Hisc-Topics: 2

@pace

Probabilistic view of ML algo

[MLE ; MAP ; Bayesian ; Non-param

Ops:

chunk

↳ Q&A tab ← all questions

↓

"END: - - - - - "

Y/N; Interactivity → chat

Q1 Linear-reg → Squared-loss: $\sum_{i=1}^{n} \left[\overbrace{y_i - (w^T x_i + b)}^{e_i} \right]^2$

why not → abs. val loss $|e_i|$

→ e_i^4 or e_i^6 ?

(Q2) Regularization $\longrightarrow$ $\lambda \sum_{j=1}^{d} w_j^2$ L_2-reg

why not L_1-reg?

Linear-reg:- $\epsilon_i \sim$ Gaussian noise [why?]

(Q3) reg:

$x_q \rightarrow$ Model $\rightarrow \hat{y}_q$: point estimate

<loc; sqft ....>

<price>

$\hat{y}_q \rightarrow$ 1cr

C.I on $\hat{y}_q$ → Bayesian

→ Non-param

Practical

Case 1: [95L, 1.05Cr]

Case 2: [70L, 1.5Cr]

C.I $\longrightarrow$ CLT (pop mean & std-dev are finite)

$\searrow$ bootstrapping (non-param)

(Q4) post-read:

Logistic reg → log-loss?

Cond. prob

Bayes.

PDF of Gaussian dist

Probability & Eqns

Simplify it

Linear-reg

[squared-loss]

$\in \mathbb{R}$ $\in \mathbb{R}^d$ $\in \mathbb{R}^d$ Noise

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

[ignoring w_0]

A1: Linearity

A2: Noise is Gaussian

$$\epsilon \sim N(0, \sigma^2)$$

$$\hat{y}_i = w^T x_i$$

$$y_i - \hat{y}_i = \epsilon_i$$

Hack:

$$w^T x_i + w_0$$

$$\begin{bmatrix} w_1 & w_2 & \cdots & w_j & w_d \end{bmatrix} \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ij} \\ x_{id} \end{bmatrix} = w'^T x_i'$$

w_0

Noise

$$\epsilon \sim N(0, \sigma^2)$$

$$P(\epsilon_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{\epsilon_i^2}{2\sigma^2}\right\}$$

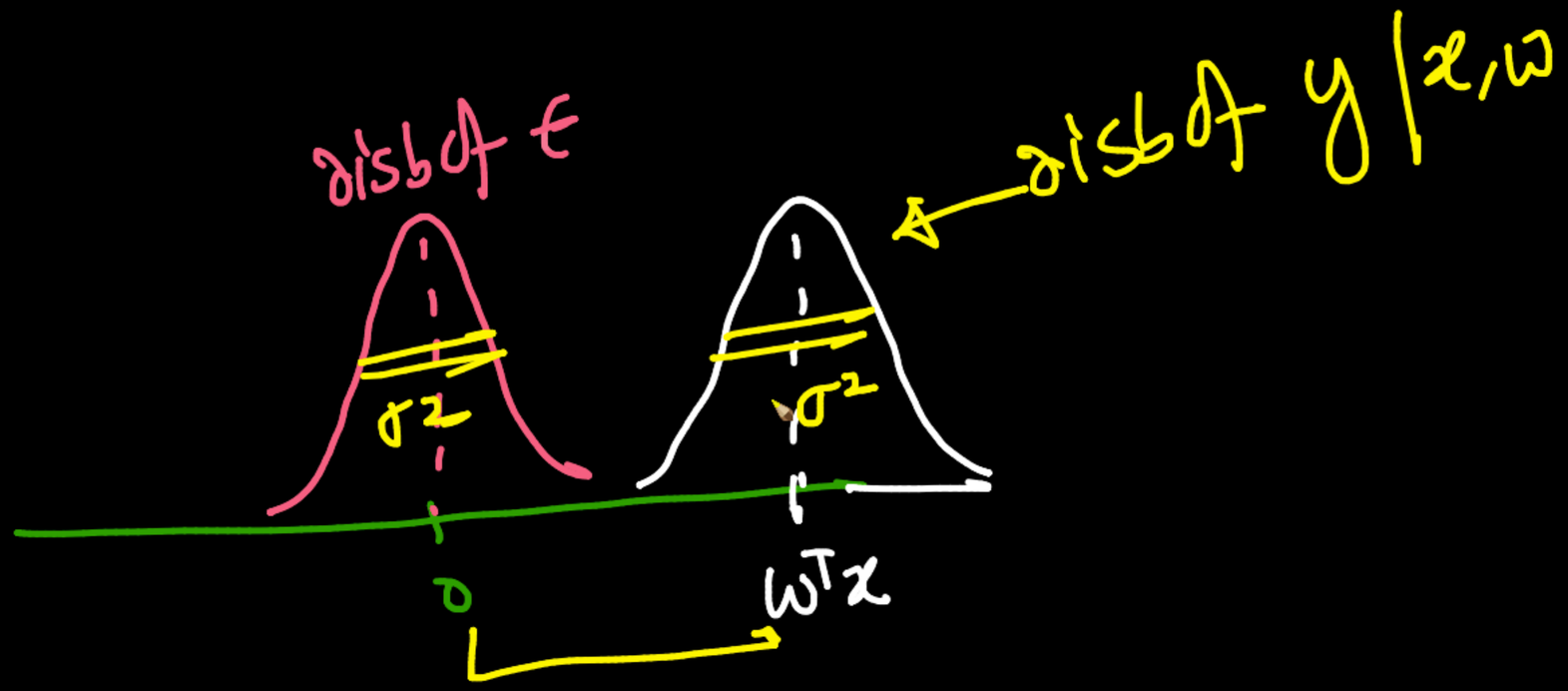
PDF of

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

$$y_i = \omega^T x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:22



$y \mid w, x$

$y = w^T x + \epsilon$

13 / 14

$$y \mid x, w \sim N\left(w^T x, \sigma^2\right)$$

$$p\left(y_i \mid x_i, w\right) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{\left(y_i - w^T x_i\right)^2}{2\sigma^2}\right\}$$

