Homework 1 by 161220097 戚赟

2019 年 3 月 16 日

1 [20pts] Basic Probability and Statistics

The probability distribution of random variable X follows:

$$f_X(x) = \begin{cases} \frac{1}{2} & 0 < x < 1; \\ \frac{1}{6} & 2 < x < 5; \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

(1) [5pts] Please give the cumulative distribution function $F_X(x)$ for X;

(2) [5pts] Define random variable Y as $Y = 1/(X^2)$, please give the probability density function $f_Y(y)$ for Y;

(3) [10pts] For some random non-negative random variable Z, please prove the following two formulations are equivalent:

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} z f(z) \mathrm{d}z, \tag{2}$$

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} \Pr[Z \geq z] \mathrm{d}z, \tag{3}$$

Meantime, please calculate the expectation of random variable X and Y by these two expectation formulations to verify your proof.

Solution:
(1)The cumulative distribution function $F_X(x)$ fo X follows:

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{2} & 0 < x \leq 1; \\ \frac{1}{2} & 1 < x \leq 2; \\ \frac{1}{6} + \frac{1}{6}x & 2 < x \leq 5; \\ 1 & 5 < x; \end{cases} \tag{4}$$

(2)The the probability density function $f_Y(y)$ for Y follows:

$$f_Y(y) = \begin{cases} 0 & y \leq \frac{1}{25}; \\ \frac{1}{12}y^{-\frac{3}{2}} & \frac{1}{25} < y \leq \frac{1}{4}; \\ 0 & \frac{1}{4} < y \leq 1; \\ \frac{1}{4}y^{-\frac{3}{2}} & 1 < y; \end{cases} \tag{5}$$

(3)

$$\because \mathbb{E}[Z] = \int_{z=0}^{\infty} \Pr[Z \geq z]\mathrm{d}z = \int_{z=0}^{\infty}\int_{x=z}^{\infty} f_Z(z)\mathrm{d}x\mathrm{d}z \tag{6}$$

$$After\ change\ the\ order\ of\ integration \tag{7}$$

$$\mathbb{E}[Z] = \int_{z=0}^{\infty}\int_{x=z}^{\infty} f_Z(x)\mathrm{d}x\mathrm{d}z = \int_{x=0}^{\infty}\int_{z=0}^{x} f_Z(x)\mathrm{d}z\mathrm{d}x \tag{8}$$

$$= \int_{x=0}^{\infty} x f_Z(x)\mathrm{d}x \tag{9}$$

$$= \int_{z=0}^{\infty} z f(z)\mathrm{d}z \tag{10}$$

By equation(2):

$$\mathbb{E}[X] = 2 \tag{11}$$

$$\mathbb{E}[Y]\ does\ not\ exist \tag{12}$$

By equation(3):

$$\mathbb{E}[X] = 2 \tag{13}$$

$$\mathbb{E}[Y]\ does\ not\ exist \tag{14}$$

So the proof is verified

2 [20pts] Strong Convexity

Let $D \in \mathbb{R}^2$ be a finite set. Define a function $E : \mathbb{R}^3 \to \mathbb{R}$ by

$$E(a, b, c) = \sum_{x \in \mathcal{D}} (ax_1^2 + bx_1 + c - x_2)^2. \tag{15}$$

(1) [10pts] Show that E is convex.

(2) [10pts] Does there exist a set D such that E is strongly convex? Proof or a counterexample.

