Review about Machine Learning

Chou Pham

April 2021

1 Probability

1.1 Problem 1.1

1. $A = \{1, 2\}$ and $B = \{1, 3\}$.
Consider:

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

$P(A) = P(\{1, 2\}) = P(\{1\}) + P(\{2\}) = \frac{1}{2}$.
$P(B) = P(\{1, 3\}) = P(\{1\}) + P(\{3\}) = \frac{1}{2}$
Then $P(A \cap B) = P(A)P(B)$, thus, A and B are independent.
2. Consider:

$$P(A \cap B \cap C) = P(\emptyset) = 0$$

Easy to verify $P(A)P(B)P(C) = 1/8$, thus A, B, C are not mutual independent.
3. Consider:

$$P(A \cap B|C) = \frac{P(A \cap B \cap C)}{p(C)} = 0$$

and $P(A \cap B)P(B \cap C) = 1/16$, then A, B are not conditional independent given C.

1.2 Problem 1.2

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

1.3 Problem 1.3

:

1.
$X \sim \mathcal{N}(\mu, C)$, then:

$$p(x) = \eta \exp(-0.5(x-\mu)^T C^{-1}(x-\mu)) \tag{1}$$

Since C is a symmetric positive definite matrix, we can write $C = U\Lambda U^T$, with Λ is a diagonal matrix, and U is a orthogonal matrix ($U^T = U^{-1}$). Note that:

$$(U\Lambda U^T)(U\Lambda^{-1}U^T) = U\Lambda(U^T U)\Lambda^{-1}U^T = UU^T = I$$

So we have:

$$C^{-1} = U\Lambda^{-1}U^T$$

Rewrite the distribution $p(x)$ in (1):

$$p(x) = \eta \exp(-0.5(x-\mu)^T U\Lambda^{-1}U^T(x-\mu))$$

Change of variable, set $u = U^T(x-\mu)$, then the distribution of u is:

$$p(u) = \eta \exp(-0.5u^T\Lambda^{-1}u)$$

$$\implies U \sim \mathcal{N}(0, \Lambda)$$

2.
Consider:

$$U = \begin{bmatrix} X \\ Y \end{bmatrix}$$

U is a normal distribution with mean μ and covariance matrix Σ are determined by:

$$\mu = \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} C & 0_{d\times d} \\ 0_{d\times d} & D \end{bmatrix}$$

The term: $X + Y = [I_d, I_d]U$. We apply the affine transform properties:

$$p(AX + b) = \mathcal{N}(A\mu + b, A\Sigma A^T) \tag{2}$$

with assumption: $X \sim \mathcal{N}(\mu, \Sigma)$.

$$p(U) = p([I_d, I_d][X, Y]^T) = \mathcal{N}(a + b, C + D)$$

2 Decision Theory

2.1 Problem 2.1

$y \in \{0,1\}$. Denote the predict of classifier: $\hat{y} = f(x)$. The loss function between y and $\hat{y}$:

$$\mathcal{L}(y,\hat{y}) = \begin{cases} 0, \text{ if } \hat{y} = y \\ 1, \text{ if } \hat{y} \neq y \text{ and } \hat{y} \in \{0,1\} \\ \lambda, \text{ if } \hat{y} = \text{reject} \end{cases}$$

We compute the expectation of loss:

$$E(L(Y, f(x))|X = x) = p(y = 0|x)L(y = 0, \hat{y}) + p(y = 1|x)L(y = 1, \hat{y})$$

Denote: $\alpha = p(y = 1|x)$, then: $p(y = 0|x) = 1 - \alpha$. Then:

$$E[L(y,\hat{y})|x] = \alpha L(y = 1, \hat{y}) + (1 - \alpha)L(y = 0, \hat{y})$$

Then we have 3 decisions:
- Case 1: $\hat{y} = 1$, then $E[L(y,\hat{y})|x] = 1 - \alpha$.
- Case 2: $\hat{y} = 0$, then $E[L(y,\hat{y})|x] = \alpha$.
- Case 3: $\hat{y} =$reject, then $E[L(y,\hat{y})|x] = \lambda$.