Sample

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

$\mathcal{D}_u$: all pairs of (x_i, y_i)

w, x_i

Likelihood:

$$\mathcal{L} = P(y_1, y_2 \ldots y_n \mid x_1, x_2, \ldots x_n, w)$$

high

params

Model: $w^T x_i$

Instabs

$$\mathcal{L}(\theta) = P(\mathcal{D} \mid \theta)$$

high

data

params

$\mathcal{D}$: dataset

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:38

intuition

$$P(\mathcal{D} \mid \theta)$$

→ Params

→ Model

$$X \rightarrow \mathcal{D} = \{\underbrace{x_1, x_2 \ldots x_n}_{\text{scalars}}\}$$

$$X \sim N(\underbrace{\mu, \sigma}_{\theta = [\mu, \sigma]}) \leftarrow \text{Model}$$

17 / 19

Q

Maximize

$$P(y_1, y_2, \ldots \mid \underbrace{x_1, x_2 \ldots}, \boxed{w})$$

$$\left[\begin{array}{l} y_i = w^T x_i \quad : \text{Model} \\ \epsilon_i \sim N(0, \sigma^2) \end{array}\right.$$

A1

A2

(x_1, y_1)

(x_2, y_2)

$\vdots$

$$\mathcal{L}(\overset{\theta}{w}) = P\left(y_1, y_2, \dots y_n \,\middle|\, x_1, x_2, \dots x_n, w\right)$$

$$\mathcal{D} = \left\{ (x_i, \overset{\text{actual}}{y_i}) \right\}_{i=1}^{n}$$

$$P(X, Y) = P(X)\,P(Y)$$

if X and Y are indep

$$\mathcal{L}(w) = P(y_1, y_2 \dots y_n \mid x_1, x_2 \dots x_n, w)$$

$$\mathcal{L}(w) = \prod_{i=1}^{n} P(y_i \mid x_i, w)$$

Obj: Find w which maximizes $\mathcal{L}(w)$ given my D

$$\log(a \cdot b) = \log(a) + \log(b)$$

$$\log(e^x) = x$$

$$e^{(\log(x))} = x$$

$$\max_{w} \mathcal{L}(w) = \max_{w} \prod_{i=1}^{n} \underbrace{P(y_i \mid x_i, w)}_{0 \text{ to } 1}$$

log-likelihood

$$\max_{w} \ell(w) = \max_{w} \sum_{i=1}^{n} \log \left[\underbrace{P(y_i \mid x_i, w)}_{} \right]$$

$$= \max_{w} \sum_{i=1}^{n} \log\left(\underbrace{\frac{1}{\sqrt{2\pi}\,\sigma}} \cdot \exp\left\{ -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right\} \right)$$

$$= \max_{w} \left[\underbrace{n \log \frac{1}{\sqrt{2\pi}\,\sigma}}_{\text{const}} - \sum_{i=1}^{n} \overset{\text{const}}{\boxed{\frac{1}{2\sigma^2}}} (y_i - w^T x_i)^2 \right]$$

log → simplify
→ monotonic

$$\min_{w} \cancel{\max} \sum_{i=1}^{n} + \left(y_i - w^T x_i\right)^2 \longrightarrow \text{sq-loss}$$

$$\max_{x} f(x) = \min_{x} -f(x)$$

Gradient-descent

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:52

SQ. loss

assumptions

→ Linearly $y_i = w^T x_i + \epsilon_i$

→ $\epsilon_i \sim N(0, \sigma^2)$

→ all (x_i, y_i) pairs are indep

25 / 26

$$\underbrace{\text{Maximum likelihood}}_{\mathcal{L}(\omega)} \quad \underline{\underline{\text{Est}}} \rightarrow \omega : \text{params}$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:54
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2 Update

Example 1-1

Suppose we have a random sample $X_1, X_2, \cdots, X_n$ where:

- $X_i = 0$ if a randomly selected student does not own a sports car, and
- $X_i = 1$ if a randomly selected student does own a sports car.

Assuming that the X_i are independent Bernoulli random variables with unknown parameter p, find the maximum likelihood estimator of p, the proportion of students who own a sports car.

$X_i \sim Ber(p)$

$p = P(X_i = 1)$

$1-p = P(X_i = 0)$

Answer

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$$

for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i; p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

27 / 27

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:55
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2 Update

Example 1-1

$\boxed{10}$ page 1

$n=10$

Suppose we have a random sample $X_1, X_2, \cdots, X_n$ where:

- $X_i = 0$ if a randomly selected student does not own a sports car, and
- $X_i = 1$ if a randomly selected student does own a sports car.

Assuming that the X_i are independent Bernoulli random variables with unknown parameter p, find the maximum likelihood estimator of p, the proportion of students who own a sports car.

point est:

$P(X_i = 1) = p$

$= \dfrac{\sum_{i=1}^{n} x_i}{n}$

Answer

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$$

for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i; p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

28 / 28

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:56
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2 Update

Example 1-1

Suppose we have a random sample $X_1, X_2, \cdots, X_n$ where:

- $X_i = 0$ if a randomly selected student does not own a sports car, and
- $X_i = 1$ if a randomly selected student does own a sports car.

Assuming that the X_i are independent Bernoulli random variables with unknown parameter p, find the maximum likelihood estimator of p, the proportion of students who own a sports car.

Answer

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$$

for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i; p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

29 / 29

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:56
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2
Update

Answer

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$$

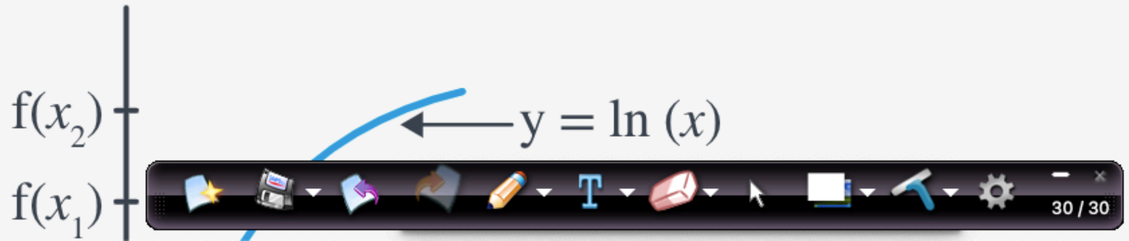
for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i;p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Now, in order to implement the method of maximum likelihood, we need to find the p that maximizes the likelihood $L(p)$. We need to put on our calculus hats now, since in order to maximize the function, we are going to need to differentiate the likelihood function with respect to p. In doing so, we'll use a "trick" that often makes the differentiation a bit easier. Note that the natural logarithm is an increasing function of x:



p

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:57
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2 Update

