3 LDA "linear discriminant analysis"

Logistic regression involves direction modeling $P(Y = k|X = x)$ using the logistic function for the case of two response classes. We now consider a less direct approach.

Idea:

Model the distribution of the predictors X separately in each of the response classes (given Y) and then use Bayes theorems to flip these around and get estimates for $P(Y=k|X=x)$.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Why do we need another method when we have logistic regression?

* 1. We might have more than 2 response classes.

even with just 2 class in the response

2. If n is small and the distribution of the predictors is approximately normal in each class, LDA is more stable than logistic regression.

3. When classes are well-separated, the parameter estimates in logistic regression are surprisingly unstable.

3.1 Bayes' Theorem for Classification

notation

Suppose we wish to classify an observation into one of K classes, where $K \geq 2$.

categorical Y with K classes (possible distinct and unordered values).

π_k - overall or "prior" probability that a randomly chosen oservation falls into the k^{th} class.

→ could know this from domain knowledge
could estimate from training data

$f_k(x) = P(X=x \mid Y=k)$ ← only works in discrete case

probability that X falls into a small region around x given $Y=k$ (cts).

conditional density function of X for an observation that comes from class k.

$$\overset{A}{P}(Y = k|\overset{B}{X = x}) = \frac{\overset{P(\overset{A}{Y=k})}{\pi_k} \; \overset{P(\overset{B}{X=x}|\overset{A}{Y=k})}{f_k(x)}}{\underbrace{\sum_{l=1}^{K} \pi_l \; f_l(x)}_{\underset{B}{P(X=x)}}}$$

Bayes theorem

Use the same abbreviation as before
$p_k(x) = P(Y=k \mid X=x)$
"posterior probability" that an observation $X=x$ comes from the k^{th} class.

In general, estimating π_k is easy if we have a random sample of Y's from the population.

could get from domain knowledge

computing the fraction of training observations that come from the k^{th} class.

Estimating $f_k(x)$ is more difficult unless we assume some particular forms.

If we can estimate $f_k(x)$ we can classifier that is close to the "best" classifier (more later).

"optimal" classifier: assuming we know $p_k(x) = P(Y=k \mid X=x)$

- assignment to class with the highest posterior probability $p_k(x)$.
- "Bayes classifier" and is known to be optimal in terms of overall error rate.

i.e. we can do no better than the Bayes classifier.

3.2 p = 1

Let's (for now) assume we only have 1 predictor. We would like to obtain an estimate for $f_k(x)$ that we can plug into our formula to estimate $p_k(x)$. We will then classify an observation to the class for which $\hat{p}_k(x)$ is greatest.

$\hookrightarrow \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}$.

estimating the Bayes classifier!

Suppose we assume that $f_k(x)$ is normal. In the one-dimensional setting, the normal density takes the form

Gaussian

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

variance parameter for k^{th} class

mean parameter for k^{th} class

Let's also (for now) assume $\sigma_1^2 = \ldots = \sigma_K^2 = \sigma^2$ (shared variance term).

Plugging this into our formula to estimate $p_k(x)$,

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{\ell=1}^{K} \pi_\ell \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_\ell)^2\right)}$$

not 3.14159..., this denotes the prior prob. that observation falls into ℓ^{th} class.

take log, rearrange.

We then assign an observation $X = x$ to the class which makes $p_k(x)$ tthe largest. This is equivalent to

assign obs. to class which makes

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

largest.

LDA decision criteria is linear in x $\Rightarrow$ "Linear discriminant analysis"

Example 3.1 Let $K = 2$ and $\pi_1 = \pi_2$. When does the Bayes classifier assign an observation to class 1?

When $\delta_1(x) > \delta_2(x)$?

what x values will make this happen?

$$\Leftrightarrow x\frac{\mu_1}{\cancel{\sigma^2}} - \frac{\mu_1^2}{2\cancel{\sigma^2}} + \cancel{\log(\pi_1)} > x\frac{\mu_2}{\cancel{\sigma^2}} - \frac{\mu_2^2}{\cancel{\sigma^2}} + \cancel{\log(\underbrace{\pi_2}_{=\pi_1})}$$

