Introduction to Machine Learing: Homework II

Due on Oct 26th, 2022 at 11:59pm

Professor Ziping Zhao

Bingnan Li
2020533092

1. [Bayesian Decision Theory]

(a) Specify decision rule via likelihood and posterior respectively, and give example with $x = light$ to use the rules.
Solution:

Likelihood decision rule:

$$choose \begin{cases} C = C_1, \; if \; P(x|C_1) > P(x|C_2) \\ C = C_2, \; elsewhere \end{cases}$$

Posterior decision rule:

$$choose \begin{cases} C = C_1, \; if \; P(C_1|x) > P(C_2|x) \\ C = C_2, \; elsewhere \end{cases}$$

If $x = light$, then likelihood is

$$P(x = light|C_1) = \frac{2125}{2125 + 6375} = 0.25$$

$$P(x = light|C_2) = \frac{1000}{1000 + 500} = 0.667$$

Thus, by likelihood decision rule, we choose C_2. However, the posterior probability is

$$P(C_1|x = light) = \frac{2125}{2125 + 1000} = 0.68$$

$$P(C_2|x = light) = \frac{1000}{2125 + 1000} = 0.32$$

Thus, by posterior decision rule, we choose C_1.

(b) (i) Find the optimal decision rule that will give the minimum expected loss.
Solution:

Let action α_i be classifying input $\boldsymbol{x}$ into class i and $R(\alpha_i|\boldsymbol{x})$ be the expected loss of taking action α_i, then we have

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{K} \lambda_{i,j} P(C_k|\boldsymbol{x})$$

Then the optimal decision rule is

$$\begin{cases} choose \; class \; i, & if \; R(\alpha_i|\boldsymbol{x}) = \min\{R(\alpha_j|\boldsymbol{x})\}_{j=1}^{K} < \lambda \\ reject \; \boldsymbol{x}, & if \; R(\alpha_i|\boldsymbol{x}) = \min\{R(\alpha_j|\boldsymbol{x})\}_{j=1}^{K} \geq \lambda \end{cases}$$

(ii) State the relationship between λ and the rejection threshold θ.
Solution:

Given that $\lambda_{i,k} = \begin{cases} 1, & if \; i \neq k \\ 0, & if \; i = k \end{cases}$ Then the corresponding expected loss is as following

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{K} \lambda_{i,j} P(C_k|\boldsymbol{x}) = 1 - P(C_i|\boldsymbol{x})$$

In order to get minimum expected loss, we hope:

- When $P(C_i|\boldsymbol{x}) > \theta$, we hope that $R(\alpha_i|\boldsymbol{x}) = \min\{R(\alpha_j|\boldsymbol{x})\}_{j=1}^K = 1 - P(C_i|\boldsymbol{x}) \leq \lambda$ so that we won't get expected loss any lower by manually choosing to reject $\boldsymbol{x}$.
  In order to achieve that, we need:
  $$\lambda \geq max(1 - P(C_i|\boldsymbol{x})) = 1 - \theta$$
- When $P(C_i|\boldsymbol{x}) \leq \theta$, we hope that $\lambda \leq R(\alpha_i|\boldsymbol{x}) = \min\{R(\alpha_j|\boldsymbol{x})\}_{j=1}^K = 1 - P(C_i|\boldsymbol{x})$ so that we won't get expected loss any lower by selecting to classifying $\boldsymbol{x}$ into class i.
  In order to achieve that, we need
  $$\lambda \leq min(1 - P(C_i|\boldsymbol{x})) = 1 - \theta$$

Based on that, we have
$$\lambda \geq 1 - \theta$$
$$\lambda \leq 1 - \theta$$

Therefore, we have
$$\lambda = 1 - \theta \Rightarrow \lambda + \theta = 1$$

2. [Parameter Estimation]

(a) Let $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T$ and $\boldsymbol{y} = [y_1, \cdots, y_n]^T$. Assume that $e_i \sim \mathcal{N}(0, \sigma^2)$ and $\boldsymbol{X^T X}$ is a full-rank matrix. Derive the maximum likelihood estimate $\boldsymbol{w}_{ML} = (\boldsymbol{X^T X}^{-1}\boldsymbol{X^T y})$.

