Problem Set 2

1 Gaussian Discriminant Analysis Model

Given m training data $\{x^{(i)}, y^{(i)}\}_{i=1,\cdots,m}$, assume that $y \sim Bernoulli(\psi)$, $x \mid y = 0 \sim \mathcal{N}(\mu_0, \Sigma)$, $x \mid y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$. Hence, we have

- $p(y) = \psi^y (1-\psi)^{1-y}$
- $p(x \mid y = 0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)\right)$
- $p(x \mid y = 1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)\right)$

The log-likelihood function is

$$
\begin{aligned}
\ell(\psi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \psi, \mu_0, \mu_1, \Sigma) \\
&= \log \prod_{i=1}^{m} p(x^{(i)} \mid y^{(i)}; \psi, \mu_0, \mu_1, \Sigma) p(y^{(i)}; \psi)
\end{aligned}
$$

Solve ψ, μ_0, μ_1 and Σ by maximizing $\ell(\psi, \mu_0, \mu_1, \Sigma)$. (Please refer to page 13 of Lecture 5, and use the results about *trace* presented in Lecture 2.)

Hint: If $y = tr(AX^{-1}B)$, then $\frac{dy}{dX} = -X^{-1}BAX^{-1}$

Solution: The log-likelihood function can be written as

$$\begin{aligned}
&\ell(\psi, \mu_0, \mu_1, \Sigma) \\
&= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \psi, \mu_0, \mu_1, \Sigma) \\
&= \log \prod_{i=1}^{m} p(x^{(i)} \mid y^{(i)}; \psi, \mu_0, \mu_1, \Sigma) p(y^{(i)}, \psi) \\
&= \sum_{i=1}^{m} \Big[\log p(x^{(i)} \mid y^{(i)}; \psi, \mu_0, \mu_1, \Sigma) + \log p(y^{(i)}, \psi)\Big] \\
&= \sum_{i=1}^{m} [-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) - \frac{n}{2}\log(2\pi) \\
&\quad - \frac{1}{2}\log|\Sigma| + y^{(i)} \log \psi + (1 - y^{(i)}) \log(1 - \psi)]
\end{aligned}$$

We calculate the derivatives of $\ell(\psi, \mu_0, \mu_1, \Sigma)$ with respect to ψ, and let it be zero.

$$\begin{aligned}
\frac{\partial}{\partial \psi} \ell(\psi, \mu_0, \mu_1, \Sigma) &= \frac{\partial}{\partial \psi} \sum_{i=1}^{m} [y^{(i)} \log \psi + (1 - y^{(i)}) \log(1 - \psi)] \\
&= \sum_{i=1}^{m} \left(\frac{y^{(i)}}{\psi} + \frac{1 - y^{(i)}}{1 - \psi} \right) \\
&= \sum_{i=1}^{m} \frac{y^{(i)} - \psi}{\psi(1 - \psi)} \\
&= 0
\end{aligned}$$

We thus have

$$\psi = \frac{\sum_{i=1}^{m} y^{(i)}}{m} = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = 1)}{m}$$

Since

$$\begin{aligned}
&\frac{\partial}{\partial \mu_0} \ell(\psi, \mu_0, \mu_1, \Sigma) \\
&= \frac{\partial}{\partial \psi} \sum_{i=1}^{m} \left[-\frac{1}{2} \mathbf{1}(y^{(i)} = 0)(x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0) \right] \\
&= \frac{\partial}{\partial \psi} \sum_{i=1}^{m} -\frac{1}{2} \mathbf{1}(y^{(i)} = 0) \cdot Tr \left(\mu_0^T \Sigma^{-1} \mu_0 - \mu_0^T \Sigma^{-1} x^{(i)} - (x^{(i)})^T \Sigma^{-1} \mu_0 \right) \\
&= \sum_{i=1}^{m} \mathbf{1}(y^{(i)} = 0) \Sigma^{-1} (x^{(i)} - \mu_0) \\
&= 0
\end{aligned}$$

we have

$$\mu_0 = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = 0) x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = 0)}$$