Solution:
(1)To prove E is convex, according to the definition of convex.$\forall x_1, x_2, \forall t \in [0,1], f(tx_1+(1-t)x_2) \leq tf(x_1)+(1-t)f(x_2)$.Set $g(a,b,c) = \sqrt{E(a,b,c)}, \forall x_1, x_2, \forall t \in [0,1]$.

$$\because tE(x_1) + (1-t)E(x_2) - E(tx_1 + (1-t)x_2) \tag{16}$$
$$= tg(x_1)^2 + (1-t)g(x_2)^2 - (tg(x_1) + (1-t)g(x_2))^2 \tag{17}$$
$$= t(1-t)(g(x_1) - g(x_2))^2 \geq 0 \tag{18}$$
$$\therefore tE(x_1) + (1-t)E(x_2) \geq E(tx_1 + (1-t)x_2) \tag{19}$$
$$\therefore E \ is \ convex \tag{20}$$

Another way is to proof Hessian Matrix is positive semi-definite, And

$$H(E) = \sum_i E_i, E_i = \begin{bmatrix} x_i^4 & x_i^3 & x_i^2 \\ x_i^3 & x_i^2 & x_i \\ x_i^2 & x_i & 1 \end{bmatrix} \tag{21}$$

Because E_i is positive semi-definite,so $H(E)$ is positive semi-definite and E is a convex function.

(2) If the function $f$is twice continuously differentiable, then it is strongly convex with parameter m if and only if $\nabla^2 f(x) \succeq mI$ for all x in the domain, where I is the identity and $\nabla^2 f$ is the Hessian matrix, and the *inequality*$\succeq$ means that$\nabla^2 f(x) - mI$is positive semi-definite. **[1]From Wikipedia**

So according to this theorem,calculate the Hessiaan matrix first.

$$H(E) = \begin{bmatrix} \frac{\partial E}{\partial a \partial a} & \frac{\partial E}{\partial a \partial b} & \frac{\partial E}{\partial a \partial c} \\ \frac{\partial E}{\partial b \partial a} & \frac{\partial E}{\partial b \partial b} & \frac{\partial E}{\partial b \partial c} \\ \frac{\partial E}{\partial c \partial a} & \frac{\partial E}{\partial c \partial b} & \frac{\partial E}{\partial c \partial c} \end{bmatrix} \tag{22}$$

So

$$H(E) = \begin{bmatrix} \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^2 & \sum_i x_i & \sum_i 1 \end{bmatrix} \tag{23}$$

And set $D = \{(1,0),(2,0),(3,0)\}, m = 0.01$,the $\nabla^2 f(x) - mIis$

$$H(E) = \begin{bmatrix} 97.99 & 36 & 14 \\ 36 & 13.99 & 6 \\ 14 & 6 & 2.99 \end{bmatrix} \tag{24}$$

Because H(E) is a positive semi-definite,so accroding to the therom, E is strongly convex.

3 [20pts] Transition Probability Matrix

Suppose x_k is the fraction of NJU students who prefer course A at year k. The remaining fraction $y_k = 1 - x_k$ prefers course B.

At year $k+1$, $\frac{1}{5}$ of those who prefer course A change their mind. Also at the same year, $\frac{1}{10}$ of those who prefer course B change their mind (possibly after taking the problem 3 last year).

Create the matrix P to give $[x_{k+1} \quad y_{k+1}]^\top = P[x_k \quad y_k]^\top$ and find the limit of $P^k[1 \quad 0]^\top$ as $k \to \infty$.

Solution

It's easy to get matrix P accroding to the description.

$$P = \begin{bmatrix} \frac{4}{5} & \frac{1}{10} \\ \frac{1}{5} & \frac{9}{10} \end{bmatrix} \tag{25}$$

The matrix eigenvalues of P is 1 and $\frac{7}{10}$. The corresponding vector is

$\alpha_1 = [1 \quad 2]^\top$ and $\alpha_2 = [1 \quad -1]^\top$.Because $[1 \quad 2]^\top = \frac{1}{3}(\alpha_1 + 2\alpha_2)$.So

$$\lim_{k\to+\infty} P^k[1 \quad 0]^\top = \lim_{k\to+\infty} \frac{1}{3}(1^k\alpha_1 + 2(\frac{7}{10})^k\alpha_2) = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \end{bmatrix}^\top \tag{26}$$

4 [20pts] Hypothesis Testing

Yesterday, a student was caught by the teacher when tossing a coin in class. The teacher is very nice and did not want to make things difficult. S(he) wished the student to determine *if the coin is biased for heads* with $\alpha = 0.05$.