Then the decision theory is:

$$f(x) = \begin{cases} 1, \text{ if } p(y = 1|x) \geq 0.5 \text{ and } \lambda > p(y = 0|x) \\ 0, \text{ if } p(y = 0|x) > 0.5 \text{ and } \lambda > p(y = 0|x) \\ \text{reject, otherwise} \end{cases}$$

3 Maximum likelihood

3.1 Problem 3.3

1.

$$p(y = 1|x) = \frac{p(x, y = 1)}{p(x)} = \frac{p(x|y = 1)p(y = 1)}{p(x|y = 0)p(y = 0) + p(x|y = 1)p(y = 1)} \tag{3}$$

We have:

$$\begin{cases} p(y = 1) = \Phi \\ p(y = 0) = 1 - \Phi \end{cases} \tag{4}$$

$$\begin{cases} p(x|y = 1) = \lambda_1 \exp(-\lambda_1 x) \\ p(x|y = 0) = \lambda_0 \exp(-\lambda_0 x) \end{cases} \tag{5}$$

Substitute (4),(5) to (3):

$$\begin{aligned}
p(y=1|x) &= \frac{\Phi\lambda_1\exp(-\lambda_1 x)}{\Phi\lambda_1\exp(-\lambda_1 x)+(1-\Phi)\lambda_0\exp(-\lambda_0 x)} \\
&= \frac{1}{1+\frac{1-\Phi}{\Phi}\frac{\lambda_0}{\lambda_1}\exp(-(\lambda_0-\lambda_1)x)} \\
&= \frac{1}{1+\exp(-(\theta_0+\theta_1 x))}
\end{aligned}$$

We choose:

$$\begin{cases}\theta_1 = \lambda_0 - \lambda_1 \\ \theta_0 = \log(\frac{\Phi\lambda_1}{\lambda_0(1-\Phi)})\end{cases}$$

2. Maximum likelihood

$$\begin{aligned}
l(\Phi,\lambda_0,\lambda_1) &= \sum_{i=1}^{n}\log(p(x_i,y_i|\Phi,\lambda_0,\lambda_1)) \\
&= \sum_{i=1}^{n}[\log(p(x_i|y_i,\Phi,\lambda_0,\lambda_1))+\log(y_i|\Phi,\lambda_0,\lambda_1)] \qquad (6)\\
&= \sum_{i=1}^{n}[\log(p(x_i|y_i,\lambda_0,\lambda_1))+\log(p(y_i|\Phi))]
\end{aligned}$$

We have:

$$p(x_i|y_i,\lambda_0,\lambda_1) = (\lambda_1\exp(-\lambda_1 x_i))^{y_i}(\lambda_0\exp(-\lambda_0 x_i))^{1-y_i}$$

$$p(y_i|\Phi) = \Phi^{y_i}(1-\Phi)^{1-y_i}$$

Substitute to (6):

$$\begin{aligned}
l(\Phi,\lambda_0,\lambda_1) &= \sum_{i=1}^{n}[y_i(\log(\lambda_1)-\lambda_1 x_i)+(1-y_i)(\log(\lambda_0)-\lambda_0 x_i)+y_i\log(\Phi)+(1-y_i)\log(1-\Phi)] \\
&= N_1\log(\lambda_1)+N_0\log(\lambda_0)-\lambda_1\sum_{i:y_i=1}x_i-\lambda_0\sum_{i:y_i=0}x_i+N_1\log(\Phi)+N_0\log(1-\Phi)
\end{aligned} \qquad (7)$$

where N_1 is the number of $y_i=1$ and N_0 is the number of $y_i=0$. We have the partial derivatives:

$$\frac{\partial l(\Phi,\lambda_0,\lambda_1)}{\partial\Phi} = \frac{N_1}{\Phi}-\frac{N_0}{1-\Phi}$$

$$\frac{\partial l}{\partial\lambda_0} = \frac{N_0}{\lambda_0}-\sum_{i:y_i=0}x_i$$

$$\frac{\partial l}{\partial \lambda_1} = \frac{N_1}{\lambda_1} - \sum_{i:y_i=1} x_i$$

Second order derivatives:

$$\frac{\partial^2 l}{\partial \phi^2} = -\frac{N_1}{\Phi^2} - \frac{N_0}{(\Phi - 1)^2} < 0$$

$$\frac{\partial^2 l}{\partial \lambda_1^2} = -\frac{N_1}{\lambda_1^2} < 0$$

$$\frac{\partial^2 l}{\partial \lambda_0^2} = -\frac{N_0}{\lambda_0^2} < 0$$

Then solving the derivative equation will help us to find the optimal value:

$$\begin{cases} \Phi = N_1/(N_1 + N_0) \\ \lambda_0 = \frac{N_0}{\sum_{i:y_i=0} x_i} \\ \lambda_1 = \frac{N_1}{\sum_{i:y_i=1} x_i} \end{cases}$$