Proof:

Given that $e_i \sim \mathcal{N}(0, \sigma^2)$ and $y_i = \boldsymbol{w^T x_i} + e_i$, we have that
$$y_i \sim \mathcal{N}(\boldsymbol{w^T x_i}, \sigma^2)$$

Thus, the likelihood function is
$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}) = \log(P(\boldsymbol{y|X, w})) &= -\sum_{i=1}^n \left(\log\sqrt{2\pi}\sigma + \frac{\left(y_i - \boldsymbol{w^T x_i}\right)^2}{2\sigma^2}\right) \\
&= -n\log\sqrt{2\pi}\sigma - \frac{\sum_{i=1}^n \left(y_i - \boldsymbol{w^T x_i}\right)^2}{2\sigma^2} \\
&= -n\log\sqrt{2\pi}\sigma - \frac{||\boldsymbol{y - Xw}||_2^2}{2\sigma^2} \\
\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}} &= -\frac{1}{2\sigma^2}\frac{\partial \boldsymbol{w^T X^T X w - 2 w^T X^T y + y^T y}}{\partial\boldsymbol{w}} \\
&= -\frac{1}{2\sigma^2}\left(2\boldsymbol{X^T X w} - 2\boldsymbol{X^T y}\right) = 0 \\
&\Rightarrow \boldsymbol{X^T X w = X^T y} \\
&\Rightarrow \boldsymbol{w}_{ML} = \left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{X^T y}
\end{aligned}$$

(b) Assume $\boldsymbol{w} \sim \mathcal{N}(0, \nu^2\boldsymbol{I})$. Derive the maximum a posteriori estimate
$$\boldsymbol{w}_{MAP} = \left(\boldsymbol{X^T X} + \frac{\sigma^2}{\nu^2}\boldsymbol{I}\right)^{-1}\boldsymbol{X^T y}$$

Proof:

Given that $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \nu^2\boldsymbol{I})$, we know that
$$P(\boldsymbol{w}) = \frac{1}{(2\pi)^{n/2}\nu}\exp[-\frac{1}{2}(\boldsymbol{w^T}\frac{1}{\nu^2}\boldsymbol{Iw})] = \frac{1}{(2\pi)^{n/2}\nu}\exp\left(-\frac{\boldsymbol{w^T w}}{2\nu^2}\right)$$

Then by definition, we know that

$$\begin{aligned}
\boldsymbol{w}_{MAP} &= \arg\max_{\boldsymbol{w}} P(\boldsymbol{w}|\mathcal{D}) \\
&= \arg\max_{\boldsymbol{w}} \frac{P(\mathcal{D}|\boldsymbol{w})P(\boldsymbol{w})}{P(\mathcal{D})} \\
&= \arg\max_{\boldsymbol{w}} P(\boldsymbol{y},\boldsymbol{X}|\boldsymbol{w})P(\boldsymbol{w}) \\
&= \arg\max_{\boldsymbol{w}} P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})P(\boldsymbol{X}|\boldsymbol{w})P(\boldsymbol{w})
\end{aligned}$$

Since the value of $\boldsymbol{X}$ does not rely on $\boldsymbol{w}$, we have $P(\boldsymbol{X}|\boldsymbol{w}) = P(\boldsymbol{X})$. Sequentially, we have

$$\boldsymbol{w}_{MAP} = \arg\max_{\boldsymbol{w}} P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})P(\boldsymbol{X})P(\boldsymbol{w}) = \arg\max_{\boldsymbol{w}} P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})P(\boldsymbol{w})$$

Next, let $\mathcal{L}(\boldsymbol{w}) = \log P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})P(\boldsymbol{w}) = \log P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}) + \log P(\boldsymbol{w})$, we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}) &= -n\log\sqrt{2\pi}\sigma - \frac{||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}||_2^2}{2\sigma^2} + \log{(2\pi)^{n/2}}\nu - \frac{\boldsymbol{w^T}\boldsymbol{w}}{2\nu^2} \\
\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}} &= -\frac{\boldsymbol{X^T}\boldsymbol{X}\boldsymbol{w}-\boldsymbol{X^T}\boldsymbol{y}}{\sigma^2} - \frac{\boldsymbol{w}}{\nu^2} = 0 \\
&\Rightarrow \left(\boldsymbol{X^T}\boldsymbol{X} + \frac{\sigma^2}{\nu^2}\boldsymbol{I}\right)\boldsymbol{w} = \boldsymbol{X^T}\boldsymbol{y}
\end{aligned}$$

Given that $\boldsymbol{X^T}\boldsymbol{X}$ is full-rank and $\frac{\sigma^2}{\nu^2}\boldsymbol{I}$ is positive-definite, we know that $\boldsymbol{X^T}\boldsymbol{X} + \frac{\sigma^2}{\nu^2}\boldsymbol{I}$ is invertible.