Also, according to the student' s desk mate, the coin was tossed for 50 times and it got 35 heads.

(1) [10pts] Show all calculate and rules (hint: using z-test).

(2) [10pts] Calculate the p-value and interpret it.

Solution

(1)By using Z-test, we assume $H_0 = 0.5, H_1 > 0.5$.So

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\frac{35}{50} - \frac{1}{2}}{\sqrt{p(1-p)}/\sqrt{n}} = \frac{\frac{1}{5}}{\frac{1}{2}/\sqrt{50}} = 2\sqrt{2} = 2.8284 \tag{27}$$

(2)According to the Z-table, $-z \leq -2.8284, P(Z \geq z) = 0.0023 < \alpha = 0.05$, that means we refuse the H_0 hypothesis and the coin is biased for heads with $\alpha = 0.05$. The meanings of P is to decide if null hypothesis is acceptable. Because $P < 0.05$, the null hypothesis is not the same as reality and the coins is biased for heads.

5 [20pts] Performance Measures

We have a set of samples that we wish to classify in one of two classes and a ground truth class of each sample (denoted as 0 and 1). For each example a classifier gives us a score (score closer to 0 means class 0, score closer to 1 means class 1). Below are the results of two classifiers (C_1 and

C_2) for 8 samples,their ground truth values (y) and the score values for both classifiers (y_{C_1} and y_{C_2}).

y	1	0	1	1	1	0	0	0
y_{C_1}	0.5	0.3	0.6	0.22	0.4	0.51	0.2	0.33
y_{C_2}	0.04	0.1	0.68	0.22	0.4	0.11	0.8	0.53

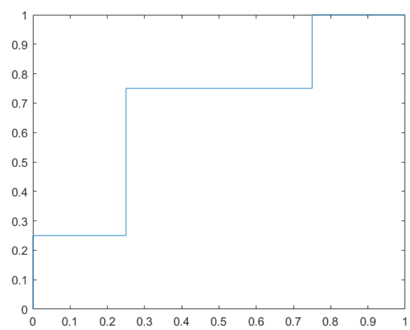
(1) [8pts] For the example above calculate and draw the ROC curves for classifier C_1 and C_2. Also calculate the area under the curve (AUC) for both classifiers.

(2) [8pts] For the classifier C_1 select a decision threshold $th_1 = 0.33$ which means that C_1 classifies a sample as class 1, if its score $y_{C_1} > th_1$, otherwise it classifies it as class 0. Use it to calculate the confusion matrix and the F_1 score. Do the same thing for the classifier C_2 using a threshold value $th_2 = 0.1$.

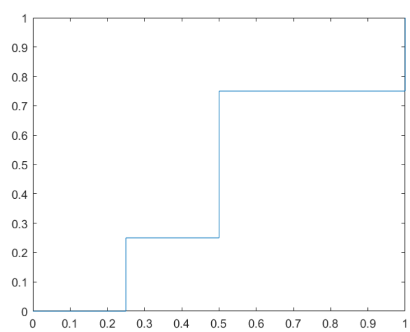
(3) [4pts] Prove Eq.(2.22) in Page 35. (AUC $= 1 - \ell_{rank}$).

Solution

(1)



(a) ROC of y_{C_1}, x means FPR



(b) ROC of y_{C_2}, x means FPR

图 1: pics

$$AUC\ of\ C_1 = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)(y_i+y_{i+1}) = \frac{11}{16} \tag{28}$$

$$AUC\ of\ C_2 = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)(y_i+y_{i+1}) = \frac{7}{16} \tag{29}$$

(2) The confusion matrix of C_1 and C_2 is:

表 1:

Reality	Production	
	positive	negative
positive	3	1
negative	1	3

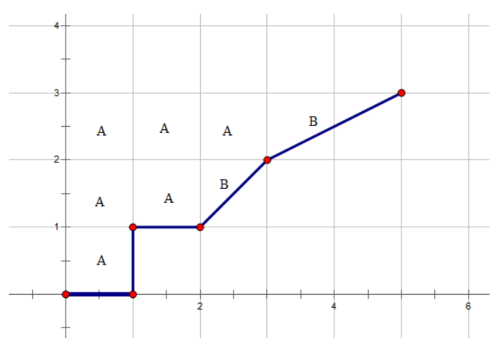
表 2: 2

Reality	Production	
	positive	negative
positive	3	1
negative	3	1

$F_{C_1} = \frac{2PR}{P+R} = \frac{3}{4}$ when $th1 = 0.33$.
$F_{C_2} = \frac{2PR}{P+R} = \frac{3}{5}$ when the $th2 = 0.1$.

(3) We just need to prove the ℓ_{rank} is the area above the ROC. Because the x_{max} and y_{max} is 1. So set $x = x \times m^-, y = y \times m^+$. And x of one point means the curruent numbers of false positive. So count of x reperesents the area of unique y as each rectangle's width and height is 1.
So after one step of increasing th , set $a = \triangle x,\ b = \triangle y$.
case one: $a = 0$,the area increase $1 \times x = x$,means count of x whose value is less than y's, just the $\sum x^- \in D^-\ \mathbb{I}(f(y) < f(x^-))$ (The area A in (a) for the same y).
case two: $b = 0$,the area increase 0.
case three: $a \neq 0,\ b \neq 0$,the area increase $x + \frac{ab}{2}$,the $\frac{ab}{2}$ means the count of $f(x^+) = f(y^-)$ (like the area B). So the increase is $\sum x^- \in D^-\ (\mathbb{I}(f(y) < f(x^-)) + \frac{1}{2}\ \mathbb{I}(f(y) = f(x^-)))$.

To conclue, the ℓ_{rank} is $\frac{1}{m^+m^-} \sum x^+ \in D^+ \sum x^- \in D^-\ (\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2}\ \mathbb{I}(f(x^+) = f(x^-)))$.And AUC $= 1 - \ell_{rank}$.



(a) Example

6 [Bonus 10pts]Expected Prediction Error

For least squares linear regression problem, we assume our linear model as:

$$y = x^T \beta + \epsilon, \tag{30}$$

where ϵ is noise and follows $\epsilon \sim N(0, \sigma^2)$. Note the instance feature of training data $\mathcal{D}$ as $\boldsymbol{X} \in \mathbb{R}^{p \times m}$ and note the label as $\boldsymbol{Y} \in \mathbb{R}^n$, where n is the number of instance and p is the feature dimension. So the estimation of model parameter is:

$$\hat{\beta} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{Y}. \tag{31}$$

For some given test instance x_0, please proof the expected prediction error $\mathbf{EPE}(x_0)$ follows:

$$\mathbf{EPE}(x_0) = \sigma^2 + \mathbb{E}_{\mathcal{D}}[x_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}x_0\sigma^2]. \tag{32}$$

Please give the steps and details of your proof.(Hint: $\mathbf{EPE}(x_0) = \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathcal{D}}[(y_0 - \hat{y}_0)^2]$, you can also refer to the proof progress of variance-bias decomposition on the page 45 of our reference book)

Solution:

$$\hat{y}_0 = x_0^T\hat{\beta} = x_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{Y} \tag{33}$$

$$= x_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}(\boldsymbol{X^T\beta} + \boldsymbol{\epsilon}) \tag{34}$$

$$= x_0^T\beta + x_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon \tag{35}$$

So $\mathbb{E}_{\mathcal{D}}[\hat{y}_0] = \mathbb{E}_{\mathcal{D}}[x_0^T\beta + x_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon]$
Because $\epsilon \sim N(0, \sigma^2)$, $\mathbb{E}_{\mathcal{D}}[x_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\epsilon] = 0$ and $\mathbb{E}_{\mathcal{D}}[\hat{y}_0] = x_0^T\beta$.

