

1. 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  $\Sigma$  equals to 0:  $\Sigma_{i,j} = 0, \forall i \neq j$  with the **NB assumption**.

2. 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  $P(C_1)$ .

- (b) In 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.

- (c) In (b), we model the class conditional distribution with same covariance matrix  $\Sigma$ . 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  $a = x^T A x + w^T x + b$  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.

3. We are given a training set  $\{(x^{(i)}, y^{(i)}); i = \{1, \dots, m\}\}$ , 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:

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

In this exercise, we want to maximize  $L(\phi, \mu_0, \mu_1, \Sigma)$  with respect to  $\phi, \mu_0$ . The maximization over  $\Sigma$  is left for discussion.

- (a) Write down the explicit expression for  $P(x^{(1)}, \dots, x^{(m)}, y^{(1)}, \dots, y^{(m)})$  and  $L(\phi, \mu_0, \mu_1, \Sigma)$ .
- (b) Find the maximum likelihood estimate for  $\phi$ . How do you know such  $\phi$  is the “best” but not the “worst”? Hint: Show that the second derivative of  $L(\phi, \mu_0, \mu_1, \Sigma)$  with respect to  $\phi$  is negative.
- (c) Find the maximum likelihood estimate for  $\mu_0$ . How do you know such  $\mu_0$  is the “best” but not the “worst”? Hint: Show that the Hessian Matrix of  $L(\phi, \mu_0, \mu_1, \Sigma)$  with respect to  $\mu_0$  is negative definite. You may use the following: if  $A$  is positive definite, then  $A^{-1}$  is also positive definite.

4. 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?
  - (b) In GDA model, we assume the class label follow a Bernoulli distribution and we model the class conditional distribution as multivariate Gaussian with same covariance matrix ( $\Sigma$ ) and different means ( $\mu_0$  and  $\mu_1$ ). Find the maximum likelihood estimate of the parameters  $P(y = 0)$ ,  $\mu_0$ ,  $\mu_1$  and  $\Sigma$  given this data set.
  - (c) Using the result you find in Question 2 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.
  - (d) Visualize your results by plotting the contour of the two distributions  $P(x, y = 0)$  and  $P(x, y = 1)$ . For consistency, use `contour(X1,X2,Your Joint Probability Matrix,'LevelList', logspace(-3,-1,7))`. Your decision boundary should pass through points where the two distribution have equal probabilities. Explain why?

5. Suppose we have a data set  $\{x_1, \dots, x_N\}$  where  $x_n \in \mathbf{R}^M$  and our goal is to partition the data set into  $K$  clusters with  $\mu_k$  representing the center of the  $k$ -th cluster. Recall that in K-means clustering we are attempting to minimize an objective function defined as follows:

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|_2^2,$$

where  $r_{nk} \in \{0, 1\}$  and  $r_{nk} = 1$  only if  $x_n$  is assigned to cluster  $k$ .

- (a) What is the minimum value of the objective function when  $K = n$  (the number of clusters equals to the number of samples)?
- (b) Adding a regularization term, the objective function now becomes:

$$J = \sum_{k=1}^K \left[ \lambda \|\mu_k\|_2^2 + \sum_{n=1}^N r_{nk} \|x_n - \mu_k\|_2^2 \right].$$

Consider the optimization of  $\mu_k$  with all  $r_{nk}$  known. Find the optimal  $\mu_k$  for

$$\operatorname{argmin}_{\mu_k} \lambda \|\mu_k\|_2^2 + \sum_{n=1}^N r_{nk} \|x_n - \mu_k\|_2^2.$$