# Introduction to Machine Learning
# Homework 1

March 19, 2019

## Academic integrity

Our lesson cares much more on academic integrity. No matter who should do our utmost to handle the establishment of academic integrity standard including the host teacher and assistants of this lesson. We hope you will have the same faith with us.

(1) Discussion between students is allowing. The work named by yourself must be completed by your own hands. Any kind of Copying from existing documents will be seen as illegal.

(2) Any kind of Copying from other people's fruits of labour(Publication or Internet documents) will be accused of plagiarism. The score of plagiarists will be canceled. Please mark the authors if you cited any public documents of them;

(3) Highly resemble homework will be seen as Coping. No matter who you are, the one who copy or the one who is copied, both of your score will be canceled. Please protect your homework not to be copied by others actively.

## Homework submission notes

(1) Please follow the submission methods on the website;

(2) If you are not follow the methods or your submission format are not correct. we will deduct some score of your homework;

(3) Unless some special cases(such as illness), the submission over deadline will not be accepted and your score will be set as zero.

# 1 [20pts] Basic Probability and Statistics

The probability distribution of random variable $X$ follows:

$$f_X(x) = \begin{cases} \frac{1}{2} & 0 < x < 1; \\ \frac{1}{6} & 2 < x < 5; \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

(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:

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} z f(z) \mathrm{d}z, \tag{1.2}$$

$$\mathbb{E}[Z] = \int_{z=0}^{\infty} \Pr[Z \geq z] \mathrm{d}z, \tag{1.3}$$

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

(1) 按照定义，$F_X(x)$ 是 $f_X(x)$ 在 $x$ 定义域上的积分，所以 $F_X(x)$ 的函数表示为：

$$F_X(x) = \begin{cases} 0 & x \leq 0; \\ \frac{1}{2}x & 0 < x < 1; \\ \frac{1}{2} & 1 \leq x \leq 2; \\ \frac{1}{6}x + \frac{1}{6} & 2 < x < 5; \\ 1 & x \geq 5. \end{cases} \tag{1.4}$$

(2) 设 $Y$ 的累计分布函数为 $F_Y(y)$，则：

$$\begin{aligned} F_Y(y) &= \Pr[Y \leq y] \\ &= \Pr[\frac{1}{X^2} \leq y] \\ &= \Pr[X \geq \frac{1}{\sqrt{y}}] \\ &= 1 - F_X(\frac{1}{\sqrt{y}}) \end{aligned} \tag{1.5}$$

故 $Y$ 的概率密度函数为:

$$
\begin{aligned}
f_Y(y) &= \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} \\
&= \frac{1}{2} f_X(\frac{1}{\sqrt{y}}) y^{-\frac{3}{2}}
\end{aligned}
\tag{1.6}
$$

因此可得

$$
f_Y(y) = \begin{cases} \frac{1}{12} y^{-\frac{3}{2}} & \frac{1}{25} < y < \frac{1}{4}; \\ \frac{1}{4} y^{-\frac{3}{2}} & y > 1; \\ 0 & \text{otherwise.} \end{cases}
\tag{1.7}
$$

(3) 证明:

$$
\begin{aligned}
\int_{z=0}^{\infty} \Pr[Z \geq z] \mathrm{d}z &= \int_{z=0}^{\infty} \int_{x=z}^{\infty} f(x) \mathrm{d}x \mathrm{d}z \\
&= \int_{x=0}^{\infty} \int_{z=0}^{x} f(x) \mathrm{d}z \mathrm{d}x \\
&= \int_{z=0}^{\infty} z f(z) \mathrm{d}z
\end{aligned}
\tag{1.8}
$$