3.2 Problem 3.4

1. MLE for θ: Given $X_1, X_2, ..., X_n \sim U(0, \theta)$.

Uniform distribution pdf:

$$p(x) = \begin{cases} \frac{1}{\theta} \text{ if } x \in [0, \theta] \\ 0 \text{ otherwise} \end{cases}$$

Likelihood:

$$p(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

Case 1: if $\theta < \max_i\{x_i\}$, then there exist a i index satisfies $x_i > \theta$. Then $p(x_i|\theta) = 0$. We conclude the likelihood $p(x_1, x_2, ..., x_n|\theta) = 0$.

Case 2: if $\theta \geq \max_i\{x_i\}$, the for each value x_i: $p(x_i|\theta) = \frac{1}{\theta}$. The likelihood becomes:

$$p(x_1, x_2, ..., x_n|\theta) = \frac{1}{\theta^n} \leq \frac{1}{x_{max}^n}$$

where $x_{max} = \max_i\{x_i\}$.

Then the solution for MLE is $\theta = x_{max}$.

2. Pareto prior and posterior distribution.

We choose the prior for θ:

$$p(\theta) = \alpha\beta^\alpha\theta^{-\alpha-1} I_{\beta,\infty}(\theta)$$

where:

$$I(\theta, \infty)(\theta) = \begin{cases} 1 \text{ if } \theta > \beta \\ 0 \text{ otherwise} \end{cases}$$

The posterior distribution can be computed as:

$$p(\theta|x_1, x_2, ..., x_n) = \frac{p(x_1, x_2, ..., x_n|\theta)p(\theta)}{p(x_1, x_2, ..., x_n)} \tag{8}$$

Case 1: if $\theta < \max_i\{x_i\}$, then $p(\theta|x_1, x_2, ..., x_n) = 0$.
Case 2: if $\theta \geq \max_i\{x_i\}$:

$$p(x_1, x_2, ..., x_n|\theta) = \frac{1}{\theta^n} = \theta^{-n}$$

$$p(\theta) = \begin{cases} 0 \text{ if } \theta \leq \beta \\ \alpha\beta^\alpha\theta^{-\alpha-1} \text{ otherwise} \end{cases}$$

The distribution:

$$\begin{aligned} p(x_1, x_2, ..., x_n) &= \int p(x_1, x_2, ..., x_n|\theta)p(\theta)d\theta \\ &= \int_{u=\max\{\beta, x_{max}\}}^{\infty} \frac{1}{\theta^n}\alpha\beta^\alpha\theta^{-\alpha-1}d\theta \\ &= \int_u^\infty \alpha\beta^\alpha\theta^{-\alpha-1-n}d\theta \\ &= \alpha\beta^\alpha u^{-\alpha-n}/(\alpha+n) \end{aligned} \tag{9}$$

The the posterior distribution:

$$p(\theta|x_1, x_2, ..., x_n) = \frac{\theta^{-\alpha-n-1}(\alpha+n)}{u^{-\alpha-n}} \text{ if } \theta \geq u$$

the we have the positerior distribution form:

$$p(\theta|x_1, x_2, ..., x_n) = \begin{cases} 0 \text{ if } \theta < u \\ \theta^{-\alpha-n-1}(\alpha+n)u^{\alpha+n} \text{ if } \theta \geq u \end{cases} \tag{10}$$

with $u = \max\{x_{max}, \beta\}$.
3. MAP for θ
If $\theta < u$, then $p(\theta|x_1, x_2, ..., x_n) = 0$.
If $\theta \geq u$, then $p(\theta|x_1, x_2, ..., x_n) = \theta^{-\alpha-n-1}(\alpha+n)u^{\alpha+n}$. Then the solution for MAP is $\theta = u$. Then we have two cases:
Case1: $\beta \leq x_{max}$, then $\theta = x_{max}$. This is also the solution using MLE.
Case2: $\beta > x_{max}$, then $\theta = \beta$.
4. Optimal θ under square loss.

