ECE M146 Homework 6 Solution

Introduction to Machine Learning

Instructor: Lara Dolecek

TA: Zehui (Alex) Chen, Ruiyi (John) Wu

Please upload your homework to Gradescope by May 21, 11:59 pm.
Please submit a single PDF directly on Gradescope
You may type your homework or scan your handwritten version. Make sure all the work is discernible.

1. The pdf for two jointly Gaussian random variables X and Y is of the following form parameterized by the scalars m_1, m_2, σ_1, σ_2 and ρ_{XY}:

$$f_{X,Y}(x,y) = \frac{\exp\left\{\frac{-1}{2(1-\rho_{XY}^2)}\left[(\frac{x-m_1}{\sigma_1})^2 - 2\rho_{XY}\left(\frac{x-m_1}{\sigma_1}\right)\left(\frac{y-m_2}{\sigma_2}\right) + (\frac{y-m_2}{\sigma_2})^2\right]\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{XY}^2}}. \tag{1}$$

The pdf for multivariate jointly Gaussian random variable $Z \in \mathbb{R}^k$ is of the following form parameterized by $\mu \in \mathbb{R}^k$ and $\Sigma \in \mathbb{R}^{k\times k}$.

$$f_Z(z) = \frac{\exp\left\{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right\}}{\sqrt{(2\pi)^k|\Sigma|}}. \tag{2}$$

Suppose $Z = [X,Y]^T$, i.e., $z = [x,y]^T$.

(a) Find μ, Σ^{-1} and Σ in terms of m_1, m_2, σ_1, σ_2 and ρ_{XY}.
Solution: We find the following result by directly comparing (1) and (2):

$$\mu = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix},$$

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho_{XY}^1)}\begin{bmatrix} \sigma_1^2 & -\rho_{XY}\sigma_1\sigma_2 \\ -\rho_{XY}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{XY}\sigma_1\sigma_2 \\ \rho_{XY}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

One can verify that by plugging the above expressions into (2), we get (1) back.

(b) Suppose $\rho_{XY} = 0$, what is Σ in this case? Can you write $f_{X,Y}(x,y)$ as the product of two single variate Gaussian distributions? Are X and Y independent?
Solution:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}.$$

$$f_{X,Y}(x,y) = \frac{\exp\left\{-\frac{1}{2}(x-m_1)^2\right\}}{\sqrt{2\pi}\sigma_1} \times \frac{\exp\left\{-\frac{1}{2}(y-m_2)^2\right\}}{\sqrt{2\pi}\sigma_2}. \tag{3}$$

X and Y are independent by definition.

2. The Gaussian Discriminant Analysis (GDA) models the class conditional distribution as multivariate Gaussian, i.e, $P(X|Y) \sim \mathcal{N}(\mu_Y, \Sigma)$. Suppose we want to enforce the **Naive Bayes (NB) assumption**, i.e. $P(X_i|Y, X_j) = P(X_i|Y), \forall j \neq i$, to GDA. Show that all off diagonal elements of Σ equal to 0: $\Sigma_{i,j} = 0, \forall i \neq j$ with the **NB assumption**.

   Solution: By definition:

$$
\begin{aligned}
\Sigma_{i,j} &= E\left[(X_i|Y - E[X_i|Y])\,(X_j|Y - E[X_j|Y])\right] \\
&= E\left[X_iX_j|Y + E[X_i|Y]E[X_j|Y] - E[X_i|Y]X_j|Y - X_i|YE[X_j|Y]\right] \\
&= 2E[X_i|Y]E[X_j|Y] - 2E[X_i|Y]E[X_j|Y] \\
&= 0.
\end{aligned}
$$

   The second last step comes from the NB assumption.

3. Consider the classification problem for two classes, C_0 and C_1. In the generative approach, we model the class-conditional distribution $P(x|C_0)$ and $P(x|C_1)$, as well as the class priors $P(C_0)$ and $P(C_1)$.The posterior probability for class C_0 can be written as

$$P(C_0|x) = \frac{P(x|C_0)P(C_0)}{P(x|C_0)P(C_0) + P(x|C_1)P(C_1)}.$$

(a) Show that $P(C_0|x) = \sigma(a)$ where $\sigma(a)$ is the *sigmoid* function defined by

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

Find a in terms of $P(x|C_0)$, $P(x|C_1)$, $P(C_0)$ and $P(C_1)$.
Solution:

$$a = \ln \frac{P(x|C_0)P(C_0)}{P(x|C_1)P(C_1)}.$$

(b) In the GDA model, we have the class conditional distribution as follows

$$P(x|C_0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T\Sigma^{-1}(x - \mu_0)\right),$$

$$P(x|C_1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T\Sigma^{-1}(x - \mu_1)\right).$$