对于随机变量 $X$ ：

$$
\begin{aligned}
\int_{x=0}^{\infty} x f(x) \mathrm{d}x &= \int_{x=0}^{1} \frac{1}{2} x \mathrm{d}x + \int_{x=2}^{5} \frac{1}{6} x \mathrm{d}x = 2 \\
\int_{x=0}^{\infty} \Pr[X \geq x] \mathrm{d}x &= \int_{x=0}^{\infty} 1 - F_X(x) \mathrm{d}x \\
&= \int_{x=0}^{1} 1 - \frac{1}{2} x \mathrm{d}x + \int_{x=1}^{2} \frac{1}{2} \mathrm{d}x + \int_{x=2}^{5} \frac{5}{6} - \frac{1}{6} x \mathrm{d}x \\
&= 2
\end{aligned}
\tag{1.9}
$$

对于随机变量 $Y$ ：

$$
\begin{aligned}
\int_{y=0}^{\infty} y f(y) \mathrm{d}y &= \int_{y=\frac{1}{25}}^{\frac{1}{4}} \frac{1}{12} y^{-\frac{1}{2}} \mathrm{d}y + \int_{y=1}^{\infty} \frac{1}{4} y^{-\frac{1}{2}} \mathrm{d}y = \infty \\
\int_{y=0}^{\infty} \Pr[Y \geq y] \mathrm{d}y &= \int_{y=0}^{\infty} 1 - F_Y(y) \mathrm{d}y \\
&= \int_{y=0}^{\infty} F_X(\frac{1}{\sqrt{y}}) \mathrm{d}y \\
&= \int_{y=0}^{\frac{1}{25}} 1 \mathrm{d}y + \int_{y=\frac{1}{25}}^{\frac{1}{4}} \frac{1}{6\sqrt{y}} + \frac{1}{6} \mathrm{d}y + \int_{y=\frac{1}{4}}^{1} \frac{1}{2} \mathrm{d}y + \int_{y=1}^{\infty} \frac{1}{2\sqrt{y}} \mathrm{d}y \\
&= \infty
\end{aligned}
\tag{1.10}
$$

故得证。

## 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

$$E(a,b,c) = \sum_{x \in \mathcal{D}} (ax_1^2 + bx_1 + c - x_2)^2. \tag{2.1}$$

(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.

(1) 函数 $E(a,b,c) = \sum_{x \in \mathcal{D}} (ax_1^2 + bx_1 + c - x_2)^2$ 的 Hessian 矩阵为：

$$H = \begin{bmatrix} 2\sum_{x\in\mathcal{D}} x_1^4 & 2\sum_{x\in\mathcal{D}} x_1^3 & 2\sum_{x\in\mathcal{D}} x_1^2 \\ 2\sum_{x\in\mathcal{D}} x_1^3 & 2\sum_{x\in\mathcal{D}} x_1^2 & 2\sum_{x\in\mathcal{D}} x_1 \\ 2\sum_{x\in\mathcal{D}} x_1^2 & 2\sum_{x\in\mathcal{D}} x_1 & 2\sum_{x\in\mathcal{D}} 1 \end{bmatrix} \tag{2.2}$$

即证矩阵 $H$ 是半正定的，这等价于证明矩阵 $H$ 的各阶顺序主子式均非负。而矩阵 $H$ 的各阶顺序主子式依次为：

$$\left| 2\sum_{x\in\mathcal{D}} x_1^4 \right| \geq 0 \tag{2.3}$$

由柯西-施瓦茨不等式，可知对欧几里得空间 $\mathrm{R}^n$：

$$(x_1^2 + x_2^2 + \cdots + x_n^2)(y_1^2 + y_2^2 + \cdots + y_n^2) \geq (x_1y_1 + x_2y_2 + \cdots + x_ny_n)^2 \tag{2.4}$$

