HW2 - Handwritten Assignment Answer

1.

Consider a generative classification model for K classes defined by prior class probabilities $p(C_k) = \pi_k$ and general class-conditional densities $p(x|C_k)$, where x is the input feature vector. Suppose we are given a training data set $\{x_n, \mathbf{t_n}\}$ where $n = 1, \ldots, N$,and $\mathbf{t_n}$ is a binary target vector of length K that uses the $1 - of - K$ coding scheme, so that it has components $t_{nk} = 1$ if pattern n is from class C_k, otherwise $t_{nj} = 0$. Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probablities is given by

$$\pi_k = \frac{N_k}{N}$$

where N_k is the number of data points assigned to class C_k.

ans1:

目標：求 $maximize \ \ P_c(x_1, x_2, \ldots, x_N)$
(期待可以準確判斷每個 x_i 最有可能所屬的class) (0.4pt)

假設 C_{x_i} 表示 x_i 所屬的類別，對於 $i = 1, 2, \ldots, N$

依據獨立假設性：

$$\begin{aligned} \text{令 } F &= P_c(x_1, x_2, \ldots, x_N) \\ &= P_c(x_1)P_c(x_2)\ldots P_c(x_N) \\ &= P(C_{x_1}, x_1)P(C_{x_2}, x_2)\ldots P(C_{x_N}, x_N) \\ &= P(C_{x_1})P(x_1|C_{x1})P(C_{x2})P(x_2|C_{x2})\ldots P(C_{xN})P(x_N|C_{xN}) \\ &= \prod_{i=1}^{N} P(C_{xi})P(x_i \mid C_{xi}) \end{aligned}$$

對 F 取log:

$f = logF = \sum_{i=1}^{N} logP(C_{xi}) \ + \ \sum_{i=1}^{N} logP(x_i \mid C_{xi})$
(0.2pt)
從題目中，我們知道每個類別C含有N_k個資料，對於$k = 1, \ldots, K$

$\sum_{x_i \in C_k} logP(x_i) = N_k logP(C_k)$

改寫 f 中的第一項：

$$\begin{aligned} \sum_{i=1}^{N} logP(C_{x_i}) &= \sum_{x \in C_1} logP(C_x) \ + \ \ldots \ + \sum_{x \in C_N} logP(C_x) \\ &= N_1 logP(C_1) \ + \ \ldots \ + N_k logP(C_k) \\ &= \sum_{k=1}^{K} N_k logP(C_k) \end{aligned}$$

重寫f：

$$f = logF = \sum_{k=1}^{K} N_k logP(C_k) + \sum_{i=1}^{N} logP(x_i \mid C_{x_i})$$

我們的目標是要 $maximize \, logF$
但由於有 $\sum_{k=1}^{K} \pi_k = 1$ 之條件，因此可以使用$lagrange \ multiplier$。

令 $g = \sum_{k=1}^{K} \pi_k - 1, \lambda \in \mathbb{R}^+$

$$L(\pi_1, \pi_2, \ldots, \pi_k, \lambda) = f + \lambda g$$

(0.2pt)
對特定的π_k做微分可得：

$$\frac{\partial L}{\partial \pi_k} = \frac{\partial f}{\partial \pi_k} + \frac{\partial g}{\partial \pi_k} = \frac{N_k}{\pi_k} + \lambda \overset{def}{=} 0$$
$$\pi_k = -\frac{N_k}{\lambda}$$

(0.1pt)
將 $\pi_k = -\frac{N_k}{\lambda}$ 代入 $\sum_{k=1}^{K} \pi_k = 1$，得到 $\lambda = -N$，故得 $\pi_k = \frac{N_k}{N}$。(0.1pt)

2.

Show that

$$\frac{\partial log(det\ \mathbf{\Sigma})}{\partial \sigma_{ij}} = \mathbf{e_j \Sigma^{-1}\ e_i^T}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{m \times m}$ is a (non-singular) covariance matrix and $\mathbf{e_j}$ is a row vector(ex: $e_3 = [0, 0, 1, 0, \ldots, 0]$).

ans2:

令對 $det\Sigma$ 第 i 列第 j 個 $node$ 作展開，其餘因子為 $[adj(\Sigma)]_{ji} = C_{ij}$

$$\begin{aligned}
\frac{\partial}{\partial \sigma_{ij}} log(det\Sigma) &= \frac{1}{det\Sigma} \frac{\partial\ det\Sigma}{\partial \sigma_{ij}} \\
&= \frac{1}{det\Sigma}[\sigma_{i1}C_{i1} + \ldots + \sigma_{ij}C_{ij} + \ldots \sigma_{im}C_{im}] \\
&= \frac{1}{det\Sigma} C_{ij} \\
&= \frac{1}{det\Sigma}[adj(\Sigma)]_{ji} \\
&= [\frac{1}{det\Sigma} adj(\Sigma)]_{ji} \\
&= [\Sigma^{-1}]_{ji} \\
&= \mathbf{e_j \Sigma^{-1}\ e_i^T}
\end{aligned}$$

3.