Suppose we are able to find the maximum likelihood estimation of $\mu_0, \mu_1, \Sigma, P(C_0)$, and $P(C_1)$. Show that $a = w^T x + b$ for some w and b. Find w and b in terms of $\mu_0, \mu_1, \Sigma, P(C_0)$, and $P(C_1)$. This shows that the decision boundary is linear.
Solution: We plug the class conditional distribution into the equation of a in (a). Simplify the equation and we have

$$a = \ln \frac{P(C_0)}{P(C_1)} + x^T\Sigma^{-1}\mu_0 - x^T\Sigma^{-1}\mu_1 - \frac{\mu_0^T\Sigma^{-1}\mu_0}{2} + \frac{\mu_1^T\Sigma^{-1}\mu_1}{2}.$$

From above, we identify:

$$w = \Sigma^{-1}\mu_0 - \Sigma^{-1}\mu_1;$$

and

$$b = \ln \frac{P(C_0)}{P(C_1)} - \frac{\mu_0^T\Sigma^{-1}\mu_0}{2} + \frac{\mu_1^T\Sigma^{-1}\mu_1}{2}.$$

This can be interpreted as a special case for the solution of (c).

(c) In (b), we modeled the class conditional distribution with same covariance matrix Σ. Now let us consider two classes that have difference covariance matrix as follows

$$P(x|C_0) = \frac{1}{(2\pi)^{n/2}|\Sigma_0|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T\Sigma_0^{-1}(x - \mu_0)\right),$$

$$P(x|C_1) = \frac{1}{(2\pi)^{n/2}|\Sigma_1|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right).$$

Suppose we are able to find the maximum likelihood estimation of $\mu_0, \mu_1, \Sigma_0, \Sigma_1, P(C_0)$, and $P(C_1)$. Show that $a = x^TAx + w^Tx + b$ for some A, w and b. Find w and b in terms of $\mu_0, \mu_1, \Sigma_0, \Sigma_1, P(C_0)$, and $P(C_1)$. This shows that the decision boundary is quadratic.

Solution: We plug the class conditional distribution into the equation of a in (a). Simplify the equation and we have

$$a = \ln\frac{P(C_0)}{P(C_1)} + \ln\frac{|\Sigma_1|^{1/2}}{|\Sigma_0|^{1/2}} - \frac{1}{2}x^T\Sigma_0^{-1}x + \frac{1}{2}x^T\Sigma_1^{-1}x + x^T\Sigma_0^{-1}\mu_0 - x^T\Sigma_1^{-1}\mu_1 - \frac{\mu_0^T\Sigma_0^{-1}\mu_0}{2} + \frac{\mu_1^T\Sigma_1^{-1}\mu_1}{2}.$$

From above, we identify:

$$A = \frac{1}{2}\Sigma_1^{-1} - \frac{1}{2}\Sigma_0^{-1};$$

$$w = \Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1;$$

and

$$b = \ln\frac{P(C_0)}{P(C_1)} + \ln\frac{|\Sigma_1|^{1/2}}{|\Sigma_0|^{1/2}} - \frac{\mu_0^T\Sigma_0^{-1}\mu_0}{2} + \frac{\mu_1^T\Sigma_1^{-1}\mu_1}{2}.$$

4. 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, we want to maximize $L(\phi, \mu_0, \mu_1, \Sigma)$ with respect to ϕ, μ_0. The maximization over Σ is left for discussion.

(a) Write down the explicit expression for $P(x^{(1)}, \cdots, x^{(m)}, y^{(1)}, \cdots, y^{(m)})$ and $L(\phi, \mu_0, \mu_1, \Sigma)$.
Solution:

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

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

(b) Find the maximum likelihood estimate for ϕ. How do you know such ϕ is the "best" but not the "worst"? Hint: Show that the second derivative of $L(\phi, \mu_0, \mu_1, \Sigma)$ with respect to ϕ is negative.
Solution: We only care about the terms that contains ϕ and treat other terms as constant:

$$L(\phi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{m} \{y^{(i)} \ln(\phi) + (1 - y^{(i)}) \ln(1-\phi)\} + const.$$

We set the derivative to 0:

$$\frac{\partial L}{\partial \phi} = \frac{N_1}{\phi} - \frac{N_0}{1-\phi} = 0.$$

where $N_1 = \sum_{i=1}^m y^{(i)}$ and $N_0 = \sum_{i=1}^m (1 - y^{(i)})$. We find $\phi = \frac{N_1}{N_0+N_1}$. Why not the "worst"? We take the second derivative.

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

This shows that the log likelihood function is concave with respect to ϕ and therefore have a unique maximum.

(c) Find the maximum likelihood estimate for μ_0. How do you know such μ_0 is the "best" but not the "worst"? Hint: Show that the Hessian Matrix of $L(\phi, \mu_0, \mu_1, \Sigma)$ with respect to μ_0 is negative definite. You may use the following: if A is positive definite, then A^{-1} is also positive definite. Also B is negative definite if $-B$ is positive definite.