当 $\frac{x_1}{y_1} = \frac{x_2}{y_2} = \cdots = \frac{x_n}{y_n}$ 时等式成立。

因此

$$\begin{vmatrix} 2\sum_{x\in\mathcal{D}} x_1^4 & 2\sum_{x\in\mathcal{D}} x_1^3 \\ 2\sum_{x\in\mathcal{D}} x_1^3 & 2\sum_{x\in\mathcal{D}} x_1^2 \end{vmatrix} = 4\left(\sum_{x\in\mathcal{D}} x_1^4 * \sum_{x\in\mathcal{D}} x_1^2 - \left(\sum_{x\in\mathcal{D}} x_1^3\right)^2\right) \geq 0 \tag{2.5}$$

$$\begin{aligned} \begin{vmatrix} 2\sum_{x\in\mathcal{D}} x_1^4 & 2\sum_{x\in\mathcal{D}} x_1^3 & 2\sum_{x\in\mathcal{D}} x_1^2 \\ 2\sum_{x\in\mathcal{D}} x_1^3 & 2\sum_{x\in\mathcal{D}} x_1^2 & 2\sum_{x\in\mathcal{D}} x_1 \\ 2\sum_{x\in\mathcal{D}} x_1^2 & 2\sum_{x\in\mathcal{D}} x_1 & 2\sum_{x\in\mathcal{D}} 1 \end{vmatrix} &= 4\Big(\sum_{x\in\mathcal{D}} x_1^4 * \sum_{x\in\mathcal{D}} x_1^2 * \sum_{x\in\mathcal{D}} 1 + \sum_{x\in\mathcal{D}} x_1^3 * \sum_{x\in\mathcal{D}} x_1 * \sum_{x\in\mathcal{D}} x_1^2 \\ &\quad + \sum_{x\in\mathcal{D}} x_1^2 * \sum_{x\in\mathcal{D}} x_1^3 * \sum_{x\in\mathcal{D}} x_1 - \sum_{x\in\mathcal{D}} x_1^4 * \sum_{x\in\mathcal{D}} x_1 * \sum_{x\in\mathcal{D}} x_1 \\ &\quad - \sum_{x\in\mathcal{D}} x_1^3 * \sum_{x\in\mathcal{D}} x_1^3 * \sum_{x\in\mathcal{D}} 1 - \sum_{x\in\mathcal{D}} x_1^2 * \sum_{x\in\mathcal{D}} x_1^2 * \sum_{x\in\mathcal{D}} x_1^2\Big) \\ &\geq 0 \end{aligned} \tag{2.6}$$

因此可得矩阵 $H$ 是半正定的，故 $E$ 是凸函数。

(2) 存在集合 $D$ 使得 $E$ 是严格凸函数。从 (1) 中可知矩阵 $H$ 的二三阶顺序主子式等号成立的条件是对于所有 $x \in \mathcal{D}$，$x_1$ 均相等。显然只需要构造 $\mathcal{D} = \{(1,2),(3,4)\}$，即可得 $H$ 的二三阶顺序主子式均大于 0 ，此时 $H$ 是正定矩阵，即 $E$ 是凸函数。

# 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 $k \to \infty$.

由题意可知

$$x_{k+1} = \frac{4}{5}x_k + \frac{1}{10}y_k \tag{3.1}$$

$$y_{k+1} = \frac{9}{10}y_k + \frac{1}{5}x_k \tag{3.2}$$

因此

$$P = \begin{bmatrix} \frac{4}{5} & \frac{1}{10} \\ \frac{1}{5} & \frac{9}{10} \end{bmatrix} \tag{3.3}$$

矩阵 $P$ 的特征多项式为

$$P - \lambda I = \begin{vmatrix} \frac{4}{5} - \lambda & \frac{1}{10} \\ \frac{1}{5} & \frac{9}{10} - \lambda \end{vmatrix} \tag{3.4}$$

令多项式为 0 ，解得特征值 $\lambda = 1$ 和 $\lambda = 0.7$
特征值 $\lambda = 1$ 对应的特征向量为 $[2 \quad 1]^\top$, $\lambda = 0.7$ 对应的特征向量为 $[1 \quad -1]^\top$，故可知矩阵对角化时所用的 $Q$ 与 $Q^{-1}$ 分别为：

$$Q = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix} \tag{3.5}$$

$$Q^{-1} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{2}{3} \end{bmatrix} \tag{3.6}$$