Therefore, we know that

$$\boldsymbol{w}_{MAP} = \left(\boldsymbol{X^T}\boldsymbol{X} + \frac{\sigma^2}{\nu^2}\boldsymbol{I}\right)^{-1}\boldsymbol{X^T}\boldsymbol{y}$$

3. [Parameter Estimation]
Solution:

Let $\boldsymbol{z}_l = sigmoid(\boldsymbol{W}_1\boldsymbol{x}_l + \boldsymbol{b}_1)$ and $\boldsymbol{f}_l = softmax(\boldsymbol{W_2}\boldsymbol{z}_l + \boldsymbol{b}_2)$, then

$$\begin{aligned}
\ell(\boldsymbol{W_1},\boldsymbol{W_2},\boldsymbol{b_1},\boldsymbol{b_2}) &= -\sum_{l=1}^{L}\boldsymbol{y}_l^T\log\boldsymbol{f}_l \\
&= -\sum_{l=1}^{L}\sum_{k=1}^{M} y_{l,k}\log f_{l,k}
\end{aligned}$$

where $f_{l,k} = softmax(\boldsymbol{w}_{2,k}^T\boldsymbol{z}_l + b_{2,k})$. Then, we have

$$\begin{aligned}
\frac{\partial\ell}{\partial\boldsymbol{w_{2,i}^T}\boldsymbol{z_l} + b_{2,i}} &= \frac{\partial\ell}{\partial f_{l,k}}\frac{\partial f_{l,k}}{\boldsymbol{w_{2,i}^T}\boldsymbol{z_l} + b_{2,i}} \\
&= -\sum_{l=1}^{L}\sum_{k=1}^{M}\frac{y_{l,k}}{f_{l,k}}f_{l,k}(\delta_{ki} - f_{l,i}) \\
&= -\sum_{l=1}^{L}\sum_{k=1}^{M}y_{l,k}(\delta_{ki} - f_{l,i}) \\
&= -\sum_{l=1}^{L}(y_{l,i} - f_{l,i}) \\
where\ \delta_{ki} &= \begin{cases} 1, & k = i \\ 0, & k \neq i \end{cases}
\end{aligned}$$

Since

$$\frac{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}}{\partial w_{2,i,j}} = z_{l,j}$$

$$\frac{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}}{\partial b_{2,i}} = 1$$

$$\frac{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}}{\partial w_{1,p,q}} = \sum_{j=1}^{H} w_{2,i,j} \frac{\partial z_{l,j}}{\partial w_{1,p,q}}$$

$$\frac{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}}{\partial b_{1,p}} = \sum_{j=1}^{H} w_{2,i,j} \frac{\partial z_{l,j}}{\partial b_{1,p}}$$

$$\frac{\partial z_{l,j}}{\partial w_{1,p,q}} = \begin{cases} z_{l,q}(1 - z_{l,q})x_{l,q},\ j = p \\ 0,\ j \neq q \end{cases}$$

$$\frac{\partial z_{l,j}}{\partial b_{1,p}} = \begin{cases} z_{l,q}(1 - z_{l,q}),\ j = p \\ 0,\ j \neq q \end{cases}$$

Where $z_{l,q} = sigmoid(\boldsymbol{w_{1,q}^T x_l} + b_{1,q})$.

Then, we have:

$$\begin{aligned} \Delta w_{2,i,j} &= -\eta \frac{\partial \ell}{\partial w_{2,i,j}} = -\eta \frac{\partial \ell}{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}} \frac{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}}{\partial w_{2,i,j}} \\ &= \eta \sum_{l=1}^{L} (y_{l,i} - f_{l,i}) z_{l,j} \end{aligned}$$

$$\begin{aligned} \Delta b_{2,i} &= -\eta \frac{\partial \ell}{\partial b_{2,i}} = -\eta \frac{\partial \ell}{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}} \frac{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}}{\partial b_{2,i}} \\ &= \eta \sum_{l=1}^{L} (y_{l,i} - f_{l,i}) \end{aligned}$$

$$\begin{aligned} \Delta w_{1,p,q} &= -\eta \frac{\partial \ell}{\partial w_{1,p,q}} = -\eta \sum_{i=1}^{M} \frac{\partial \ell}{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}} \frac{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}}{\partial w_{1,p,q}} \\ &= \eta \sum_{l=1}^{L} \left[\sum_{i=1}^{M} (y_{l,i} - f_{l,i}) w_{2,i,p} \right] z_{l,p}(1 - z_{l,p}) x_{l,q} \end{aligned}$$

$$\begin{aligned} \Delta b_{1,p} &= -\eta \frac{\partial \ell}{\partial b_{1,p}} = -\eta \sum_{i=1}^{M} \frac{\partial \ell}{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}} \frac{\partial \boldsymbol{w_{2,i}^T z_l} + b_{2,i}}{\partial b_{1,p}} \\ &= \eta \sum_{l=1}^{L} \left[\sum_{i=1}^{M} (y_{l,i} - f_{l,i}) w_{2,i,p} \right] z_{l,p}(1 - z_{l,p}) \end{aligned}$$

Where

$$f_{l,i} = softmax(\boldsymbol{w}_{2,i}^T \boldsymbol{z}_l + b_{2,i})$$
$$z_{l,p} = sigmoid(\boldsymbol{w}_{1,q}^T \boldsymbol{x}_l + b_{1,q})$$

4. [Bayesian Decision Theory, Linear Discrimination]

   (a) minimize the expected loss v.s. minimize the misclassification error.