Answer

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$$

X_i are 0 or 1

↳ PMF

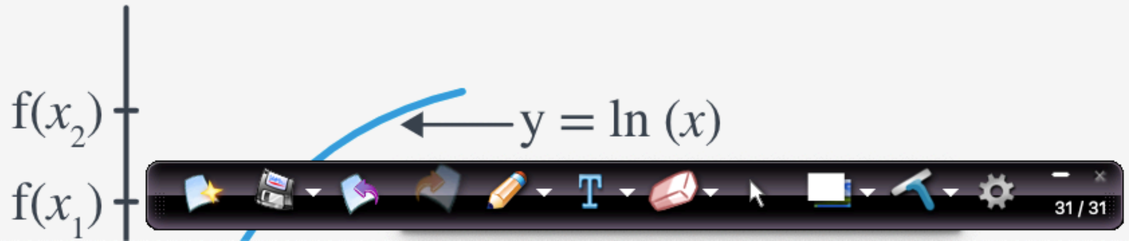
for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i;p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Now, in order to implement the method of maximum likelihood, we need to find the p that maximizes the likelihood $L(p)$. We need to put on our calculus hats now, since in order to maximize the function, we are going to need to differentiate the likelihood function with respect to p. In doing so, we'll use a "trick" that often makes the differentiation a bit easier. Note that the natural logarithm is an increasing function of x:



Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:57
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2 Update

Answer

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$$

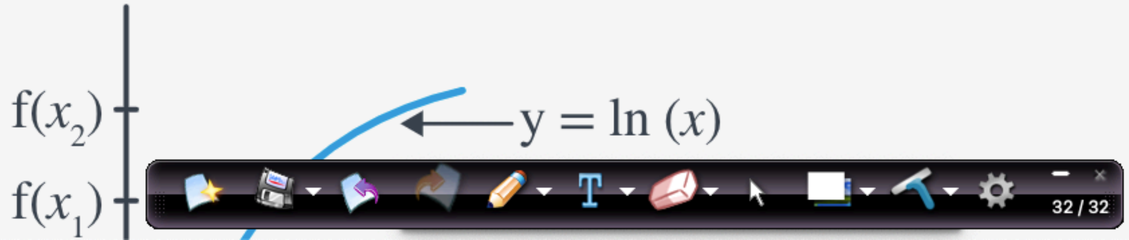
for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i;p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Now, in order to implement the method of maximum likelihood, we need to find the p that maximizes the likelihood $L(p)$. We need to put on our calculus hats now, since in order to maximize the function, we are going to need to differentiate the likelihood function with respect to p. In doing so, we'll use a "trick" that often makes the differentiation a bit easier. Note that the natural logarithm is an increasing function of x:



Model: Ber
params: p
$X_1, X_2 \ldots$
are indep

Chrome File Edit View History Bookmarks Profiles Tab Window Help 9:59
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2 Update

Answer

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$$

for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i;p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

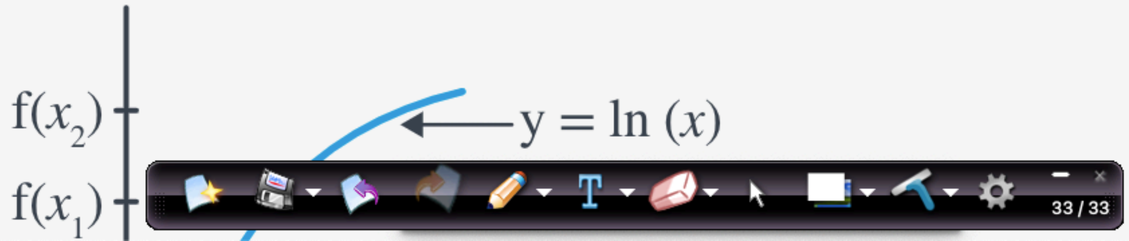
$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Now, in order to implement the method of maximum likelihood, we need to find the p that maximizes the likelihood $L(p)$. We need to put on our calculus hats now, since in order to maximize the function, we are going to need to differentiate the likelihood function with respect to p. In doing so, we'll use a "trick" that often makes the differentiation a bit easier. Note that the natural logarithm is an increasing function of x:

$x_1 \; x_2 \; \cdots \; x_n$

$P(x_1 x_2 \cdots x_n \mid p) = \mathcal{L}(p)$

$P(\mathcal{D} \mid p)$



33 / 33

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:00
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2 Update

throughout the remainder of this course, I will use either $\ln L(p)$ or $\log L(p)$ to denote the natural logarithm of the likelihood function.)

In this case, the natural logarithm of the likelihood function is:

$$\log L(p) = (\sum x_i)\log(p) + (n - \sum x_i)\log(1-p)$$

$$\log a^b = b \log a$$

Now, taking the derivative of the log likelihood, and setting to 0, we get:

$$\frac{\partial \log L(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p} \overset{SET}{\equiv} 0$$

Now, multiplying through by $p(1-p)$, we get:

$$(\sum x_i)(1-p) - (n - \sum x_i)p = 0$$

$$\mathcal{L}(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Upon distributing, we see that two of the resulting terms cancel each other out:

$$\sum x_i - \cancel{p\sum x_i} - np + \cancel{p\sum x_i} = 0$$

leaving us with:

$$\sum x_i - np = 0$$

Now, all we have to do is solve for p. In doing so, you'll want to make sure that you always put a hat ("^") on the parameter, in this case p, to indicate it is an estimate:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{}$$

34 / 34

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:01
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2 Update

$$\frac{\partial \log L(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p} \overset{SET}{\equiv} 0$$

Now, multiplying through by $p(1-p)$, we get:

$$(\sum x_i)(1-p) - (n - \sum x_i)p = 0$$

Upon distributing, we see that two of the resulting terms cancel each other out:

$$\sum x_i - \cancel{p\sum x_i} - np + \cancel{p\sum x_i} = 0$$

leaving us with:

$$\sum x_i - np = 0$$

Now, all we have to do is solve for p. In doing so, you'll want to make sure that you always put a hat ("^") on the parameter, in this case p, to indicate it is an estimate:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

or, alternatively, an estimator:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Oh, and we should technically verify that we indeed did obtain a maximum. We can do that by verifying that the second derivative of the log likelihood with respect to p is negative. It is, but you might want to do the work to convince yourself!

