EE M146 Discussion 5 Solution
Introduction to Machine Learning Monday, April 5, 2019
Instructor: Lara Dolecek
TA: Zehui (Alex) Chen, Ruiyi (John) Wu

1. Discriminative v.s. Generative So far, we have learned two approaches for binary classification in class. The generative approach model the prior $P(C_i)$ and class conditional distribution $P(x|C_i)$. The discriminative approach model $P(C_i|x)$ directly. Taking the Gaussian Discriminant Analysis model as an example. It models the class conditional distribution with two mean vectors μ_0, μ_1 and a shared covariance matrix Σ. In class, you learned that the resulting posterior probability for each class can be written as a logistic sigmoid function on a linear function: $P(C_i|x) = \sigma(w^T x)$.

   (a) If $x \in R^M$, how many parameters do we need for logistic regression?
   Solution: $2M + 1$.

   (b) How many parameters do we need for GDA model?
   Solution:$1 + 2M + M(M+1)/2$.

   (c) How many parameters do we need for GDA with Naive Bayes assumption?
   Solution:$1 + 3M$.

2. We are given a training set $\{(x^{(i)}, y^{(i)}); i = \{1, \cdots, m\}\}$, where $x^{(i)} \in R^n$ and $y^{(i)} \in \{0, 1\}$. We consider the Gaussian Discriminant Analysis (GDA) model, which models $P(x|y)$ using multivariate Gaussian. Writing out the model, we have:

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

   The log-likelihood of the data is given by:

$$L(\phi, \mu_0, \mu_1, \Sigma) = \ln P(x^{(i)}, \cdots, x^{(m)}, y^{(i)}, \cdots, y^{(m)}) = \ln \prod_{i=1}^{m} P(x^{(i)}|y^{(i)})P(y^{(i)}).$$

   In this exercise, suppose we already find μ_0 and μ_1, we want to maximize $L(\phi, \mu_0, \mu_1, \Sigma)$ with respect to Σ.

   (a) Write down the explicit expression for $P(x^{(i)}, \cdots, x^{(m)}, y^{(i)}, \cdots, y^{(m)})$ and $L(\phi, \mu_0, \mu_1, \Sigma)$.
   Solution:

$$\begin{aligned}
&P(x^{(i)}, \cdots, x^{(m)}, y^{(i)}, \cdots, y^{(m)} \\
=&\prod_{i=1}^{m}\left[\frac{1-\phi}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)\right]^{1-y^{(i)}} \\
&\times\left[\frac{\phi}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)\right]^{y^{(i)}}
\end{aligned}$$

$$
\begin{aligned}
&L(\phi, \mu_0, \mu_1, \Sigma) \\
&= \sum_{i=1}^{m} \Bigg\{ (1 - y^{(i)}) \left[\ln(1-\phi) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0) \right] \\
&\quad + y^{(i)} \left[\ln(\phi) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^{(i)} - \mu_1)^T \Sigma^{-1} (x^{(i)} - \mu_1) \right] \Bigg\}.
\end{aligned}
$$

(b) Differentiate $L(\phi, \mu_0, \mu_1, \Sigma)$ with respect to Σ and set it to 0. Show that the maximum likelihood result for Σ is:

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.$$

Hints: You may use the following properties without proof: $a = Tr(a)$ for scalar a; $Tr(A) + Tr(B) = Tr(A+B)$; $\frac{\partial \ln |A|}{\partial A} = A^{-T}$; $\frac{\partial Tr(A^{-1}B)}{\partial A} = -(A^{-1}BA^{-1})^T$.
Solution: We pick out the terms in $L(\phi, \mu_0, \mu_1, \Sigma)$ and treat other terms as constant:

$$
\begin{aligned}
L(\phi, \mu_0, \mu_1, \Sigma) &= -\frac{m}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) + const \\
&= -\frac{m}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{m} Tr\left((x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})\right) + const \\
&= -\frac{m}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{m} Tr\left(\Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T\right) + const \\
&= -\frac{m}{2}\ln(|\Sigma|) - \frac{m}{2} Tr\left(\Sigma^{-1}\frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T\right) + const \\
&= -\frac{m}{2}\ln(|\Sigma|) - \frac{m}{2} Tr\left(\Sigma^{-1} S\right) + const,
\end{aligned}
$$

where

$$S = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.$$

We then take the derivative with respect to Σ and set to 0:

$$-\Sigma^{-T} + (\Sigma^{-1} S \Sigma^{-1})^T = 0.$$

We find $\Sigma = S$ which is the desired result.