$$
\begin{aligned}
E_{\theta\sim p(\theta|x_1,...,x_n)}[(\theta-\hat{\theta})^2] &= \int p(\theta|x_1,...,x_n)(\theta-\hat{\theta})^2 d\theta \\
&= (\alpha+n)u^{\alpha+n}\int_u^{\infty}\theta^{-\alpha-n-1}(\theta-\hat{\theta})^2 d\theta
\end{aligned}
$$

We will drop the constant term $(\alpha+n)u^{\alpha+n}$ here since it doesn't contribute to find the optimal value of $\hat{\theta}$.

$$
\begin{aligned}
\hat{\theta} &= \arg\min_{\hat{\theta}} F(\hat{\theta}) \\
&= \arg\min_{\hat{\theta}} \hat{\theta}^2\int_u^{\infty}\theta^{-\alpha-n-1}d\theta - 2\hat{\theta}\int_u^{\infty}\theta^{-\alpha-n}d\theta \\
&= \arg\min_{\hat{\theta}} a\hat{\theta}^2 - 2b\hat{\theta}
\end{aligned}
$$

with $a = \int_u^{\infty}\theta^{-\alpha-n-1}d\theta$ and $b = \int_u^{\infty}\theta^{-\alpha-n}d\theta$.
This is a quadratic function of $\hat{\theta}$ and the optimal value of $\hat{\theta}$ is:

$$\hat{\theta} = \frac{b}{a}$$

(the computation of b and a is left for the readers).

4 Bayesian Inference

4.1 Problem 4.1

The distribution:

$$p(x|\theta,\lambda) = \mathcal{N}(x|\theta,\lambda) = \sqrt{\frac{\lambda}{2\pi}}\exp(-\frac{\lambda}{2\pi}(x-\theta)^2)$$

the prior:

$$p(\theta) = \mathcal{N}(\theta|\mu_0,\lambda_0) = \sqrt{\frac{\lambda_0}{2\pi}}\exp(-\frac{\lambda_0}{2\pi}(\theta-\mu_0)^2)$$

then the posterior distribution:

$$
\begin{aligned}
p(\theta|x_1,x_2,...,x_n) &= \frac{p(x_1,x_2,...,x_n|\theta)p(\theta)}{p(x_1,x_2,...,x_n)} \\
&= \frac{\prod_{i=1}^n p(x_i|\theta)p(\theta)}{p(x_1,x_2,...,x_n)} \\
&= \eta\exp(-\frac{\lambda}{2}\sum_{i=1}^n(x_i-\theta)^2 - \frac{\lambda_0}{2}(\theta-\mu_0)^2)
\end{aligned}
\tag{11}
$$

The inner exp term in (11) is a square term of θ.

We consider:

$$f(\theta) = \frac{\lambda}{2}\sum_{i=1}^{n}(\theta - x_i)^2 + \frac{\lambda_0}{2}(\theta - \mu_0)^2 = a\theta^2 + b\theta + c$$

and a can be solved by finding:

$$\frac{\partial f}{\partial \theta} = \lambda\sum_{i=1}^{n}(\theta - x_i) + \lambda_0(\theta - \mu_0)$$

$$\frac{\partial^2 f}{\partial \theta^2} = n\lambda + \lambda_0 > 0$$

then $a > 0$, then (11) is a normal distribution. The mean of that normal can be found by solving the derivative equation:

$$\frac{\partial L}{\partial \theta} = 0$$

then

$$\theta = \frac{\lambda \sum_{i=1}^{n} x_i + \lambda_0\mu_0}{\lambda_0 + n\lambda}$$

Then the posterior distribution is:

$$p(\theta|x_1, x_2, ..., x_n) = \mathcal{N}(\theta|M, L^{-1})$$

where:

$$L = n\lambda + \lambda_0$$

and

$$M = \frac{\lambda_0\mu_0 + \lambda \sum_{i=1}^{n} x_i}{\lambda_0 + n\lambda}$$

Now after computing $p(\theta|x_1, x_2, ..., x_n)$, we can use it to compute $p(x_1, x_2, ..., x_n)$.

$$p(x_1, x_2, ..., x_n) = \frac{p(x_1, x_2, ..., x_n|\theta)p(\theta)}{p(\theta|x_1, x_2, ..., x_n)} \tag{12}$$