$$\left\{ \hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} \right\}$$

$x_i \sim Ber(p)$

$f(x_1 x_2 \ldots x_n | p)$

35 / 35

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:02
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2
Update

$$\frac{\partial \log L(p)}{\partial p} = \frac{\sum x_i}{p} - \frac{(n - \sum x_i)}{1-p} \overset{SET}{\equiv} 0$$

Now, multiplying through by $p(1-p)$, we get:

$$(\sum x_i)(1-p) - (n - \sum x_i)p = 0$$

Upon distributing, we see that two of the resulting terms cancel each other out:

$$\sum x_i - \cancel{p\sum x_i} - np + \cancel{p\sum x_i} = 0$$

leaving us with:

$$\sum x_i - np = 0$$

Now, all we have to do is solve for p. In doing so, you'll want to make sure that you always put a hat ("^") on the parameter, in this case p, to indicate it is an estimate:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

or, alternatively, an estimator:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Oh, and we should technically verify that we indeed did obtain a maximum. We can do that by verifying that the second derivative of the log likelihood with respect to p is negative. It is, but you might want to do the work to convince yourself!

36 / 36

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:02
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2 Update

Assuming that the X_i are independent Bernoulli random variables with unknown parameter p, find the maximum likelihood estimator of p, the proportion of students who own a sports car.

Answer

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$$

→ PMF of bernouli → Model

for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i;p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

$\boxed{p}$

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Now, in order to implement the method of maximum likelihood, we need to find the p that maximizes the likelihood $L(p)$. We need to put on our calculus hats now, since in order to maximize the function, we are going to need to differentiate the likelihood function with respect to p. In doing so, we'll use a "trick" that often makes the differentiation a bit easier. Note that the natural logarithm is an increasing function of x:

37 / 37

Chrome File Edit View History Bookmarks Profiles Tab Window Help
1.2 - Maximum Likelihood Estim
New Tab
online.stat.psu.edu/stat415/lesson/1/1.2
Update

Assuming that the X_i are independent Bernoulli random variables with unknown parameter p, find the maximum likelihood estimator of p, the proportion of students who own a sports car.

Answer

If the X_i are independent Bernoulli random variables with unknown parameter p, then the probability mass function of each X_i is:

$$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$$

for $x_i = 0$ or 1 and $0 < p < 1$. Therefore, the likelihood function $L(p)$ is, by definition:

$$L(p) = \prod_{i=1}^{n} f(x_i;p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \times \cdots \times p^{x_n}(1-p)^{1-x_n}$$

for $0 < p < 1$. Simplifying, by summing up the exponents, we get :

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Now, in order to implement the method of maximum likelihood, we need to find the p that maximizes the likelihood $L(p)$. We need to put on our calculus hats now, since in order to maximize the function, we are going to need to differentiate the likelihood function with respect to p. In doing so, we'll use a "trick" that often makes the differentiation a bit easier. Note that the natural logarithm is an increasing function of x:

38 / 39

Chrome File Edit View History Bookmarks Profiles Tab Window Help

10:07

40 / 40

Chrome File Edit View History Bookmarks Profiles Tab Window Help
1.2 - Maximum Likelihood Estim
New Tab
online.stat.psu.edu/stat415/lesson/1/1.2
Update

Example 1-2

Suppose the weights of randomly selected American female college students are normally distributed with unknown mean μ and standard deviation σ. A random sample of 10 American female college students yielded the following weights (in pounds):

x_1 x_2 x_3 . x_n

115 122 130 127 149 160 152 138 149 180

Model:

$x_i \sim N(\underbrace{\mu, \sigma^2}_{\theta})$

$\mathcal{L}(\theta) = P(x_1 x_2 \ldots x_n \mid \theta)$

Based on the definitions given above, identify the likelihood function and the maximum likelihood estimator of μ, the mean weight of all American female college students. Using the given sample, find a maximum likelihood estimate of μ as well.

Answer

The probability density function of X_i is:

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

for $-\infty < x < \infty$. The parameter space is $\Omega = \{(\mu, \sigma) : -\infty < \mu < \infty \text{ and } 0 < \sigma < \infty\}$. Therefore, (you might want to convince yourself that) the likelihood function is:

$$L(\mu, \sigma) =$$

40 / 40

Chrome File Edit View History Bookmarks Profiles Tab Window Help
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2
Update

pounds):

115 122 130 127 149 160 152 138 149 180

Based on the definitions given above, identify the likelihood function and the maximum likelihood estimator of μ, the mean weight of all American female college students. Using the given sample, find a maximum likelihood estimate of μ as well.

$[\mu, \sigma^2]$

$\mathcal{L}(\theta) = P(x_1 x_2 \cdots x_n | \theta)$

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

Model: N

Answer

The probability density function of X_i is:

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

for $-\infty < x < \infty$. The parameter space is $\Omega = \{(\mu, \sigma) : -\infty < \mu < \infty \text{ and } 0 < \sigma < \infty\}$. Therefore, (you might want to convince yourself that) the likelihood function is:

$$L(\mu, \sigma) = \sigma^{-n}(2\pi)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

for $-\infty < \mu < \infty$ and $0 < \sigma < \infty$. It can be shown (we'll do so in the next example!), upon maximizing the likelihood function with respect to μ, that the maximum likelihood estimator of μ is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Based on the given sample, a maximu

41 / 41

Chrome File Edit View History Bookmarks Profiles Tab Window Help
1.2 - Maximum Likelihood Estim New Tab
online.stat.psu.edu/stat415/lesson/1/1.2
Update

Answer

The probability density function of X_i is:

$$f(x_i;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

for $-\infty<x<\infty$. The parameter space is $\Omega=\{(\mu,\sigma):-\infty<\mu<\infty \text{ and } 0<\sigma<\infty\}$. Therefore, (you might want to convince yourself that) the likelihood function is:

$$L(\mu,\sigma)=\sigma^{-n}(2\pi)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

for $-\infty<\mu<\infty \text{ and } 0<\sigma<\infty$. It can be shown (we'll do so in the next example!), upon maximizing the likelihood function with respect to μ, that the maximum likelihood estimator of μ is:

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}$$

Based on the given sample, a maximum likelihood estimate of μ is:

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{10}(115+\cdots+180)=142.2$$

pounds. Note that the only difference between the formulas for the maximum likelihood estimator and the maximum likelihood estimate is that:

- the estimator is defined using capital letters (to denote that its value is random), and
- the estimate is defined using lowercase letters (to denote that its value is fixed and based on an obtained sample)

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:12

MLE:

$x_1 \dots x_n$ (obs)