   Solution:

   Let λ_{ij} be the loss for misclassifying $\boldsymbol{x}$ into class i while it actually belongs to class j, α_i be the action of classifying $\boldsymbol{x}$ into class i and $R(\alpha_i|\boldsymbol{x})$ be the expected loss of classifying $\boldsymbol{x}$ into class i.

   Therefore,

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{2} \lambda_{ij} P(C_j|\boldsymbol{x})$$

   Let $g(\boldsymbol{x}) = R(\alpha_1|\boldsymbol{x}) - R(\alpha_2|\boldsymbol{x})$, in order to minimize the expected loss, we have the following decision rule:

$$\begin{cases} choose\ C_1,\ if\ g(\boldsymbol{x}) < 0 \\ choose\ C_2,\ elsewhere \end{cases}$$

   Then the decision boundary is the solution of $g(\boldsymbol{x}) = 0$.

$$\begin{aligned}
& g(\boldsymbol{x}) = 0 \\
& \Downarrow \\
& (\lambda_{11} - \lambda_{21}) \frac{P(\boldsymbol{x}|C_1)P(C_1)}{P(\boldsymbol{x})} \\
& = (\lambda_{22} - \lambda_{12}) \frac{P(\boldsymbol{x}|C_2)P(C_2)}{P(\boldsymbol{x})} \\
& \Downarrow \\
& \log(|\lambda_{11} - \lambda_{21}|) + \log P(\boldsymbol{x}|C_1) + \log P(C_1) \\
& = \log(|\lambda_{22} - \lambda_{12}|) + \log P(\boldsymbol{x}|C_2) + \log P(C_2) \\
& \Downarrow \\
& \log(|\lambda_{11} - \lambda_{21}|) + \frac{1}{2\sigma^2}(\boldsymbol{2\mu_1^T x - \mu_1^T \mu_1}) + \log P(C_1) \\
& = \log(|\lambda_{22} - \lambda_{12}|) + \frac{1}{2\sigma^2}(\boldsymbol{2\mu_2^T x - \mu_2^T \mu_2}) + \log P(C_2) \\
& \Downarrow
\end{aligned}$$

$$\frac{1}{\sigma^2}(\boldsymbol{\mu_1^T} - \boldsymbol{\mu_2^T})\boldsymbol{x} = \log \frac{(\lambda_{22} - \lambda_{12})P(C_2)}{(\lambda_{11} - \lambda_{21})P(C_1)} + \frac{1}{2\sigma^2}(\boldsymbol{\mu_1^T \mu_1 - \mu_2^T \mu_2})$$

   Define $\boldsymbol{w} = \frac{1}{\sigma^2}(\boldsymbol{\mu_1 - \mu_2})$ and $w_0 = \log \frac{(\lambda_{11}-\lambda_{21})P(C_1)}{(\lambda_{22}-\lambda_{12})P(C_2)} - \frac{1}{2\sigma^2}(\boldsymbol{\mu_1^T \mu_1 - \mu_2^T \mu_2})$, we have

$$\boldsymbol{w^T x} + w_0 = \boldsymbol{w^T (x - x_0)} = 0$$

   where $\boldsymbol{x_0} = \frac{1}{2}(\boldsymbol{\mu_1 + \mu_2}) - \frac{1}{||\boldsymbol{\mu_1 - \mu_2}||_2^2} \sigma^2 \log \frac{(\lambda_{11}-\lambda_{21})P(C_1)}{(\lambda_{22}-\lambda_{12})P(C_2)} (\boldsymbol{\mu_1 - \mu_2})$