We compute each term in the first stage:

$$p(x_1, x_2, ..., x_n|\theta) = \eta \exp(-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i - \theta)^2)$$

$$p(\theta) = \eta \exp(-\frac{\lambda_0}{2}(\theta - \mu_0)^2)$$

$$p(\theta|x_1, x_2, ..., x_n) = \eta \exp(-\frac{1}{2L}(\theta - M)^2)$$

4.2 Problem 4.2

$X_i \sim W_d(X_i|S^{-1}, \nu)$ implies:

$$p(X_i|S^{-1}, \nu) = \frac{|S|^{n/2}|X_i|^{(\nu-d-1)/2}\exp(-\frac{1}{2}trace(SX_i))}{2^{\nu d/2}\Gamma_d(\frac{\nu}{2})}$$

The prior: $p(S) = W_d(S|S_0^{-1}, \nu_0)$:

$$p(S) = \frac{|S_0|^{n/2}|S|^{(\nu_0-d-1)/2}\exp(-\frac{1}{2}trace(S_0S))}{2^{\nu_0 d/2}\Gamma_d(\frac{\nu_0}{2})}$$

The distribution:

$$p(S|X_1, X_2, ..., X_n) = \frac{p(X_1, X_2, ..., X_n|S)p(S)}{p(X_1, X_2, ..., X_n)} = \eta p(X_1, X_2, ..., X_n|S)p(S) \tag{13}$$

The likelihood can be computed as:

$$\begin{aligned} p(X_1, X_2, ..., X_n|S) &= \prod_{i=1}^{n} p(X_i|S) \\ &= \frac{|S|^{n^2/2}(\prod_{i=1}^n |X_i|)^{(\nu-d-1)/2}\exp(-\frac{1}{2}trace(S\sum_{i=1}^n X_i))}{2^{n\nu d/2}(\Gamma_d(\nu/2))^n} \end{aligned} \tag{14}$$

Then, replace (14) to (13), we can compute the posterior distribution:

$$p(S|X_1, X_2, ..., X_n) = \eta\frac{|S|^{n^2/2+(\nu-d-1)/2}|S_0|^{n/2}(\prod_{i=1}^n |X_i|)^{(\nu-d-1)/2}\exp(-\frac{1}{2}trace(S(\sum_{i=1}^n X_i + S_0)))}{2^{n\nu d/2+\nu_0 d/2}\Gamma_d(\frac{\nu}{2})^n\Gamma_d(\frac{\nu_0}{2})}$$

We choose:

$$\begin{cases} S' = \sum_{i=1}^n X_i + S_0 \\ \nu' = n^2/2 + \nu \end{cases}$$

then we have:

$$p(S|X_1, X_2, ..., X_n) = \eta\frac{|S'|^{n/2}|S|^{(\nu'-d-1)/2}\exp(-\frac{1}{2}trace(S'S))}{2^{\nu' d/2}\Gamma_d(\nu'/2)}\beta = \eta\beta W_d(S|S', \nu') \tag{15}$$

with β is a constant term and does not depend on S.
We make the integral of (15), then:

$$\int_S p(S|X_1, X_2, ..., X_n)dS = \eta\beta = 1$$

then: $p(S) = W_d(S|S', \nu')$
with:

$$\begin{cases} S' = \sum_{i=1}^n X_i + S_0 \\ \nu' = n^2/2 + \nu \end{cases}$$

5 Linear Regression

1. Solving linear regression problem.
Model: $y_i = \sum_{k=0}^m c_k \cos(2\pi k x_i) + \epsilon_i$
with $\epsilon_i \sim \mathcal{N}(0, 1)$.
Set $u_i = [\cos(2\pi k x_i)]_{k=\overline{0,m}}^T$
Then $y_i = c^T u_i + \epsilon_i$.
$p(y_i|x_i, c) = \mathcal{N}(y_i|c^T u_i, \sigma^2)$.
This problem here is quite simple, and we can use the approach in the slide to solve it.

2. Conjugacy for coeficient in Linear Regression Problem
I think we can see the exercise 4.1 for more details.

6 Logistic Regression

6.1 Problem 6.1

1. $\sigma(a) = \frac{1}{1+e^{-a}}$

$$\frac{d\sigma}{da} = \sigma(a)(1 - \sigma(a))$$

2. Objective function for binary logistic regression:

$$\mathcal{L}(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

where $\hat{y} = \sigma(w^T x + b) = \sigma(u)$.
Gradient computing:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w} &= \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial u} \frac{\partial u}{\partial w} \\ &= -(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}) \sigma(u)(1 - \sigma(u)) x \\ &= -(y - \hat{y}) x \end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -(y - \hat{y})$$

3. Compute the Hessian:
We denote the weight vector here is: $w = [w_1, w_2, ..., w_d]$, where d is the dimension of w plus 1 (count b).