(Model)

θ params of the model

$$\mathcal{L}(\theta) = P(x_1 \dots x_n \mid \theta)$$

Linear-reg

Where did the L_2-reg come from?

min SQ-loss

MLE: $P(y_1 \dots y_n \mid x_1 \dots x_n, w) \to P(\mathcal{D} \mid \theta)$

Linear-model
Gaussian noise
indep obs

min SQ-loss + reg

Maximum A posteriori (MAP) $\to \max_{w}$ $P(w \mid (x_i, y_i)_{i=1}^{n})$

$P(\theta \mid \mathcal{D})$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:21

Intuition

MLE: max the prob of observing the data given my params & my models under assumptions

$$\max_{\theta} P(D|\theta)$$

MAP: Find the most probable value of θ given my Dataset, model & assumptions

$$\max_{\theta} P(\theta|D)$$

45 / 46

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:23

$$P(\overset{A}{w} \mid \overset{B}{x, y}) = \frac{P(w, x, y)}{P(x, y)}$$

Matrix

$$\begin{bmatrix} \leftarrow x_i \rightarrow \end{bmatrix}$$

vector

$$\begin{bmatrix} \uparrow \\ y_i \\ \downarrow \end{bmatrix}$$

$$= \frac{P(y \mid x, w) \cdot P(w \mid x) \cancel{P(x)}}{P(y \mid x) \cancel{P(x)}}$$

46 / 46

$$\max_{\omega} P(\omega \mid \underbrace{X, Y}_{\theta}) = \max_{\omega} \frac{P(Y \mid X, \omega)\, P(\omega \mid X)}{P(Y \mid X)}$$

$$\omega^* = \arg\max_{\omega} \frac{P(Y \mid X, \omega) \cdot P(\omega)}{\cancel{\text{const}}}$$

$$w^* = \underset{w}{\operatorname{argmax}} \; \underbrace{P(Y \mid w, X)} \cdot P(w)$$

$$w^* = \underset{w}{\operatorname{argmax}} \; \underbrace{P(y_1 \ldots y_n \mid w, x_1 \ldots x_n)}_{\text{MLE}} \cdot \underbrace{P(w)}_{\text{explain}}$$

$$w^* = \underset{\underline{w}}{\arg\max} \; \overbrace{\log P(y_1 \ldots y_n \mid x_1 \ldots x_n, w)} + \underline{\log P(w)}$$

SQ. loss

(Same as MLE)

$\underline{L_2\text{-reg}}$

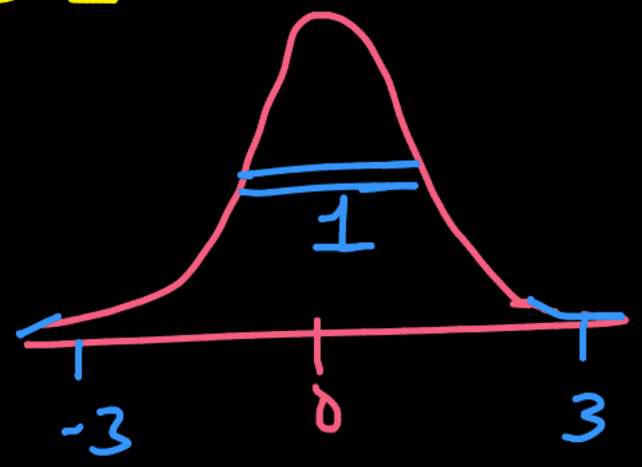
(later)

Let

i.i.d.

$$w_j \sim N(0, \sigma')$$

$\sigma' = 1$



$$P(w_j > 100) = ?$$

v.v. small

$$P(w_j < -100)$$

A: $w_j \sim N(0, \sigma^2)$

small, not too large

will not be v. large +ve or -ve value

{ regularization }

$$\max_{w} \log P(w_j) = \max \log \boxed{\frac{1}{\sqrt{2\pi}\,\sigma}} \exp\left\{ \frac{-w_j^2}{2\sigma^2} \right\}$$

$$= \max \; \frac{-w_j^2}{2\sigma^2}$$

$$= \min \; \frac{+w_j^2}{2\sigma^2}$$

$$\min_{w} \left[\underbrace{\sum_{i=1}^{n} (y_i - w^T x_i)^2}_{\prod_{i=1}^{n} P(y_i \mid x_i, w)} + \lambda \underbrace{\sum_{j=1}^{d} w_j^2}_{P(w)} \right]$$

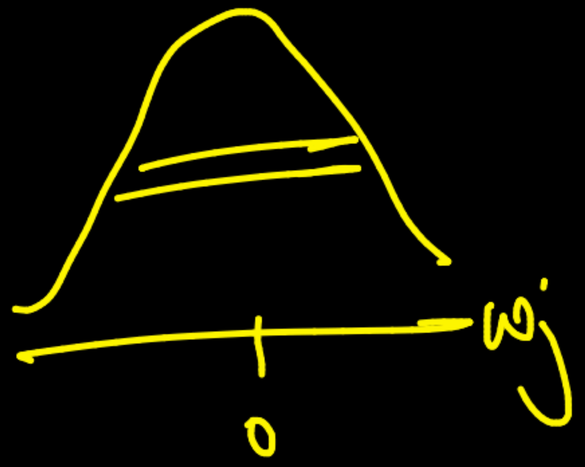
GD

$$\lambda = \frac{1}{2\sigma'^2} : \text{prob}$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:44

$$\lambda = \frac{1}{2\sigma'^2}$$

$\sigma' \uparrow$: $w_j \sim N(0, \sigma'^2)$



as $\sigma' \uparrow$; $\lambda \downarrow$

letting w_j's to be larger → less regularization

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:44

$$\begin{cases} \text{MLE} \rightarrow \overset{\text{min}}{\text{SQ}} \cdot \text{loss} \\ \text{MAP} \rightarrow \text{SQ} \cdot \text{loss} + \cancel{\text{reg}} \end{cases}$$

Math thinking

Geometry

easier

55 / 55

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:46

Max-A posterior

$$P(\theta \mid D) = P(\hat{w} \mid X, Y)$$

: Bayes rule

posterior prob.

56 / 56

Logistic reg

log-odds

Model: Sigmoid

Params: w

y_i : Bernouli

MLE

log-loss

MAP $\rightarrow$ min log-loss + L_2-reg.