Consider the classification model of $\mathbf{problem\ 1}$ & result of $\mathbf{problem\ 2}$ and now suppose that the class-condition densities are given by Gaussian distributions with a shared convariance matrix, so that

$$p(x|C_k) = \mathcal{N}(x|\mathbf{\mu_k}, \mathbf{\Sigma})$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class C_k is given by

$$\mathbf{\mu_k} = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} x_n$$

which represents the mean of those feature vectors assigned to class C_k. Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\mathbf{\Sigma} = \sum_{k=1}^{K} \frac{N_k}{N} \mathbf{S_k}$$

where

$$\mathbf{S_k} = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} (\mathbf{x_n} - \mathbf{\mu_k})(\mathbf{x_n} - \mathbf{\mu_k})^T$$

Thus $\boldsymbol{\Sigma}$ is given by a weighted average of the covariance of the data associated with each class, in which the weighting coeffients are given by the prior probabilities of the classes.

ans3:

運用第一題的結果：

$$\begin{aligned} f &= log\ P_c(x_1, x_2, \ldots, x_N) \\ &= \sum_{k=1}^{K} N_k log\pi_k + \sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk} logP(x_n|C_k) \end{aligned}$$

假設$P(x|C_k)$是來自 μ_k, Σ 的Gaussian distribution

$$\begin{aligned} P(x|C_k) &= \frac{1}{\sqrt{(2\pi)^m det\Sigma}} exp(-\frac{1}{2}(\mu_k - x)^T\Sigma^{-1}(\mu_k - x)) \\ log\ P(x|C_k) &= -\frac{1}{2}(\mu_k - x)^T\Sigma^{-1}(\mu_k - x) - \frac{1}{2} log\ det\Sigma - \frac{m}{2} log\ 2\pi \end{aligned}$$

欲求 f 對 μ_k 做偏微分

$$\begin{aligned} \frac{\partial f}{\partial \mu_k} &= \frac{\partial}{\partial \mu_k} \sum_{k=1}^{K} N_k log\pi_k + \sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk} logP(x_n|C_k) \\ &= \frac{\partial}{\partial \mu_k} \sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk} logP(x_n|C_k) \\ &= \frac{\partial}{\partial \mu_k} \sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}(-\frac{1}{2}(\mu_k - x_n)^T\Sigma^{-1}(\mu_k - x_n) - \frac{1}{2} log\ det\Sigma - \frac{m}{2} log\ 2\pi) \\ &= \sum_{n=1}^{N} t_{nk}(\Sigma^{-1}(\mu_k - x_n)) \\ &= \Sigma^{-1}(\sum_{n=1}^{N} t_{nk}x_n - t_{nk}\mu_k) \\ &= \Sigma^{-1}[(\sum_{n=1}^{N} t_{nk}x_n) - N_k\mu_k] \overset{def}{=} 0 \\ &\Rightarrow \mu_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}x_n \end{aligned}$$

(0.5pt)
利用第二題的結果，協助我們對Σ^{-1}微分：

$$\frac{\partial}{\partial \Sigma^{-1}} log\ det\Sigma = (\Sigma^{-1})^T$$

但此結果還不夠漂亮，所以需再化簡：

$$\begin{aligned} \frac{\partial}{\partial \Sigma^{-1}} log\ det\Sigma &= \frac{\partial}{\partial \Sigma^{-1}} log\ \frac{1}{det\Sigma^{-1}} \\ &= -\frac{\partial}{\partial \Sigma^{-1}} log\ det\Sigma^{-1} \\ &= -((\Sigma^{-1})^{-1})^T = -\Sigma^T = -\Sigma \end{aligned}$$

欲求 f 對 Σ^{-1}做偏微分

$$
\begin{aligned}
\frac{\partial f}{\partial \Sigma^{-1}} &= \frac{\partial}{\partial \Sigma^{-1}} \sum_{k=1}^{K} \sum_{n=1}^{N} t_{nk} log P(x_n|C_k) \\
&= \sum_{k=1}^{K} \sum_{n=1}^{N} t_{nk}(-\frac{1}{2}(\mu_k - x_n)^T \Sigma^{-1}(\mu_k - x_n) - \frac{1}{2} log\ det\Sigma - \frac{m}{2} log\ 2\pi) \\
&= \sum_{k=1}^{K} \sum_{n=1}^{N} t_{nk}(-\frac{1}{2}(\mu_k - x_n)(\mu_k - x_n)^T + \frac{1}{2}\Sigma) \\
&= \frac{1}{2} \sum_{k=1}^{K} [\sum_{n=1}^{N} t_{nk}\Sigma - \sum_{n=1}^{N} t_{nk}(\mu_k - x_n)(\mu_k - x_n)^T] \\
&= \frac{1}{2}(\sum_{k=1}^{K} N_k \Sigma - \sum_{k=1}^{K} N_k S_k) \\
&= \frac{1}{2}(N\Sigma - \sum_{k=1}^{K} N_k S_k) \overset{def}{=} 0 \\
&\Rightarrow N\Sigma = \sum_{k=1}^{K} N_k S_k \\
&\Rightarrow \Sigma = \sum_{k=1}^{K} \frac{N_k}{N} S_k
\end{aligned}
$$

(0.5pt)