We have:

$$\frac{\partial \mathcal{L}}{\partial w_i} = \sum_{k=1}^{n} (\hat{y}^{(k)} - y^{(k)}) x_i^{(k)}$$

where $y^{(j)}, x^{(j)}$ denotes label and sample j.
Second order derivative:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}}{\partial w_i w_j} &= \frac{\partial \mathcal{L}}{\partial w_j} (\sum_{k=1}^{n} (\hat{y}^{(k)} - y^{(k)}) x_i^{(k)}) \\ &= \sum_{k=1}^{n} \hat{y}^{(k)} (1 - \hat{y}^{(k)}) x_i^{(k)} x_j^{(k)}\end{aligned}$$

Denote:

$$X = \left[x^{(1)}, x^{(2)}, ..., x^{(n)}\right] \in R^{n \times d}$$

$$Y = diag(y^{(1)}(1 - y^{(1)}), y^{(2)}(1 - y^{(2)}), ..., y^{(n)}(1 - y^{(n)}))$$

Then the Hessian matrix H is:

$$H = X^T Y X$$

since $y^{(k)} = \sigma(w^T x^{(k)} + b) \in (0, 1)$
then:

$$u^T H u = u^T X^T Y H u = (Hu)^T Y (Hu) = v^T Y v$$

Since Y is a positive-definite metric, so $u^T H u \geq 0$, then H also is a positive definite matrix.

7 Estimators

7.1 Problem 7.1

Derive of biased esimator of variance:
We consider:

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{16}$$

First of all, if n random variables $X_1, X_2, ..., X_n$ follows the same distribution as variable X, then we have:

$$E[\frac{X_1 + X_2 + ... + X_n}{n}] = E[\overline{X}] = E[X] = \mu$$

we find another representation of (16):

$$
\begin{aligned}
S^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu)^2 + (\mu - \overline{X})^2 + 2(X_i - \mu)(\mu - \overline{X})] \qquad (17) \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{2}{n}\sum_{i=1}^{n}(X_i - \mu)(\mu - \overline{X}) + (\mu - \overline{X})^2
\end{aligned}
$$

We have the term:

$$
\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n}(X_i - \mu)(\mu - \overline{X}) &= \frac{2}{n}(\mu - \overline{X})(\sum_{i=1}^{n} X_i - n\mu) \\
&= -2(\overline{X} - \mu)^2
\end{aligned}
$$

Substitute to (17):

$$
S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2 \qquad (18)
$$

Then we have:

$$
\sigma^2 = E[(X - \mu)^2] = E[\sum_{i=1}^{n}\frac{1}{n}(X_i - \mu)^2]
$$

$$
E[(\overline{X} - \mu)^2] = E[(\overline{X} - E[\overline{X}])^2] = V[\overline{X}] = \frac{1}{n}\sigma^2
$$

Substitute to (18):

$$
E[S^2] = \frac{n-1}{n}\sigma^2
$$

8 Empirical Risk minimization

Let's finish 3 examples of A Toan's slide:

1. Binary classifier:

The risk:

$$
R(f) = E_{(X,Y)\sim P}[l(f(X), Y)]
$$

For a value x, we consider:

$$
\begin{aligned}
R(f(x)) &= E_{y\sim p(y|x)}[l(f(x), y)] \\
&= E_{y\sim p(y|x)}[l(\hat{y}, y)] \\
&= p(y = 0|x)l(\hat{y}, y = 0) + p(y = 1|x)l(\hat{y}, y = 1)
\end{aligned}
$$

Consider two cases:

Case 1: If $p(y=1|x) \geq p(y=0|x)$.

Then, if $\hat{y}=0$, $R(f(x)) = p(y=1|x)$.

if $\hat{y}=1$, then $R(f(x)) = p(y=0|x)$.

It's easy to find that: with $\hat{y}=1$, the risk has lower value. So in this case, we choose: $\hat{y}=1$.

Case 2: If $p(y=1|x) < p(y=0|x)$.

Then if $\hat{y}=0$, $R(f(x)) = p(y=1|x)$.

if $\hat{y}=1$, then $R(f(x)) = p(y=0|x)$.

Then in this case, $\hat{y}=0$ gives lower cost.