于是可知矩阵 $P$ 的对角化结果为

$$\begin{bmatrix} 1 & 0 \\ 0 & 0.7 \end{bmatrix} = D = Q^{-1}PQ \tag{3.7}$$

所以有

$$P = QDQ^{-1} \Rightarrow P^k = QD^kQ^{-1} \tag{3.8}$$

$$\lim_{k\to\infty} P^k = \lim_{k\to\infty} QD^kQ^{-1} = Q\left(\lim_{k\to\infty} D^k\right)Q^{-1} \tag{3.9}$$

故

$$\lim_{k\to\infty} P^k = \begin{bmatrix} \frac{2}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} \end{bmatrix} \tag{3.10}$$

因此

$$\lim_{k\to\infty} P^k[1 \quad 0]^\top = [\frac{2}{3} \quad \frac{1}{3}]^\top \tag{3.11}$$

# 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.

(1) 记原假设 $H0$ 为"硬币正面不偏重"，备择假设 $H1$ 为"硬币正面偏重"。抛硬币本身属于二项分布，即 $X \sim B(n, 0.5)$ ，由题意可知 $n = 50$，属于大样本，故可用正态分布来近似，得到：

$$z = \frac{p - 0.5}{\sqrt{\frac{0.5*(1-0.5)}{n}}} = \frac{p - 0.5}{0.5}\sqrt{n} \sim N(0,1) \tag{4.1}$$

将 $p = \frac{35}{50} = 0.7$ 代入，可得 $z = 2\sqrt{2} \approx 2.83$，查表可知 $z_{1-\alpha} = z_{0.95} \approx 1.65 < 2.83$，落入拒绝域，因此原假设不成立，即硬币正面偏重。

(2)

$$p = P(Z > z) = P(Z > 2.83) = 1 - \Phi(2.83) = 0.0023 \tag{4.2}$$

$p$ 值的意思是当原假设为真的条件下，检验统计量的观察值大于或等于其计算值的概率。此处 $p$ 值指的是在硬币正面不偏重的情况下，$z$ 值大于 2.83 的概率。因为 $p = 0.0023 < 0.05 = \alpha$，所以拒绝原假设。

## 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 $th_2 = 0.1$.

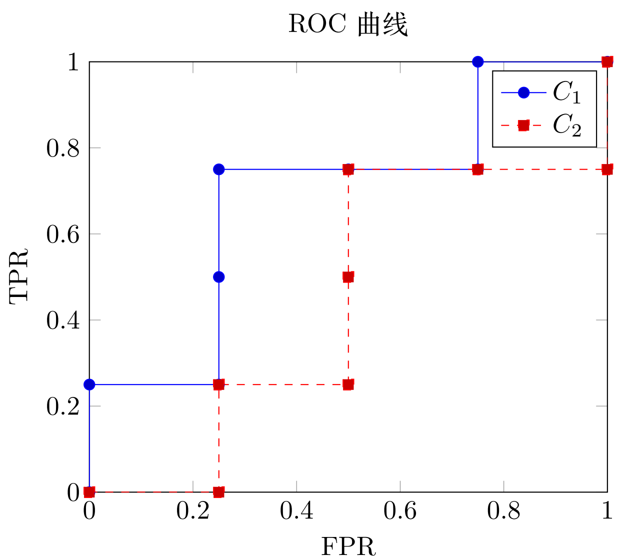
(3) [4pts] Prove Eq.(2.22) in Page 35. (AUC $= 1 - \ell_{rank}$).

(1) 将 $y_{C_1}$ 与 $y_{C_2}$ 的数据按分数从大到小排序可得表 1

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

Table 1: $C_1$ 与 $C_2$ 的分数结果排序

根据表 1 分别计算分类器 $C_1$ 与 $C_2$ 的 TPR 与 FPR，可得各自的 ROC 曲线为:




$$
\begin{aligned}
AUC_{C_1} &= 0.25 * 0.25 + 0.75 * (0.75 - 0.25) + 1 * (1 - 0.75) = 0.6875 \\
AUC_{C_2} &= 0.25 * (0.5 - 0.25) + 0.75 * (1 - 0.5) = 0.4375
\end{aligned} \tag{5.1}
$$