Similarly, we can calculate μ_1 as

$$\mu_1 = \frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = 1)x^{(i)}}{\sum_{i=1}^m \mathbf{1}(y^{(i)} = 1)}$$

By letting derivatives of $\ell(\psi, \mu_0, \mu_1, \Sigma)$ with respect to Σ be zero, we have

$$\begin{aligned}
&\nabla_\Sigma \, \ell(\psi, \mu_0, \mu_1, \Sigma) \\
&= \nabla_\Sigma \sum_{i=1}^m \left[-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) - \frac{1}{2}\log|\Sigma| \right] \\
&= \sum_{i=1}^m \nabla_\Sigma \left(-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right) - \sum_{i=1}^m \nabla_\Sigma \frac{1}{2}\log|\Sigma| \\
&= 0
\end{aligned}$$

where

$$\begin{aligned}
&\nabla_\Sigma \left(-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right) \\
&= -\frac{1}{2} \nabla_\Sigma \;\; tr\left((x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right) \\
&= \frac{1}{2}\Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1}
\end{aligned}$$

and

$$\begin{aligned}
&\nabla_\Sigma \frac{1}{2}\log|\Sigma| \\
&= \frac{1}{2|\Sigma|} \cdot |\Sigma|(\Sigma^{-1})^T \\
&= \frac{1}{2}(\Sigma^{-1})^T
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{i=1}^m \frac{1}{2}\Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} = \frac{m}{2}(\Sigma^{-1})^T \\
&\Rightarrow \sum_{i=1}^m \Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} = m(\Sigma^T)^{-1} \\
&\Rightarrow \sum_{i=1}^m \Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} = m\Sigma^{-1} \\
&\Rightarrow \sum_{i=1}^m \Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T = mI \\
&\Rightarrow \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T = m\Sigma \\
&\Rightarrow \Sigma = \frac{\sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T}{m}
\end{aligned}$$

2 MLE for Naive Bayes

Consider the following definition of **MLE problem for multinomials**. The input to the problem is a finite set $\mathcal{Y}$, and a weight $c_y \geq 0$ for each $y \in \mathcal{Y}$. The output from the problem is the distribution p^* that solves the following maximization problem.

$$p^* = \arg\max_{p \in \mathcal{P}_{\mathcal{Y}}} \sum_{y \in \mathcal{Y}} c_y \log p_y$$

Prove that, the vector p^* has components

$$p_y^* = \frac{c_y}{N}$$

for $\forall y \in \mathcal{Y}$, where $N = \sum_{y \in \mathcal{Y}} c_y$. (Hint: use the theory of Lagrange multiplier)

Using the above consequence, prove that, the maximum-likelihood estimates for Naive Bayes model are as follows

$$p(y) = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y)}{m}$$

and

$$p(x_j \mid y) = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y \wedge x_j^{(i)} = x)}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y)}$$

Solution: Our goal is to maximize the function

$$\sum_{y \in \mathcal{Y}} c_y \log p_y$$

subject to the constraints $p_y \geq 0$ and $\sum_{y \in \mathcal{Y}} p_y = 1$. We first prove the case of $c_y > 0$ for $\forall y \in \mathcal{Y}$.

We introduce a single Lagrange multiplier $\lambda \in \mathbb{R}$ corresponding to the constraint that $\sum_{y \in \mathcal{Y}} p_y = 1$. The Lagrangian is then

$$g(\lambda, p) = \sum_{y \in \mathcal{Y}} c_y \log p_y - \lambda \left(\sum_{y \in \mathcal{Y}} p_y - 1 \right)$$

According to the theory of Lagrange multiplier, the solution p^* to the maximization problem must satisfy the following condition

$$\frac{d}{dp_y} g(\lambda, p) = 0$$

for $\forall y$, and

$$\sum_{y \in \mathcal{Y}} p_y = 1$$

.