However, in order to minimize the probability of error, we define

$$\begin{aligned} g'(\boldsymbol{x}) &= \log P(C_1|\boldsymbol{x}) - \log P(C_2|\boldsymbol{x}) \\ &= \log P(\boldsymbol{x}|C_1) - \log P(\boldsymbol{x}|C_2) + \log \frac{P(C_1)}{P(C_2)} \\ &= \frac{1}{\sigma^2}(\boldsymbol{\mu_1^T} - \boldsymbol{\mu_2^T})\boldsymbol{x} - \frac{1}{2\sigma^2}(\boldsymbol{\mu_1^T\mu_1} - \boldsymbol{\mu_2^T\mu_2}) + \log\frac{P(C_1)}{P(C_2)} \\ &= \boldsymbol{w'^T x} + {w_0}' = \boldsymbol{w'^T(x - x_0{}')} = 0 \\ \Rightarrow \boldsymbol{x_0{}'} &= \frac{1}{2}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) - \frac{1}{||\boldsymbol{\mu_1} - \boldsymbol{\mu_2}||_2^2}\sigma^2 \log\frac{P(C_1)}{P(C_2)}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) \end{aligned}$$

The corresponding decision rule is:

$$\begin{cases} choose\ C_1,\ if\ g'(\boldsymbol{x}) > 0 \\ choose\ C_2,\ elsewhere \end{cases}$$

Based on that, we know

$$\begin{aligned} &\boldsymbol{w} = \boldsymbol{w'} \\ &x_0 = {x_0}' - \frac{1}{||\boldsymbol{\mu_1} - \boldsymbol{\mu_2}||_2}\sigma^2 \log\frac{\lambda_{11} - \lambda_{21}}{\lambda_{22} - \lambda_{12}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) \end{aligned}$$

Therefore, two decision boundaries have the same normal vector but different bias.

(b) By minimizing the misclassification error, obtain and draw the decision boundary when $\mu_{11} = 1, \mu_{12} = 1, \mu_{21} = 3, \mu_{22} = 5, \sigma = 1$ and $P(C_1) = P(C_2)$.

Solution:

By Bayesian rule, we have

$$P(C_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_i)P(C_i)}{P(\boldsymbol{x})}$$

Given that x_1, x_2 are independent following the Laplace distribution, we have

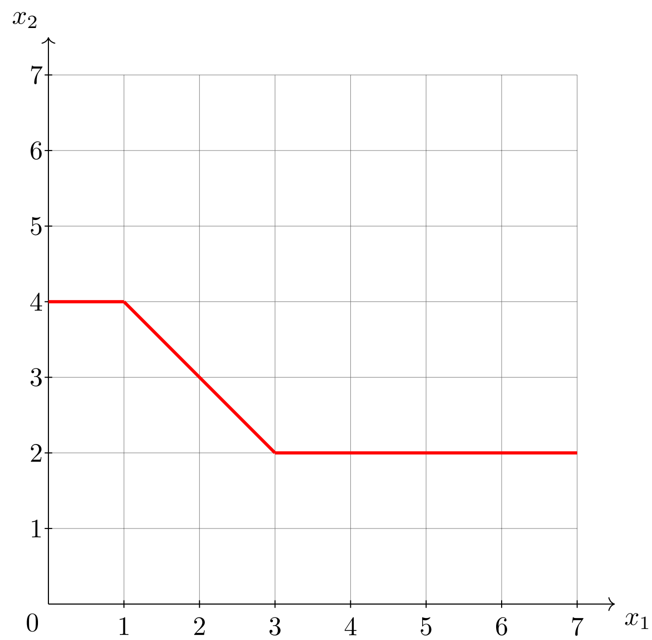
$$P(C_i|\boldsymbol{x}) = \frac{P(x_1|C_i)P(x_2|C_i)P(C_i)}{P(\boldsymbol{x})}$$

Then define discrimination function

$$\begin{aligned} g_i(\boldsymbol{x}) &= \log P(C_i|\boldsymbol{x}) \\ &= \log P(x_1|C_i) + \log P(x_2|C_i) + \log P(C_i) - \log P(\boldsymbol{x}) \\ &= -2\log 2\sigma - \frac{1}{\sigma}|x_1 - \mu_{i1}| - \frac{1}{\sigma}|x_2 - \mu_{i2}| + \log P(C_i) \\ g(\boldsymbol{x}) &= g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) \\ &= -|x_1 - 1| - |x_2 - 1| + |x_1 - 3| + |x_2 - 5| \end{aligned}$$

Then let $g(\boldsymbol{x} = 0)$, we get:

$$\begin{aligned} &-|x_1 - 1| - |x_2 - 1| + |x_1 - 3| + |x_2 - 5| = 0 \\ &\Rightarrow x_2 = \begin{cases} 4, & for\ x_1 < 1 \\ -x_1 + 5 = 0, & for\ 1 \le x_1 \ge 3 \\ 2, & for\ x_1 > 3 \end{cases} \end{aligned}$$



x_2
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
x_1