(2) 分别计算分类器 $C_1$ 与 $C_2$ 的混淆矩阵，可得：

$C_1$ 的混淆矩阵：

| | | Actual Class | |
|---|---|---|---|
| | | 1 | 0 |
| Predicted | 1 | 3 | 2 |
| Class | 0 | 1 | 2 |

$C_2$ 的混淆矩阵：

| | | Actual Class | |
|---|---|---|---|
| | | 1 | 0 |
| Predicted | 1 | 3 | 4 |
| Class | 0 | 1 | 0 |

分别计算分类器 $C_1$ 与 $C_2$ 的 F 分数，可得：

$$
\begin{aligned}
F_{1_{C_1}} &= 2 * \frac{0.6 * 0.75}{0.6 + 0.75} = 0.67 \\
F_{1_{C_2}} &= 2 * \frac{0.43 * 0.75}{0.43 + 0.75} = 0.55
\end{aligned} \tag{5.2}
$$

(3) 由 AUC 定义可知 $1-\mathrm{AUC}$ 即 ROC 曲线以上的面积。设 $S^+$ 为正例的序号集合，$S^-$ 为负例的序号集合，则

$$
\begin{aligned}
1-\mathrm{AUC} &= \frac{1}{2}\sum_{i=1}^{m-1}\left(x_i + x_{i+1}\right)\left(y_{i+1} - y_i\right) \\
&= \frac{1}{m^+}\sum_{i\in S^+} x_i + 0\sum_{i\in S^-} x_i \\
&= \frac{1}{m^+}\sum_{i\in S^+} x_i \\
&= \frac{1}{m^+}\sum_{i\in S^+}\frac{1}{m^-}\sum_{j\in S^-}\left(\mathbb{I}(f(x_i) < f(x_j)) + \frac{1}{2}\mathbb{I}(f(x_i) = f(x_j))\right) \\
&= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))\right) \\
&= l_{rank}
\end{aligned} \tag{5.3}
$$

第 1 个等号代表从水平方向逐步增加 ROC 曲线以上的面积，第 2 个等号表明负例集合增加的单位直线没有增加 ROC 曲线以上的面积。
因此 $\mathrm{AUC} = 1 - \ell_{rank}$ 成立。

# 6 [Bonus 10pts]Expected Prediction Error

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

$$
y = x^T\beta + \epsilon, \tag{6.1}
$$

where $\epsilon$ is noise and follows $\epsilon \sim N(0, \sigma^2)$. Note the instance feature of training data $\mathcal{D}$ as $\boldsymbol{X} \in \mathbb{R}^{p\times n}$ 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:

$$
\hat{\beta} = (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{Y}. \tag{6.2}
$$

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

$$
\mathbf{EPE}(x_0) = \sigma^2 + \mathbb{E}_{\mathcal{D}}[x_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}x_0\sigma^2]. \tag{6.3}
$$

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)

根据 $E(f;D) = bias^2(x) + var(x) + \epsilon^2$ 即
$E(f;D) = \mathbb{E}_D[(f(x;D) - \bar{f}(x))^2] + (\bar{f}(x) - y)^2 + \mathbb{E}_D[(y_D - y)^2]$
可知

$$\begin{aligned}\mathbf{EPE}(x_0) &= \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathcal{D}}[(y_0 - \hat{y}_0)^2] \\ &= \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathcal{D}}[(y_0 - y)^2] + \mathbb{E}_{y_0|x_0}[(\bar{\hat{y}}_0 - y)^2] + \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathcal{D}}[(\hat{y}_0 - \bar{\hat{y}}_0)^2] \\ &= \sigma^2 + 0 + \mathbb{E}_{\mathcal{D}}[(\hat{y}_0 - \bar{\hat{y}}_0)^2]\end{aligned} \tag{6.4}$$

即证

$$\mathbb{E}_{\mathcal{D}}[(\hat{y}_0 - \bar{\hat{y}}_0)^2] = \mathbb{E}_{\mathcal{D}}[x_0^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}x_0\sigma^2] \tag{6.5}$$