Differentiating with respect to p_y gives

$$\frac{d}{dp_y} g(\lambda, p) = \frac{c_y}{p_y} - \lambda$$

Letting the above derivative to zero, we have

$$p_y^* = \frac{c_y}{\lambda}$$

Recalling that $\sum_{y\in\mathcal{Y}} p_y = 1$, we have

$$p_y^* = \frac{c_y}{\sum_{y\in\mathcal{Y}} c_y} = \frac{c_y}{N}$$

We then prove that, for $\forall y \in \mathcal{Y}$, if $c_y = 0$, then $p_y^* = 0$. For $c_y = 0$, letting corresponding optimal solution $p_y^* > 0$ would decrease the objective function $g(\lambda, p)$, which is a contradiction to our goal of maximization.

We rewrite the log-likelihood function of the NB model as follows

$$\begin{aligned}
\ell(\Omega) &= \sum_{i=1}^{m} \log p(y^{(i)}) + \sum_{i=1}^{m}\sum_{j=1}^{n} \log p(x_j^{(i)} \mid y^{(i)}) \\
&= \sum_{y\in\mathcal{Y}} \mathit{count}(y) \log p(y) + \sum_{j=1}^{n}\sum_{y\in\mathcal{Y}}\sum_{x_j\in\{0,1\}} \mathit{count}_j(x \mid y) \log p(x_j \mid y)
\end{aligned}$$

where

$$\begin{aligned}
\mathit{count}(y) &= \sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y) \\
\mathit{count}_j(x \mid y) &= \sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y \wedge x_j^{(i)} = x)
\end{aligned}$$

Maximizing the above equation with respect to $p(y)$ is equivalent to maximizing

$$\sum_{y\in\mathcal{Y}} \mathit{count}(y) \log p(y)$$

subject to the constraints $p(y) \geq 0$ and $\sum_{y=1}^{k} p(y) = 1$, since the second item (in the above equation) does not depend on $p(y)$. Therefore, according to the consequence we obtained before, we have

$$p(y) = \frac{\mathit{count}(y)}{\sum_{y=1}^{k} \mathit{count}(y)} = \frac{\mathit{count}(y)}{m} = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y)}{m}$$

By a similar argument, we can maximize the log-likelihood function with respect to $p(x_j \mid y)$. As a result, our goal becomes

$$\begin{aligned}
\text{maximize} \quad & \sum_{x_j\in\{0,1\}} \mathit{count}_j(x \mid y) \log p(x_j \mid y) \\
\text{s.t.} \quad & \sum_{x_j\in\{0,1\}} \log p(x_j \mid y) = 1
\end{aligned}$$

Therefore, we have the optimal solution

$$p(x_j \mid y) = \frac{\mathit{count}_j(x \mid y)}{\sum_{x_j\in\{0,1\}} \mathit{count}_j(x \mid y)} = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y \wedge x_j^{(i)} = x)}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y)}$$

Appendix

1. Prove that if $y = tr(AX^{-1}B)$, then $\frac{dy}{dX} = -X^{-1}BAX^{-1}$.

We first derive the expression for $d(X^{-1})$. Since $X^{-1}X = I$, we have

$$d(X^{-1}X) = d(X^{-1})X + X^{-1}dX = 0$$

Thus, we get

$$d(X^{-1}) = -X^{-1}(dX)X^{-1}$$

For $y = tr(AX^{-1}B)$,

$$\begin{aligned} dy &= d(tr(AX^{-1}B)) \\ &= d(tr(A(-X^{-1}(dX)X^{-1})B)) \\ &= tr((-X^{-1}BAX^{-1})(dX)) \end{aligned}$$

Hence, we have

$$\frac{dy}{dX} = -X^{-1}BAX^{-1}$$