HWK 1 - Graphical Models

Tong ZHAO (tong.zhao@eleves.enpc.fr)

1 Exercise 1 - Learning in discrete graphical models

Given an i.i.d. sample of N observations $\{(z_i, x_i)|i \in \{1, \dots, n\}\}$, we use the maximum likelihood estimator to estimate π and θ. The log likelihood functions and their derivatives can be found in the last section. Here we show directly the result:

The MLE estimators for π is: $\hat{\pi_m} = \dfrac{N_m}{N}$ for $m \in \{1, \dots, M\}$.

The MLE estimators for θ is: $\hat{\theta}_{mk} = \dfrac{N_{mk}}{N_m}$ for $m \in \{1, \dots, M\}$ and $k \in \{1, \dots, K\}$.

2 Exercise 2.1(a) - LDA formulas

Using the result of Exercise 1, we can deduce that $\hat{\pi} = \dfrac{N_1}{N} = \dfrac{\sum_{i=1}^N y_i}{N}$.

Then we calculate the log-likelihood function. The result is obtained by setting its derivative to 0. Detailed proof can be found in the last section.

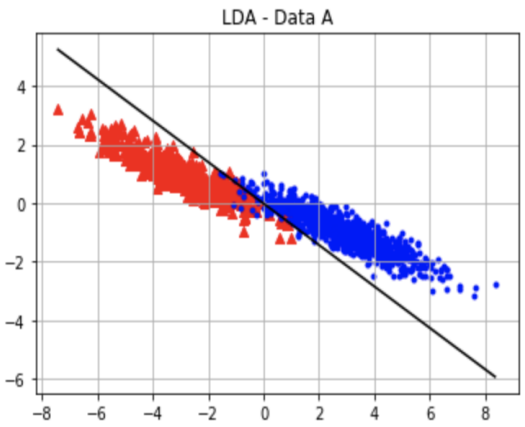
$$\begin{cases} \mu_0 = \dfrac{\sum_{i=1}^N (1-y_i)x_i}{N} \\ \mu_1 = \dfrac{\sum_{i=1}^N y_i x_i}{N} \\ \Sigma = \dfrac{\sum_{i=1}^N (x_i - \mu_{y_i})^T (x_i - \mu_{y_i})}{N} = \dfrac{\sum_{i=1}^N (x_i - (1-y_i)\mu_0 - y_i\mu_1)^T (x_i - (1-y_i)\mu_0 - y_i\mu_1)}{N} \end{cases}$$

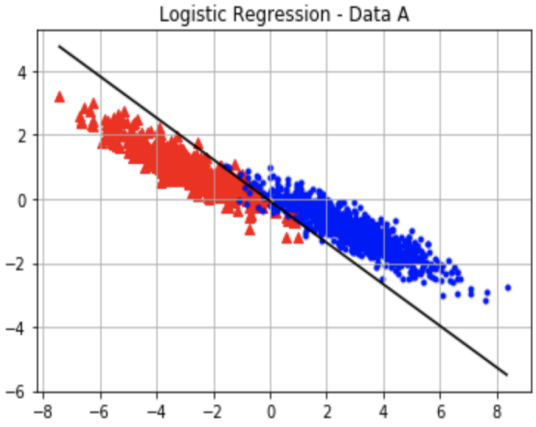
3 Exercise 2.5(a) QDA formulas

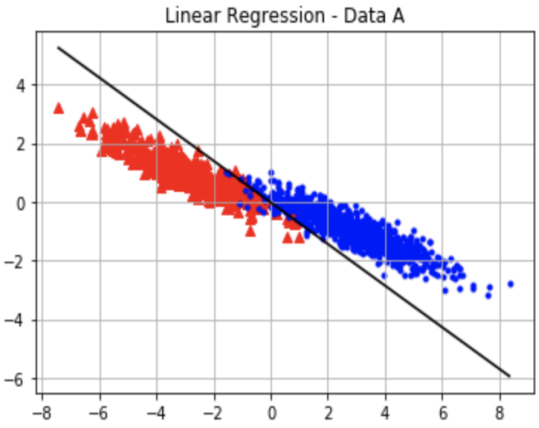
QDA model is a generalized version of LDA model, where different classes are allowed to have different covariance matrices, namely $\Sigma_0 \neq \Sigma_1$. While the maximum likelihood estimators of π, μ_0, μ_1 remain the same, the one of Σ_0 and Σ_1 can be expressed as following:

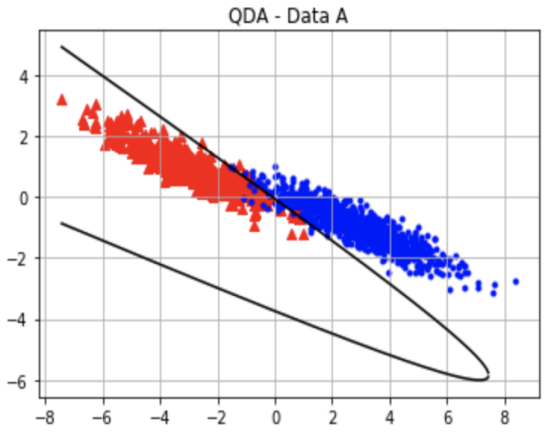
$$\begin{cases} \Sigma_0 = \dfrac{\sum_{i=1}^N (1-y_i)(x_i - \mu_0)^T (x_i - \mu_0)}{N - \sum_{i=1}^N y_i} \\ \Sigma_1 = \dfrac{\sum_{i=1}^N y_i (x_i - \mu_1)^T (x_i - \mu_1)}{\sum_{i=1}^N y_i} \end{cases}$$

4 Dataset A









Error Table

Misclassification Error	Train	Test
LDA	0.013	0.02
QDA	0.0067	0.0187
Logistic Regression	0.0067	0.0207
Linear Regression	0.013	0.0207

Comments

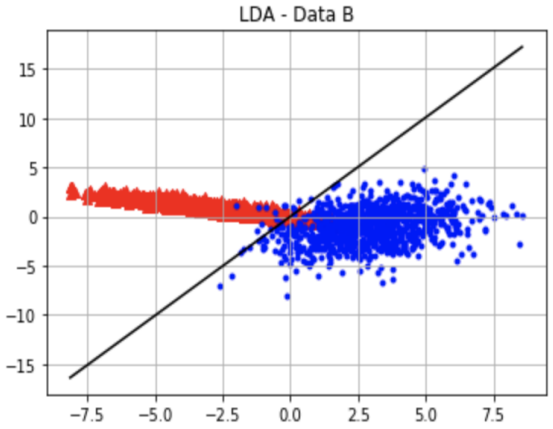
In the first case, the two classes are generated seperately from two independant gaussian distributions with the same variance. It fits very well the assumptions of LDA. We can observe from the table that its gap between the misclassification error on training dataset and the one on test dataset is the smallest.

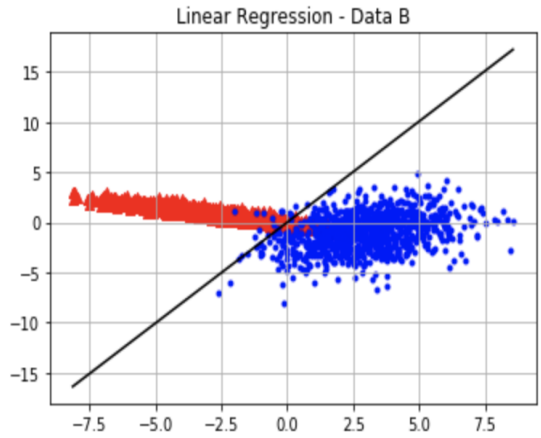
The QDA model has a looser assumption compared with LDA model. Hence it finds a better seperation on the training dataset. However from the figure we can see that the model overfits the dataset.

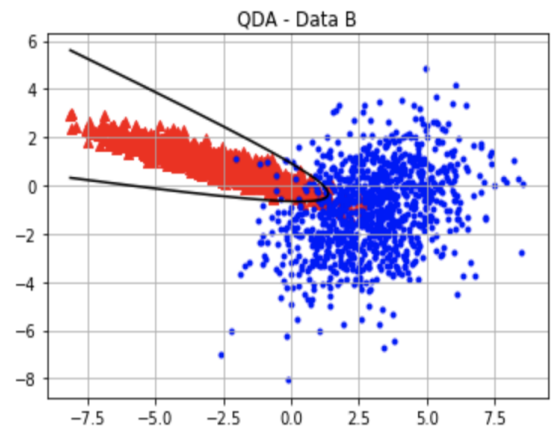
For all the 4 models, the gaps between training error and test error are obvious. It is mainly due to the distribution of the dataset. When we visualize the training dataset, we can find that it is almost linear seperable, but the test dataset is not.

We mention that the three linear models find almost the same boundary.