Chrome File Edit View History Bookmarks Profiles Tab Window Help
1.2 - Maximum Likelihood Estim | A Probabilistic Interpretation o | Laplace distribution - Wikipedi | New Tab
bjlkeng.github.io/posts/probabilistic-interpretation-of-regularization/

Selecting Priors for Linear Regression

The main idea is to select Bayesian priors on the coefficients of linear regression that get us to L1 and L2 regularization (Equation 2 and 3). Let's see how this works.

Normally Distributed Priors

We'll start with our good old friend the normal distribution and place a zero-mean normally distributed prior on *each* β_i value, all with identical variance τ^2. From Equation 6 and filling in the likelihood function from Equation 4 and our prior:

$P(w_i)$

$$
\begin{aligned}
&\arg\max_{\beta}\left[\log\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y_i-(\beta_0+\beta_1x_{i,1}+\ldots+\beta_px_{i,p}))^2}{2\sigma^2}}+\log\prod_{j=0}^{p}\frac{1}{\tau\sqrt{2\pi}}e^{-\frac{\beta_j^2}{2\tau^2}}\right]\\
&=\arg\max_{\beta}\left[-\sum_{i=1}^{n}\frac{(y_i-(\beta_0+\beta_1x_{i,1}+\ldots+\beta_px_{i,p}))^2}{2\sigma^2}-\sum_{j=0}^{p}\frac{\beta_j^2}{2\tau^2}\right]\\
&=\arg\min_{\beta}\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_{i,1}+\ldots+\beta_px_{i,p}))^2+\frac{\sigma^2}{\tau^2}\sum_{j=0}^{p}\beta_j^2\right]\\
&=\arg\min_{\beta}\left[\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_{i,1}+\ldots+\beta_px_{i,p}))^2+\lambda\sum_{j=0}^{p}\beta_j^2\right] \qquad (8)
\end{aligned}
$$

Notice that we dropped many of the constants (with respect to β) and factored a bit to simplify the expression. You can see this is the same expression as Equation 3 (L2 Regularization) with $\lambda = \sigma^2/\tau^2$ (remember σ is assumed to be constant in ordinary linear regression, and we get to pick τ for our prior). We can adjust the amount of regularization we want by cha[illegible] the priors carry on the coefficients (β). If we have a very small variance (large λ) then the coefficients will be very

58 / 59

Q

$$\begin{cases} y_i \sim N(w^T x_i, \sigma^2) \\ w_j \sim \text{Laplace-dist}(\mu, b) \end{cases}$$

MAP: min sq.loss + L1-reg

Chrome File Edit View History Bookmarks Profiles Tab Window Help 10:56

Q

$w_j \sim N(0, \sigma'^2)$

$w_1, w_2 \ldots w_d$

equivalent

$w \in \mathbb{R}^d$

$w \sim N(\vec{0}, \sigma'^2 I)$

d-dim Gaussian dist (earlier)

$w^T w = \sum_{j=1}^{d} w_j'^2$

60 / 60

Intuition

point estimates

MLE < MAP < Bayesian Inf ...

↳ study in more in DL

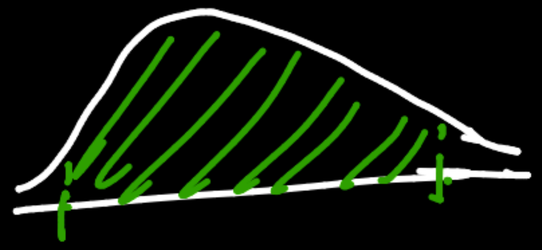
1

c.l on y_q 2

$$P(y_q \mid x_q, X, Y)$$

query

D_{Train}



$y_q \mid x_q, X, Y$

$$\left[p\left(y_q \mid \overbrace{X, Y}^{D_{TR}}, x_q\right) = \int_{\theta} p\left(y_q \mid x_q, \theta\right) \cdot p\left(\theta \mid X, Y\right) d\theta \right.$$

$\theta \to w$

monte-carlo (MCMC)

↳ Comp - very expensive

② distb of $\theta = w$

$w_j \rightarrow$ feature importance

95% C.I of $w_1 \rightarrow [0.1 \quad 0.8]$

95% C.I of $w_2 \rightarrow [-2.1 \quad +2.1]$

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

Sample

$$x_q \longrightarrow \overset{95\%}{\text{C.I on}}\ y_q$$

HINT: Bootstrapping (BS)

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:14

Idea 1:

$$\mathcal{D}^n \xrightarrow[\text{with repl}]{\text{bootstrapping}} \mathcal{D}_1^{m \le n} \xrightarrow{\text{Train}} W^{(1)}$$

