATE
PAGE

(d) We have to derive MAP solution for linear regression.

Assumption: gaussian priori distribution of weights

Let $y = w^T X$ be the best fit line.

Assuming ε to be error while making a prediction, we can write true value as

$$y_i = w^T x_i + \varepsilon$$

Since $\varepsilon \sim N(0, \sigma^2)$.

$$\Rightarrow \quad y_i \sim N(w^T x_i, \sigma^2)$$

$$\Rightarrow \quad P(y_i | \vec{x_i}, w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{w^T x_i - y_i}{\sigma}\right)^2}$$

Since we wish to derive MAP, formulae for MAP

$$P(w|D) = \frac{P(D|w)\,P(w)}{P(D)}$$

We want to maximise $P(w|D)$ i.e we want to choose w that is most likely given the Data D.

$$\bullet \quad P(w|D) = \underset{w}{\text{argmax}} \frac{P(D|w)\,P(w)}{P(D)}$$

(ignoring $P(D)$ since it remains constant)

$$= \underset{w}{\text{argmax}} \frac{P(y_1, y_2 \cdots y_n, x_1 \cdots x_n | w) P(w)}{P(y_1, y_2 \cdots y_n, x_1 \cdots x_n)}$$

Since (y_i, x_i) are iids.

$$= \underset{w}{\text{argmax}} \prod_{i=1}^{n} \frac{P(y_i x_i | w) P(w)}{P(D)}$$

(ignoring $P(D)$ since it remains constant)

$$= \underset{w}{\text{argmax}} \frac{\left(\prod_{i=1}^{n} P(y_i, x_i | w)\right) P(w)}{P(D)}$$

$$= \underset{w}{\text{argmax}} \left(\prod_{i=1}^{n} P(y_i, x_i | w)\right) P(w)$$

$$P(A, B | C) = \frac{P(A, B, C)}{P(C)}$$

$$= \frac{P(ABC)}{P(BC)} \times \frac{P(BC)}{P(C)} = P(A|BC) P(B|C)$$

Using above result

$$= \underset{w}{\text{argmax}} \left(\prod_{i=1}^{n} P(y_i | x_i, w) P(x_i | w)\right) P(w)$$

But x_i is undependent of w

$$= \underset{w}{\text{argmax}} \left(\prod_{i=1}^{n} P(y_i | x_i, w) P(x_i)\right) P(w)$$

DATE
PAGE

Since $P(x_i)$ is independent of w, $\therefore P(x_i)$ ~~is Cons~~ is constant for given Data.

Converting ~~Products~~ to log-likelihood,

$$\Rightarrow \underset{w}{\text{argmax}} \sum_{i=1}^{n} \left(\log P(y_i | x_i, w) + \log P(w) \right) \quad \text{— (2)}$$

Since it is given we have to assume gaussian prior distribution of weights.

$$P(w) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-(w^T w)/2\tau^2}$$

Pluggin values in (2)

$$\underset{w}{\text{argmax}} \left(\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w^T x_i - y_i)^2}{\sigma} \cdot \frac{1}{2}} \right) \right.$$

$$\left. + \log \left(\frac{1}{\sqrt{2\pi\tau^2}} e^{-(w^T w)/2\sigma\tau^2} \right) \right)$$

$$= \underset{w}{\text{argmax}} \left(\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2} \frac{(w^T x_i - y_i)^2}{\sigma^2} + \log \left(\frac{1}{\sqrt{2\pi\tau^2}} \right) \right.$$

$$\left. - \frac{w^T w}{2\tau^2} \right)$$

ignoring constan terms

$$\text{argmax} \left(-\frac{1}{2\sigma^2} (w^T x_i - y_i)^2 - \frac{1}{2\tau^2} w^T w \right)$$

↗ ignoring 2 as constant

DATE
PAGE

$$= \underset{w}{\operatorname{argmin}} \left(\frac{1}{\sigma^2} (w^T x_i - y_i)^2 + \frac{1}{\tau^2} (w^T w) \right)$$

Multiplying by σ^2 & dividing by n.

$$= \underset{w}{\operatorname{argmin}} \left(\frac{1}{n} (w^T x_i - y_i)^2 + \frac{\sigma^2}{n\tau^2} (w^T w) \right)$$

$$w = \underset{w}{\operatorname{argmin}} \left(\frac{1}{n} (w^T x_i - y_i)^2 + \lambda (w^T w) \right)$$

$\lambda = \sigma^2 / n\tau^2$

This is the MAP soln for linear Regression.