CS 6140: Machine Learning
College of Computer and Information Science
Northeastern University
Instructor: Bilal Ahmed
Scribe: Bilal Ahmed & Yu Wen

The Gaussian Discriminant Analysis Model

1 Introduction

Previously, we've discussed the Naive Bayes Model, where components of the input vector $\mathbf{x}_i$ are discrete-valued. When features of $\mathbf{x}_i$ are continuous-valued random variables, instead of using the Gaussian Naive Bayes Model, we could use the **Gaussian Discriminant Analysis Model** (GDA), in which we assume that, given the label y_i, $\mathbf{x}_i$ follows a multivariate Gaussian distribution. The GDA Model is also a classifier and a generative model.

2 Build the Model

GDA is also a generative model, thus we have

$$P(y_i|\mathbf{x}_i) = \frac{P(\mathbf{x}_i|y_i)P(y_i)}{P(\mathbf{x}_i)}$$

Since the task is to classify an example, here we could make an assumption of the label such that y_i follows a bernoulli distribution specified by parameter π.

$$y_i \sim Bernoulli(\pi)$$

Then, given the fact that the label is known, we can make another assumption of the input variables such that $\mathbf{x}_i|y_i = 0$ and $\mathbf{x}_i|y_i = 1$ follow the multivariate Gaussian distribution specified by (μ_0, Σ) and (μ_1, Σ), respectively.

$$\begin{aligned}
\mathbf{x}_i|y_i = 0 \quad &\sim \mathcal{N}(\mu_0, \Sigma) \\
\mathbf{x}_i|y_i = 1 \quad &\sim \mathcal{N}(\mu_1, \Sigma)
\end{aligned}$$

By giving the distributions, we have

$$\begin{aligned}
P(y_i) &= \pi^{y_i}(1-\pi)^{1-y_i} \\
P(\mathbf{x}_i|y_i = 0) &= \frac{1}{(2\pi)^{\frac{m+1}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{x}_i - \mu_0)^T\Sigma^{-1}(\mathbf{x}_i - \mu_0)) \\
P(\mathbf{x}_i|y_i = 1) &= \frac{1}{(2\pi)^{\frac{m+1}{2}}|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mathbf{x}_i - \mu_1)^T\Sigma^{-1}(\mathbf{x}_i - \mu_1))
\end{aligned}$$

, where $\mu_0, \mu_1 \in \mathbb{R}^m$ is the mean vector; $\Sigma \in \mathbb{R}^{m\times m}$ is the covariance matrix; $|\Sigma| \in \mathbb{R}$ is the determinant of Σ.

3 Train the Model

According to the assumptions we've made above, the parameters specify the GDA are followings. π, which specifies $P(y_i)$; μ_0, Σ which specify $P(\mathbf{x}_i|y_i = 0)$; μ_1, Σ which specify $P(\mathbf{x}_i|y_i = 1)$; Thus, we could write down the log-likelihood of the data

$$l(\pi, \mu_0, \mu_1, \Sigma) = ln \prod_{i=1}^{N} P(\mathbf{x}_i|y_i)P(y_i)$$

By Maximizing l, we could get the optimal parameters as followings,

$$\pi^* = \frac{\sum_{i=1}^{N} \mathbb{1}(y_i = 1)}{N}$$
$$\mu_0^* = \frac{\sum_{i=1}^{N} \mathbf{x}_i \mathbb{1}(y_i = 0)}{\sum_{i=1}^{N} \mathbb{1}(y_i = 0)}$$
$$\mu_0^* = \frac{\sum_{i=1}^{N} \mathbf{x}_i \mathbb{1}(y_i = 1)}{\sum_{i=1}^{N} \mathbb{1}(y_i = 1)}$$
$$\Sigma^* = \frac{\sum_{i=1}^{N} (\mathbf{x}_i - \mu_{y_i})(\mathbf{x}_i - \mu_{y_i})^T}{N}$$

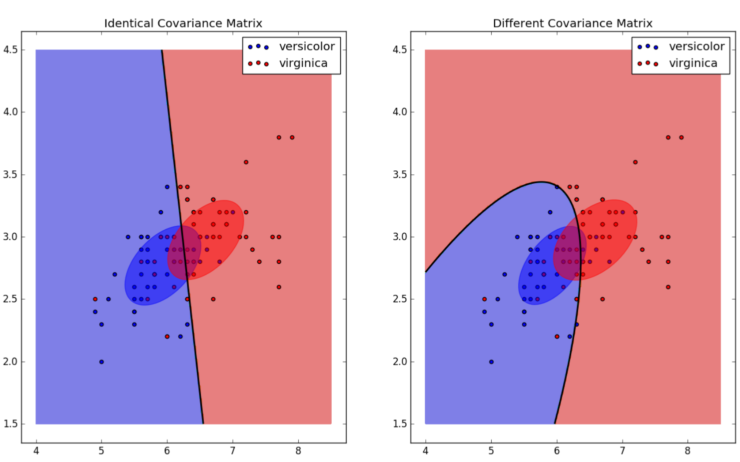
,where $\mathbb{1}(y_i = 0)$ is a indicator function such that if $y_i = 0$,$\mathbb{1}(y_i = 0)$ output 1, otherwise output 0.

4 Prediction

After a model was trained, we can make prediction on the label of a given data such that

$$y_i = \arg\max_{y_i} P(\mathbf{x}_i|y_i)P(y_i)$$

Here we give a visualization of the prediction process. Data is from the Iris flower data set.



Above two figures are two GDA models on the same data set. Decision boundaries are given by black solid line.

The left figure represents the GDA model, where $P(\mathbf{x}_i|y_i = versicolor)$ and $P(\mathbf{x}_i|y_i = virginica)$ have the identical covariance matrix as what we have discussed in above sections. The decision boundary is linear since the two covariance matrices are identical.

The right figure represents another kind of GDA model, where $P(\mathbf{x}_i|y_i = versicolor)$ and $P(\mathbf{x}_i|y_i = virginica)$ have different covariance matrices. As a consequence, the decision boundary is quadratic.

Reference

1. Andre Ng, Generative Learning Algorithms, cs229 lecture notes 2
2. `http://scikit-learn.sourceforge.net/0.5/auto_examples/plot_lda_vs_qda.html`