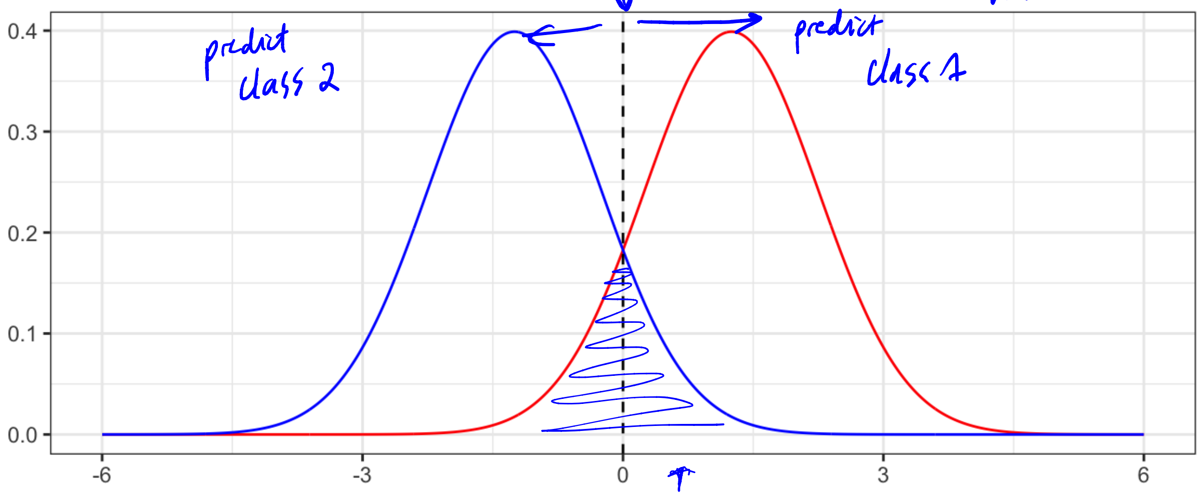
$$2x(\mu_1 - \mu_2) > \underbrace{\mu_1^2 - \mu_2^2}_{(\mu_1-\mu_2)(\mu_1+\mu_2)}$$

$$x > \frac{\mu_1+\mu_2}{2}$$

Bayes decision boundary.

$\Rightarrow$ then we will predict class 1

Bayes classifier will minimize overall error rate.



example where $\pi_1 = \pi_2 = 0.5$

$\mu_2 = -1.25$, $\mu_1 = 1.25$, $\sigma = 1$ $\Rightarrow$ Bayes decision boundary would be at 0.

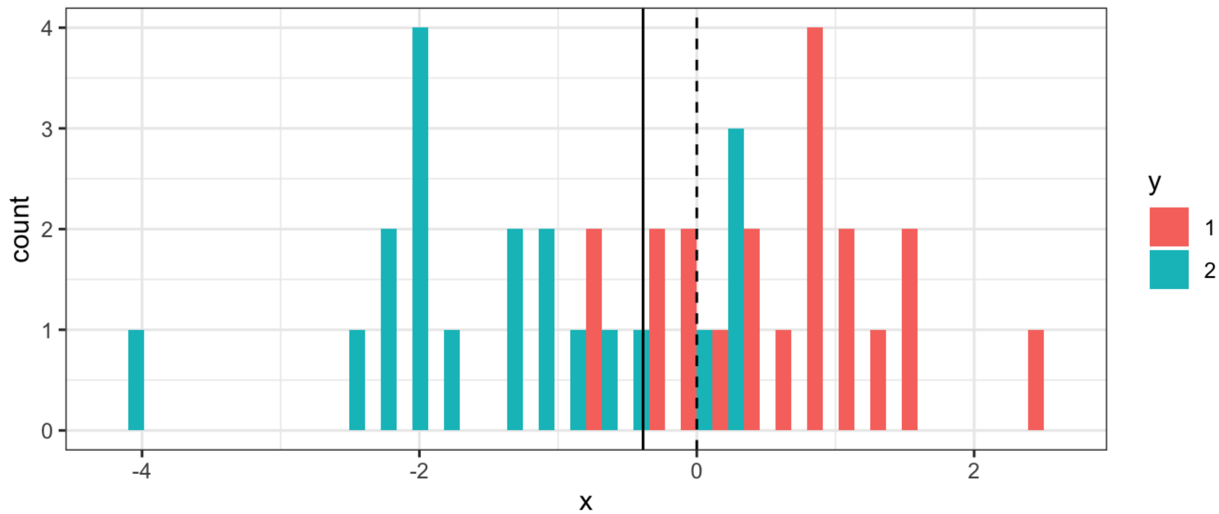
In this case we know $f_k(x) \sim N(\mu_k, \sigma^2)$ $\Rightarrow$ we can create this Bayes classifier!

In practice, even if we are certain of our assumption that X is drawn from a Gaussian distribution within each class, we still have to estimate the parameters $\mu_1, \dots, \mu_K, \pi_1, \dots, \pi_K, \sigma^2$.

The *linear discriminant analysis* (LDA) method approximated the Bayes classifier by plugging estimates in for π_k, μ_k, σ^2.

Sometimes we have knowledge of class membership probabilities $\pi_1, \dots, \pi_K$ that can be used directly. If we do not, LDA estimates π_k using the proportion of training observations that belong to the kth class.

The LDA classifier assignes an observation $X = x$ to the class with the highest value of




```
##    pred
## y       1     2
##   1 18966  1034
##   2  3855 16145
```

The LDA test error rate is approximately 12.22% while the Bayes classifier error rate is approximately 10.52%.

The LDA classifier results from assuming that the observations within each class come from a normal distribution with a class-specific mean vector and a common variance σ^2 and plugging estimates for these parameters into the Bayes classifier.