4.4 Generative Models for Classification

Dr. Lauren Perry

The Need for Alternatives

Why not just use logistic regression?

- If there is a lot of separation between the classes, logistic regression models are surprisingly unstable.
    - (Coefficient estimates can vary significantly given the same data generating process.)
- If the distribution of the predictors X is approx. normal and the sample size is small, these alternatives may be more accurate than logistic regression.
- The methods in this section have more natural extensions to three or more classes.

Idea

- Model the distribution of the predictors X separately for each response class.
- Use Bayes' theorem to work these into estimates for $\mathrm{P}(Y = k | X = x)$.

The Setup

Suppose Y can take on K distinct, unordered values.

- Let π_k represent the overall probability that a randomly chosen observation comes from the kth class.
  - Generally estimated as the proportion of training observations belonging to class k.
- Let $f_k(x) = \mathrm{P}(X|Y = k)$ denote the density function of X for an observation from the kth class.
  - So $f_k(x)$ should be relatively large if there is a high probability that an observation from the kth class has $X \approx x$.
- Then Bayes' Theorem states

$$p_k(x) = \mathrm{P}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=i}^{K} \pi_l f_l(x)}$$

  - This is the *posterior probability* that an observation belongs to the kth class, given $X = x$.

Goal: estimate $f_k(x)$ to approximate the Bayes' classifier $p_k(x)$.

Linear Discriminant Analysis for One Predictor, $p = 1$

We will classify an observation into the category for which $p_k(x) = \mathrm{P}(Y = k|X = x)$ is greatest.

- Assume $f_k(x)$ is normally distributed (Gaussian):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right]$$

  where μ_k and σ_k^2 are the mean and standard deviation parameters for the kth class.
- Also assume $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2 = \sigma^2$ (shared variance term for all classes).

Linear Discriminant Analysis for One Predictor, $p = 1$

Combining the Bayes' Theorem set up with these assumptions, we get

$$p_k(x) = \frac{\frac{\pi_k}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu_k)^2\right]}{\sum_{l=i}^{K} \frac{\pi_l}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu_l)^2\right]}$$

which looks a mess, but it can be shown this is equivalent to assigning the observation to the class for which

$$\delta_k(x) = x\left(\frac{\mu_k}{\sigma^2}\right) - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

is largest.

Linear Discriminant Analysis for One Predictor, $p = 1$

$$\delta_k(x) = x\left(\frac{\mu_k}{\sigma^2}\right) - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

If $K = 2$ and $\pi_1 = \pi_2$, this classifier assigns an observation to

- class 1 if $2x(\mu_1 - \mu_2) > \mu_1 - \mu_2$.
- class 2 *otherwise*.

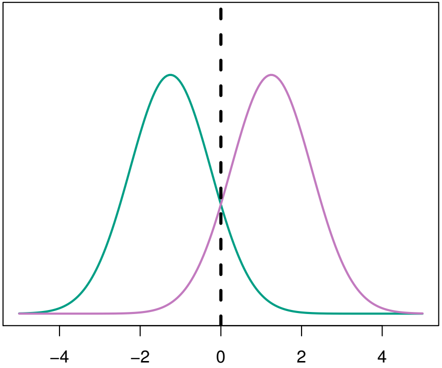
The Bayes' decision boundary is the point for which $\delta_1 = \delta_2$, which in this setting is

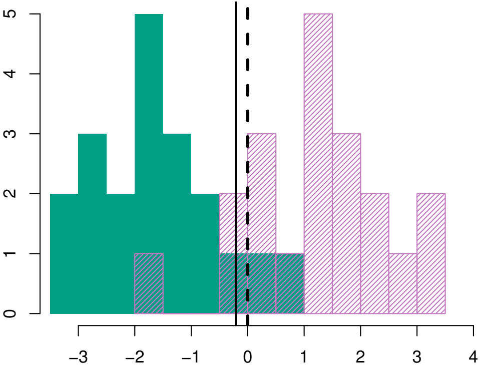
$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

Example

- Consider predictors generated from two normal distributions where $mu_1 = -1.25$, $\mu_2 = 1.25$, and $\sigma_1 = \sigma_2 = 1$.
- Assume an observation is equally likely to come from either class, i.e., $\pi_1 = \pi_2 = 0.5$.
- Then the (known) Bayes' classifier assigns an observation to class 1 if $x < 0$ and class 2 otherwise.

Example





- 20 observations drawn from each class.
- LDA decision boundary shown as solid vertical line.

Linear Discriminant Analysis for One Predictor, $p = 1$

In practice, we must estimate $\mu_1, \ldots, \mu_K$, $\pi_1, \ldots, \pi_K$, and σ.

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = \frac{n_k}{n}$$

Where n is the number of training observations and n_k is the number of training observations in the kth class.

- $\hat{\sigma}^2$ is a weighted average of sample variances across the K classes.

Linear Discriminant Analysis for One Predictor, $p = 1$

Assign an observation $X = x$ to the class for which

$$\delta_k(x) = x\left(\frac{\hat{\mu}_k}{\hat{\sigma}^2}\right) - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

is largest.

Example: Using Penguin Body Mass to Predict Species

```r
data(penguins, package = "palmerpenguins")
mod1 <- lda(species ~ body_mass_g, penguins)
predval <- predict(mod1)$class
species <- penguins$species[!is.na(penguins$species) & !is.na(penguins$body
table(predval, species)
```

```
##            species
## predval     Adelie Chinstrap Gentoo
##   Adelie       140        64     14
##   Chinstrap      0         0      0
##   Gentoo        11         4    109
```

```r
mean(predval == species)
```

```
## [1] 0.7280702
```