$$\mathbb{E}_{\mathcal{D}}[(y_0-\hat{y}_0)^2] = \mathbb{E}_{\mathcal{D}}[(y_0-\mathbb{E}\hat{y}_0+\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)^2] \tag{36}$$
$$= \mathbb{E}_{\mathcal{D}}[(y_0-\mathbb{E}_{\mathcal{D}}\hat{y}_0)^2+(\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)^2+2(y_0-\mathbb{E}_{\mathcal{D}}\hat{y}_0)(\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)] \tag{37}$$
$$= \mathbb{E}_{\mathcal{D}}[(y_0-\mathbb{E}_{\mathcal{D}}\hat{y}_0)^2+(\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)^2] \tag{38}$$
$$= \mathbb{E}_{\mathcal{D}}[(y_0-\mathbb{E}_{\mathcal{D}}\hat{y}_0)^2]+\mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)^2] \tag{39}$$
$$= \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)^2]+\mathbb{E}_{\mathcal{D}}[(y_0-x_0^T\beta+x_0^T\beta-\mathbb{E}_{\mathcal{D}}\hat{y}_0)^2] \tag{40}$$
$$= \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)^2]+\mathbb{E}_{\mathcal{D}}[(y_0-x_0^T\beta)^2+(x_0^T\beta-\mathbb{E}_{\mathcal{D}}\hat{y}_0)^2+2(y_0-x_0^T\beta)(x_0^T\beta-\mathbb{E}_{\mathcal{D}}\hat{y}_0)] \tag{41}$$
$$= \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)^2]+\mathbb{E}_{\mathcal{D}}[(y_0-x_0^T\beta)^2]+\mathbb{E}_{\mathcal{D}}[(x_0^T\beta-\mathbb{E}_{\mathcal{D}}\hat{y}_0)^2] \tag{42}$$
$$= \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)^2]+\mathbb{E}_{\mathcal{D}}[(y_0-x_0^T\beta)^2] \tag{43}$$
$$\tag{44}$$

So $\mathbf{EPE}(x_0) = \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathcal{D}}[(y_0-\hat{y}_0)^2] = var(\hat{y_0})+\sigma^2$

$$var(\hat{y_0}) = \mathbb{E}_{\mathcal{D}}[(\mathbb{E}_{\mathcal{D}}\hat{y}_0-\hat{y}_0)^2] \tag{45}$$
$$= \mathbb{E}_{\mathcal{D}}[(x_0^T\beta+x_0^T(XX^T)^{-1}X\epsilon-x_0^T\beta)^2] \tag{46}$$
$$= \mathbb{E}_{\mathcal{D}}[(x_0^T(XX^T)^{-1}X\epsilon)^2] \tag{47}$$
$$\because x_0^T(XX^T)^{-1}X\epsilon \ is \ a \ number \tag{48}$$
$$\therefore \mathbb{E}_{\mathcal{D}}[(x_0^T(XX^T)^{-1}X\epsilon)^2] \tag{49}$$
$$= \mathbb{E}_{\mathcal{D}}[x_0^T(XX^T)^{-1}X\epsilon\epsilon^TX^T(XX^T)^{-1}x_0] \tag{50}$$
$$= \mathbb{E}_{\mathcal{D}}[x_0^T(XX^T)^{-1}XII^TX^T(XX^T)^{-1}x_0\sigma^2] \tag{51}$$
$$= \mathbb{E}_{\mathcal{D}}[x_0^T(XX^T)^{-1}x_0\sigma^2] \tag{52}$$

So $\mathbf{EPE}(x_0) = \sigma^2+\mathbb{E}_{\mathcal{D}}[x_0^T(XX^T)^{-1}x_0\sigma^2]$

Reference:

[1]Wikipedia's introduction of convex function
`https://en.wikipedia.org/wiki/Convex_function`