$$\mathcal{D}_2^{m \le n} \xrightarrow{\text{Train}} W^{(2)}$$

$$\vdots$$

$$\mathcal{D}_r^{m \le n} \xrightarrow{\text{Train}} W^{(r)}$$

eval. r. models

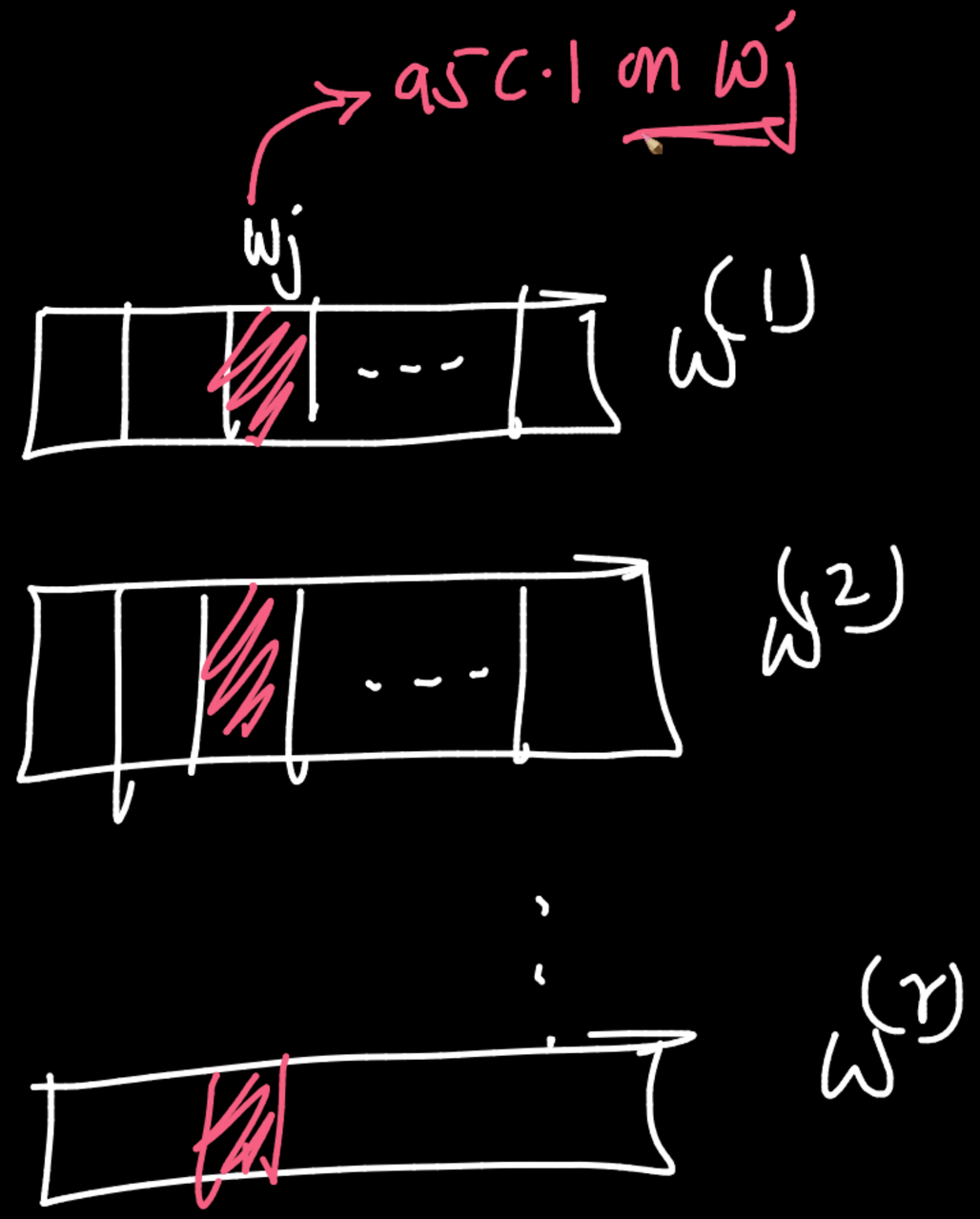
$$
\begin{tikzcd}
 & w^{(1)} \arrow[r] & y_q^{1} \\
x_q \arrow[ur] \arrow[r] & w^{(2)} \arrow[r] & y_q^{2} \\
 & \vdots & \vdots \\
 & w^{(r)} \arrow[r] & y_q^{r}
\end{tikzcd}
$$

assoc.

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:13

C.I on w_j's

Same method :-



67 / 67

Random-forest → bootstrapping

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:16

Idea 2:

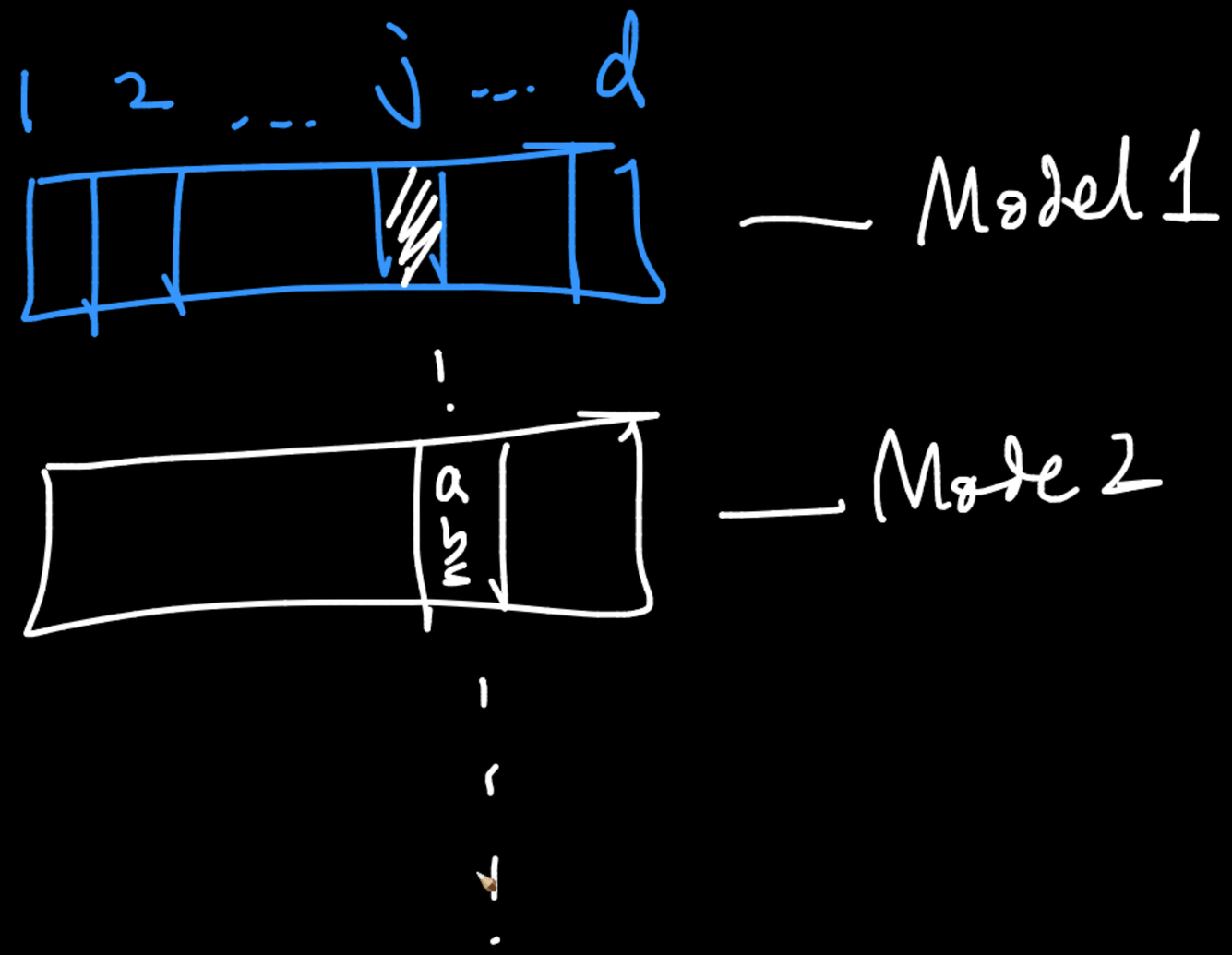
$$\mathcal{D}^{n} \xrightarrow[\text{Col-Sampl}]{\text{row \&}} \mathcal{D}_1^{m} \longrightarrow \text{Model}^{1} : w^{(1)}$$

$$\mathcal{D}_2^{m} \longrightarrow \text{Model}^{2} : w^{(2)}$$

$$\vdots$$

$$\mathcal{D}_r^{m} \longrightarrow \text{Model}^{r} : w^{(r)}$$

69 / 70


1 2 ... j ... d
Model 1
a
b
c
Mode 2

SQ

$$\text{Loss} = \|Xw - Y\|^2$$

next-slide

Linear-reg

[Normal-eqns]