We conclude:

$$\hat{y} = \begin{cases} 1, & \text{if } p(y=1|x) \geq p(y=0|x) \\ 0, & \text{otherwise.} \end{cases}$$

2. Regression with L1-loss.

The solution for this problem is quite complex, so if you can find a simpler solution, please let me know.

For a sample x (we only consider in the continous case). The risk for this sample:

$$\begin{aligned} R(f(x)) &= E_{y \sim p(y|x)}[l(\hat{y}, y)] \\ &= \int_{-\infty}^{\infty} p(y|x)|\hat{y} - y|dy \end{aligned} \tag{19}$$

For simplicity, just in this section, i will denote $p(y) = p(y|x)$.

Then:

$$\begin{aligned} R(\hat{y}) &= \int_{-\infty}^{\infty} p(y)|\hat{y} - y|dy \\ &= \int_{-\infty}^{\hat{y}} p(y)(\hat{y} - y)dy + \int_{\hat{y}}^{\infty} p(y)(y - \hat{y})dy \\ &= \hat{y}\int_{-\infty}^{\hat{y}} p(y)dy - \hat{y}\int_{\hat{y}}^{\infty} p(y)dy - \int_{-\infty}^{\hat{y}} yp(y)dy + \int_{\hat{y}}^{\infty} yp(y)dy \\ &= 2\hat{y}\int_{-\infty}^{\hat{y}} p(y)dy - \hat{y} + 2\int_{\hat{y}}^{\infty} yp(y)dy - E[y] \end{aligned} \tag{20}$$

Since $E[y]$ is a constant term and not depends on $\hat{y}$, we can ommit it in the optimization procedure of (20). We denote:

$$F(t) = \int_{\infty}^{t} p(y)dy$$

is the CDF function of $p(y)$. Denote: $u = Med(y)$, then $F(u) = 0.5$.

We will prove:

$$R(\hat{y}) \geq R(u)$$

It is equivalent to:

$$2\hat{y}F(\hat{y}) - \hat{y} + 2\int_{\hat{y}}^{\infty} yp(y)dy \geq 2\int_{u}^{\infty} yp(y)dy \tag{21}$$

We consider 2 cases:
Case 1: $\hat{y} < u$, then (21) is equivalent to:

$$2\hat{y}F(\hat{y}) - \hat{y} + 2\int_{\hat{y}}^{u} yp(y)dy \geq 0 \tag{22}$$

The integral term can be written as:

$$\begin{aligned}\int_{\hat{y}}^{u} yp(y)dy &= yF(y)_{\hat{y}}^{u} - \int_{\hat{y}}^{u} F(y)dy \\ &= uF(u) - \hat{y}F(\hat{y}) - \int_{\hat{y}}^{u} F(y)dy\end{aligned}$$

Plug into (22), it is equivalent to prove:

$$\begin{aligned}&2\hat{y}F(\hat{y}) - \hat{y} + u - 2\hat{y}F(\hat{y}) - 2\int_{\hat{y}}^{u} F(y)dy \\ &= u - \hat{y} - 2\int_{\hat{y}}^{u} F(y)dy\end{aligned} \tag{23}$$

We also have: $F(y) \leq 0.5$ for all $y \in [\hat{y}, u]$. Then:

$$\int_{\hat{y}}^{u} F(y)dy \leq \frac{1}{2}(u - \hat{y})$$

Substitute to (23), it turns out a true states.
We conclude: $R(\hat{y}) > R(u)$ for all $\hat{y} < u$.
Similar prove, we also conclude: $R(\hat{y}) > R(u)$ for all $\hat{y} > u$.
Then we conclude:

$$R(\hat{y}) \geq R(u)$$

and the optimal value for $\hat{y}$ is: $\hat{y} = Med(y)$.
3. L2 loss:
This is easier than L1-loss, for each sample x:

$$\begin{aligned}E_{y\sim p(y|x)}[l(\hat{y}, y)] &= \int_{-\infty}^{\infty} p(y|x)(\hat{y} - y)^2 dy \\ &= \hat{y}^2 - 2\hat{y}\int_{-\infty}^{\infty} p(y|x)ydy + E[y^2]\end{aligned}$$

This is a quadratic form of $\hat{y}$, then $\hat{y}$ for optimal risk can be found by:

$$\hat{y} = \int_{-\infty}^{\infty} yp(y|x)dy = E[y|x]$$