Solution: We only care about the terms that contains μ_0 and treat other terms as constant:

$$\begin{aligned} L(\phi, \mu_0, \mu_1, \Sigma) &= \sum_{i=1}^m \{-\frac{1}{2}(1 - y^{(i)})(x^{(i)} - \mu_0)^T \Sigma^{-1}(x^{(i)} - \mu_0)\} + const \\ &= -\sum_{i=1}^m [(1 - y^{(i)})(-\mu_0^T \Sigma^{-1} x^{(i)} + \frac{1}{2}\mu_0^T \Sigma^{-1} \mu_0)] + const. \end{aligned}$$

. Taking the gradient with respect to μ_0:

$$\nabla_{\mu_0} J = -\sum_{i=1}^m [(1 - y^{(i)})(-\Sigma^{-1} x^{(i)} + \Sigma^{-1} \mu_0)].$$

Setting the gradient to 0, we get

$$\mu_0 = \frac{1}{N_0} \sum_{i=1}^m (1 - y^{(i)}) x^{(i)}.$$

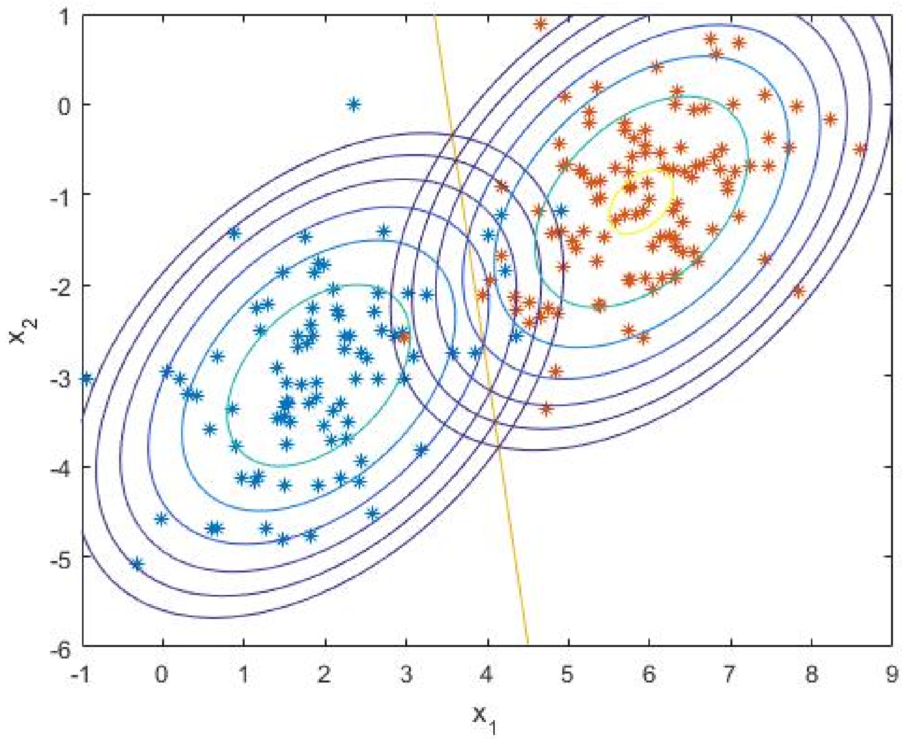
Why not "worst"? Let us calculate the Hessian matrix

$$\nabla^2_{\mu_0} J = -N_0 \Sigma^{-1}.$$

We know Σ is positive definite thus Σ^{-1} is also positive definite. The Hessian matrix is negative definite therefore there is a unique maximum.

5. In this exercise, you will implement a binary classifier using the Gaussian Discriminant Analysis (GDA) model in MATLAB. The data is given in *data.csv*. The first two columns are the feature values and the last column contains the class labels.

   (a) Visualization. Plot the data from different classes in different colors. Is the data linearly separable?



   Solution: Not linearly separable.

   (b) In the GDA model, we assume the class label follows a Bernoulli distribution and we model the class conditional distribution as multivariate Gaussian with same covariance matrix (Σ) and different means (μ_0 and μ_1). Find the maximum likelihood estimate of the parameters $P(y = 0)$ (parameter for the Bernoulli distribution), μ_0, μ_1 and Σ given this data set.
   Solution:

$$P(y=0) = 0.445, \mu_0 = \begin{bmatrix} 1.9195 \\ -2.9972 \end{bmatrix}, \mu_1 = \begin{bmatrix} 5.8982 \\ -1.0793 \end{bmatrix}, \Sigma = \begin{bmatrix} 1.0181 & 0.3887 \\ 0.3887 & 0.8036 \end{bmatrix}.$$

   (c) Using the result you find in Question 3 and your ML estimate of model parameters, find the decision boundary parameterized by $w^T x + b = 0$. Report w, b and plot the decision boundary on the same plot.
   Solution:

$$w = \begin{bmatrix} -3.6755 \\ -0.6090 \end{bmatrix}, b = 12.9050.$$

(d) Visualize your results by plotting the contour of the two distributions $P(x, y = 0)$ and $P(x, y = 1)$. For consistency, set 'LevelList' ('level' for python) to logspace(-3,-1,7). Does your decision boundary pass through the points where the two distributions have equal probabilities ? Explain why.

Solution:

$P(x, y = 0) = P(x, y = 1)$ implies $P(y = 0|x) = P(y = 1|x)$. Therefore, the equal probability points on the plot correspond to the equal probability points for the two posterior distribution which is on the decision boundary defined by $w^T x + b = 0$.