5 Dataset B







Error Table

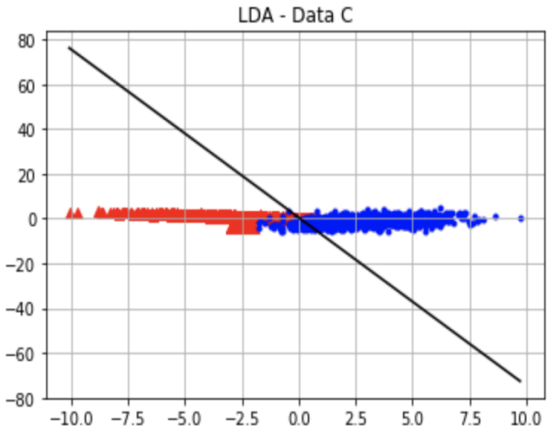
Misclassification Error	Train	Test
LDA	0.03	0.0415
QDA	0.023	0.0235
Logistic Regression	0.02	0.0395
Linear Regression	0.03	0.0415

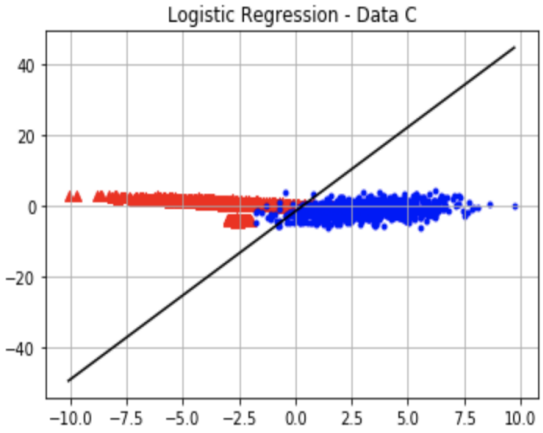
Comments

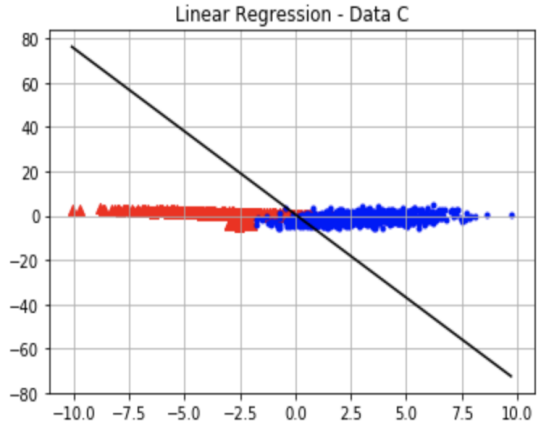
Different from the previous dataset, this one samples two classes of points seperately from two guassian distributions with different variances. This assumption best fits the one of the QDA model. From the figure and the table, we can also see the efficiency of QDA model. What's more, its gap between the train error and the test error is tiny, this proves that the proper choice of model is essential in practice.

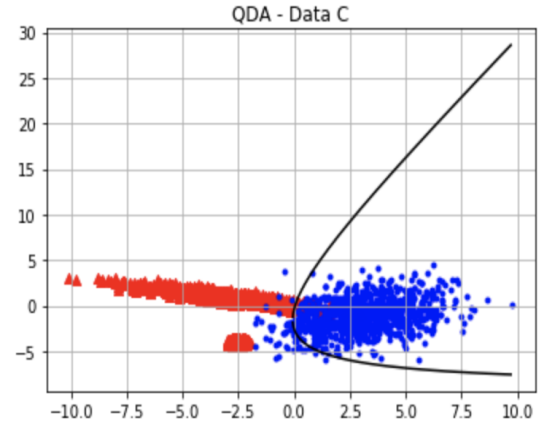
In this case, the three linear models perform poorly and the gap is larger compared to QDA model.

6 Dataset C









Error Table

Misclassification Error	Train	Test
LDA	0.055	0.0423
QDA	0.0525	0.0403
Logistic Regression	0.0425	0.0263
Linear Regression	0.055	0.0423

Comments

In this dataset, one class of points is sampled from a gaussian distribution, while the other takes a union of two different distributions. This leads to a degeneration for most of the models. We can see from the figures that the small red cluster biases the decision boundries.
The logistic regression model gives the best result in this case, since it does not assume the $p(x|y)$ directly.

In this experiments, all the models achieves a lower misclassification error on the test dataset. It is mostly because of the complicated data distributions in the training dataset.

