STATS 235: Modern Data Analysis
Classification Models– LDA, QDA & NB

Babak Shahbaba

Department of Statistics, UCI

Introduction

- Logistic regression is a discriminative model with linear boundaries
- In this lecture, we discuss several generative models, where we model $P(x|y)$
- We start with linear discriminant analysis (LDA), which also provide linear boundaries
- Next, we extend LDA to allow for nonlinear boundaries
- Finally, we discuss naive Bayes classifiers

Linear discriminant analysis

- When the set of p predictors, x, are continuos random variables, we can assume that their joint distribution is multivariate normal for each class,

$$f_k(x) \quad = \quad (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)]$$

- Note that in this setting, only the mean of the distributions, μ_k, changes from one class to another. The covariance matrix Σ remains the same for all classes.

- This assumption is of course not realistic and is made only for simplicity. We will relax it later.

Linear discriminant analysis

- Using Bayes theorem, we have

$$P(y = k|x) = \frac{\pi_k f_k(x)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}(x)}$$

where $\pi_k = P(y = k)$.

- For a given value of x, the denominator remains the same for all classes. Therefore, we can define the discriminant function based on the numerator, $\pi_k f_k(x)$, or more commonly based on its log,

$$\begin{aligned}
\delta_k(x) &= \log \pi_k + log[f_k(x)] \\
&= \log \pi_k - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)
\end{aligned}$$

Linear discriminant analysis

- With further simplification (and removing the constant parts), we have

$$\delta_k(x) \quad = \quad \log \pi_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + x^T \Sigma^{-1} \mu_k$$

- Note that the above functions are linear in x.
- Therefore, we refer to them as *linear discriminant functions*.
- Classifying cases according to these functions is called *linear discriminant analysis* (LDA).

- We can estimate π_k and μ_k for $k = 1, \dots, K$, and Σ as follows:

$$\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n} \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k}^{n_k} x_i \\
\hat{\Sigma} &= \frac{1}{n-k} \sum_{k=1}^{K} \sum_{i:y_i=k}^{n_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T
\end{aligned}$$

where n_k is the number of observed cases (training cases) that belong to class k.

Linear discriminant analysis

- After estimating the model parameters, we assign each case, i, to the class whose value of the discriminant function, $\delta_k(x_i)$, is the highest.

- Cases for which $\delta_k(x) = \delta_l(x)$ fall on the decision boundary between the two classes k and l.

- For these cases, $\delta_k(x) - \delta_l(x) = 0$, which means

$$\log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k - \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l) = 0$$

- Note that the above equation, which specifies the decision boundary, is linear in x. As the result, the decision boundaries are *hyperplanes* in the p-dimensional space. (The decision boundary is straight line if we have two predictors only.)

Linear discriminant analysis

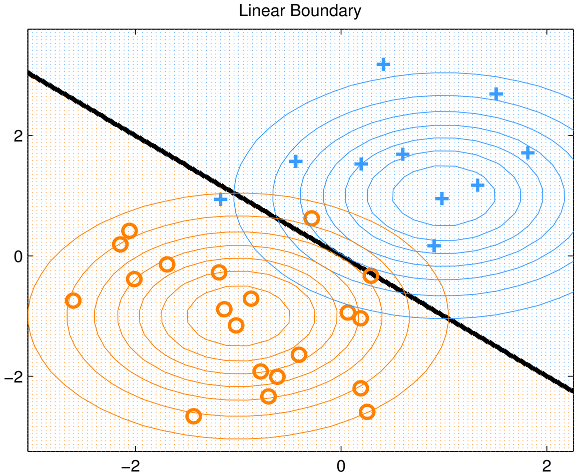

Figure 4.5a in Murphy (2012)

Quadratic discriminant analysis

- As mentioned above, the equal-covariance assumption is restrictive and is only made for convenience.
- By relaxing this assumption, the discriminant function becomes

$$\delta_k(x) \quad = \quad \log \pi_k - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)$$

  which are quadratic functions of x; hence, they are called *quadratic discriminant functions*.

- Classifying cases according to these functions is called *quadratic discriminant analysis* (QDA).
- The decision boundaries for this approach are not linear any more.

Linear discriminant analysis

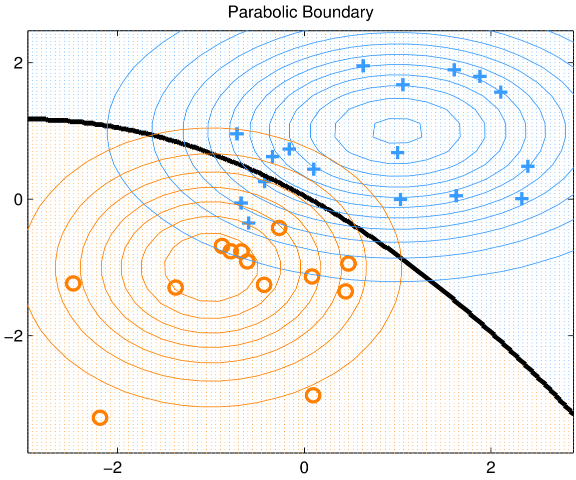

Figure 4.3a in Murphy (2012)

Naive Bayes models

- This is an alternative classification model, which is especially attractive when the dimension p is large.
- In this approach, we again use Bayes theorem to obtain the probability of each class given the observed values of predictors,

$$P(y = k|x_1, \dots, x_p) = \frac{P(y = k)P(x_1, \dots, x_p|y = k)}{\sum_{k'=1}^{K} P(y = k')P(x_1, \dots, x_p|y = k')}$$

- This time, however, we make an assumption that is naive and possibly wrong, but it simplifies the model: we assume that given a class $y = k$, the predictors are independent,

$$P(x_1, \dots, x_p|y = k) = \prod_{j=1}^{p} P(x_j|y = k)$$

Naive Bayes models

- As a result of the above naive assumption, the model simplifies to

$$P(y = k|x_1, \dots, x_p) = \frac{P(y = k) \prod_{j=1}^{p} P(x_j|y = k)}{\sum_{k'=1}^{K} P(y = k') \prod_{j=1}^{p} P(x_j|y = k')}$$

- As before, we assign each case, i, to the class with the highest conditional probability given $x_{i1}, \dots, x_{ip}$.

- It is more common to distinguish between two classes using the following logit function

$$\begin{aligned}\log \frac{P(y = k|x_1, \dots, x_p)}{P(y = l|x_1, \dots, x_p)} &= \log \frac{P(y = k) \prod_{j=1}^{p} P(x_j|y = k)}{P(y = l) \prod_{j=1}^{p} P(x_j|y = l)} \\ &= \log \frac{\pi_k}{\pi_l} + \sum_{j=1}^{p} \log \frac{P(x_j|y = k)}{P(x_j|y = l)}\end{aligned}$$

Naive Bayes models

- In practice, we estimate π_k using the proportion of observed cases that belong to class k.

- To estimate $P(x_j|k)$, we first need to assume a probability distribution model for x_j given k.

- If x_j is categorical, we can estimate $P(x_j|k)$ using the observed proportion of each category of x_j for cases with $y = k$.

- If x_j is continuous, we can assume $x_j|k$ has a Gaussian distribution and estimate its mean and variance using the cases with $y = k$.