Matrix-Mult

no-GD

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:38

$$\text{sq-loss} = \| X \cdot w - Y \|^2$$

$$\underbrace{\begin{bmatrix} \leftarrow x_1 \rightarrow \\ \leftarrow x_2 \rightarrow \\ \vdots \\ \leftarrow x_n \rightarrow \end{bmatrix}}_{n \times d} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}}_{d \times 1} \qquad \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{n \times 1}$$

$$\text{sq-loss} = \|(Xw - Y)\|^2$$

$$= (Xw - Y)^T (Xw - Y)_{n\times 1} \quad : \text{scalar}$$

$$\text{sq-loss} = Y^TY - Y^TXw - w^TX^TY + w^TX^TXw$$

$$Y^TY \neq YY^T$$

scalar matrix

$$\left[\frac{\partial\, \text{sq-loss}}{\partial w} = 0\right.$$

$$\Rightarrow \left[\; Y_{n\times 1}\right.$$

$$Y^T_{1\times n}\, Y_{n\times 1} \quad : \text{scalar}$$

$$Y_{n\times 1}\, Y^T_{1\times n} = n\times n\text{-Mat}$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:43
73 / 81

$$(Xw)^T = w^T X^T$$

$$0 = \frac{\partial \left(y^T y - y^T X w - w^T X^T y + w^T X^T X w \right)}{\partial w}$$

[Next class [Q: Matrix & vec. derivatives]]

$$= 0 - X^T y - X^T y + 2 X^T X w$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:30

Recall

Lr. sys. of eqns

$$Ax = b$$

$$x = \underline{\underline{A^{-1} b}}$$

76 / 77

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:31

Lesson:

Ex. reg using Normal eqns & no GD

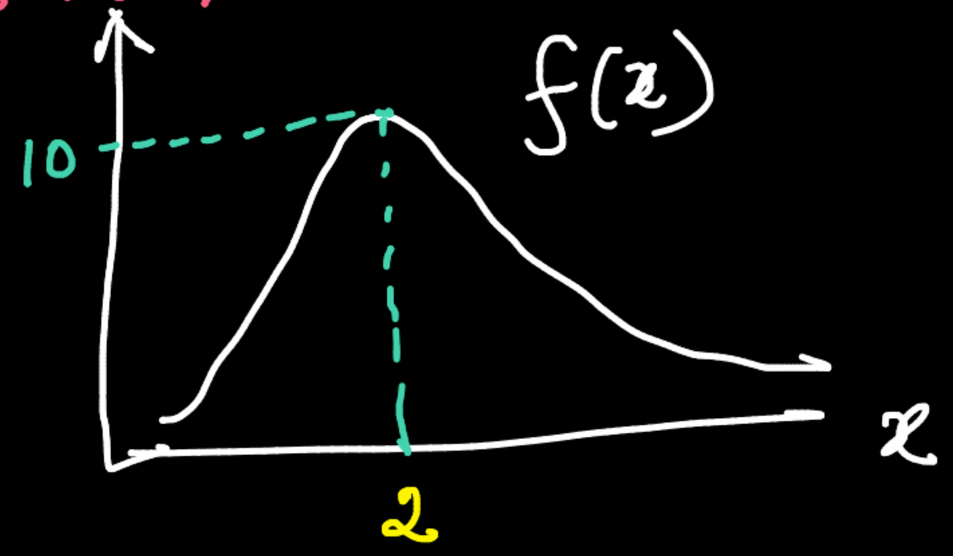
↓

pure matrix ops

77 / 77

Q

argmax & Max



$$\max_{x} f(x) = 10$$

ans

$$\underset{x}{\arg\max}\ f(x) = 2$$

Chrome File Edit View History Bookmarks Profiles Tab Window Help 11:36

Q

Bayesian Inference "θ"

non-param methods (BS)

ML

DL

y_q

MLE < MAP

79 / 79

Chrome File Edit View History Bookmarks Profiles Tab Window Help

11:36

80 / 80

Chrome File Edit View History Bookmarks Profiles Tab Window Help
1.2 - Maximum Likelihood Estim | A Probabilistic Interpretation o | Laplace distribution - Wikipedi | New Tab
bjlkeng.github.io/posts/probabilistic-interpretation-of-regularization/
Update

- Machine Learning: A Probabilistic Perspective, Kevin P. Murphy

[1] One philosophical counterpoint is that we should "let the data speak for itself". Although superficially satisfying, it is almost always the case where you inject "prior" knowledge into interpreting the data. For example, selecting a linear regression model already adds some prior knowledge or intuition to the data. In the same way, if we have some vague idea that the mean of the data should be close to zero, why not add that information into the problem? If there's enough data and we've coded things right, then the prior isn't that impactful anyways.

[2] As you might have guessed, I'm fall into the Bayesian camp. Although I would have to say that I'm much more of a pragmatist above all else. I'll use whatever works, frequentist, Bayesian, no theoretical basis, doesn't really matter as long as I can solve the desired problem in a reasonable manner. It just so happens Bayesian methods produce reasonable estimates very often.

[3] Well that's not exactly true. As a modeler, we can pick a prior that does depend on the data, although that's a bit of "double dipping". These methods are generally known as empirical Bayes methods.

[4] I got this example from Machine Learning: A Probabilistic Perspective. It has great explanations on both L1 and L2 regularization as long as you have some moderate fluency in probability. I highly recommend this textbook if you want to dig into the meat of ML techniques. It's very math (probability) heavy but really provides good high level explanations that help with intuition.

Bayesian probability regularization

Previous post Next post

Contents © 2022 Brian Keng - Powered by
81 / 81