7 Proof

7.1 Exercise 1

Given an i.i.d. sample of N observations $\{(z_i, x_i)|i \in \{1, \ldots, n\}\}$, we use the maximum likelihood estimator to estimate π and θ. Since for $m \in \{1, \ldots, M\}$, $p(z = m) = \pi_m$ is independant of x, we discard at first the variable x. The likelihood function for π is defined as:

$$f(\pi) = \prod_{i=1}^{N} \prod_{m=1}^{M} \pi_m^{\mathbb{1}_{z_i=m}} \qquad \text{s.t.} \sum_{m=1}^{M} \pi_m = 1$$

So the log likelihood can be expressed as:

$$\begin{aligned} l(\pi) &= \sum_{i=1}^{N} \sum_{m=1}^{M} \mathbb{1}_{z_i=m} \log \pi_m + \lambda(1 - \sum_{m=1}^{M} \pi_m) \\ &= \sum_{i=1}^{N} N_m \log \pi_m + \lambda(1 - \sum_{i=1}^{M} \pi_m) \end{aligned}$$

where N_m is the frequency of m in the sample. We then take the derivative for each π_m and λ:

$$\begin{cases} \dfrac{\partial f(\pi)}{\partial \pi_m} & = \dfrac{N_m}{\pi_m} - \lambda \qquad m \in \{1, \ldots, M\} \\ \dfrac{\partial f(\pi)}{\partial \lambda} & = 1 - \sum_{m=1}^{M} \pi_m \end{cases}$$

Let the derivative to be 0 and then we get the MLE estimators for π: $\hat{\pi_m} = \dfrac{N_m}{N}$ for $m \in \{1, \ldots, M\}$. Then we calculate the conditional maximum likelihood estimators for $p(x = k|z = m) = \theta_{mk}$.

$$f(\theta) = \prod_{i=1}^{N} \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{mk}^{\mathbb{1}_{z_i=m, x_i=k}} \qquad \text{s.t.} \sum_{k=1}^{K} \theta_{mk} = 1 \quad \forall m \in \{1, \ldots, M\}$$

The log likelihood is expressed as:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K} \mathbb{1}_{z_i=m, x_i=k} \log \theta_{mk} + \sum_{m=1}^{M} \lambda_m (1 - \sum_{k=1}^{K} \theta_{mk}) \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K} N_{mk} \log \theta_{mk} + \sum_{m=1}^{M} \lambda_m (1 - \sum_{k=1}^{K} \theta_{mk}) \end{aligned}$$

where N_{mk} is the freauency of $x = k, z = m$ in the sample. The derivative of the above function is:

$$\begin{cases} \dfrac{\partial l(\theta)}{\partial \theta_{mk}} & = \dfrac{N_{mk}}{\theta_{mk}} - \lambda_m \qquad m \in \{1, \ldots, M\},\ k \in \{1, \ldots, K\} \\ \dfrac{\partial l(\theta)}{\partial \lambda_m} & = 1 - \sum_{k=1}^{K} \theta_{mk} \qquad m \in \{1, \ldots, M\} \end{cases}$$

Let the derivative to be 0 and then we get the MLE estimators for θ: $\hat{\theta}_{mk} = \dfrac{N_{mk}}{N_m}$ for $m \in \{1, \ldots, M\}$ and $k \in \{1, \ldots, K\}$.

7.2 Exercise 2.1(a)

Using the result of Exercise 1, we can deduce that $\hat{\pi} = \frac{N_1}{N} = \frac{\sum_{i=1}^{N} y_i}{N}$.

Now we calculate the parameters of the Gaussian distributions. The log likelihood function is defined by:

$$\begin{aligned} l(\mu, \Sigma) &= \log \prod_{i=1}^{N} \left(\frac{1}{2\pi |\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x_i - \mu_{y_i})^T \Sigma^{-1} (x_i - \mu_{y_i})) \right) \\ &= -\frac{1}{2} \sum_{i=1}^{N} (x_i - \mu_{y_i})^T \Sigma^{-1} (x_i - \mu_{y_i}) - N(\log 2\pi + \frac{1}{2} \log |\Sigma|) \end{aligned}$$

The derivative is given by:

$$\begin{cases} \dfrac{\partial l(\mu, \Sigma)}{\partial \mu_j} & = -\sum_{i=1}^{N} \sum_{j=0}^{1} \mathbb{1}_{y_i = j} \Sigma^{-1} (x_i - \mu_j) \quad j \in \{0, 1\} \\ \dfrac{\partial l(\mu, \Sigma)}{\partial \Sigma^{-1}} & = -\frac{1}{2} \sum_{i=1}^{N} (x_i - \mu_{y_i})^T (x_i - \mu_{y_i}) + \frac{1}{2} N \Sigma \end{cases}$$

So we have,

$$\begin{cases} \mu_0 = \dfrac{\sum_{i=1}^{N} (1 - y_i) x_i}{N} \\ \mu_1 = \dfrac{\sum_{i=1}^{N} y_i x_i}{N} \\ \Sigma = \dfrac{\sum_{i=1}^{N} (x_i - \mu_{y_i})^T (x_i - \mu_{y_i})}{N} = \dfrac{\sum_{i=1}^{N} (x_i - (1 - y_i)\mu_0 - y_i \mu_1)^T (x_i - (1 - y_i)\mu_0 - y_i \mu_1)}{N} \end{cases}$$