False positive rate = type I error = 1 - specificity
True positive rate = 1 - type II error = power = sensitivity = recall
Positive predictive value = precision = 1 - false discovery proportion
Negative predictive value

2/11/23
Day 10

4.4.4: Quadratic Discriminant Analysis

- QDA assumes that each class has its own covariance matrix; i.e., an observation X from class k is s.t. $X \sim N(\mu_k, \Sigma_k)$.

Now, the discriminant function is

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

Note that δ_k is quadratic in x.

- We prefer LDA or QDA based on the bias-variance trade-off.
  - With p predictors, estimating the covariance matrix requires estimating $p(p+1)/2$ parameters. Then with K classes, that becomes $Kp(p+1)/2$ parameters.
  - LDA only requires Kp parameters to be estimated.

- LDA has less variance but more potential bias.
- LDA better for smaller training sets. QDA requires much more data.

4.5: A Comparison of Classification Methods

- Logistic regression can outperform LDA if the Gaussian assumption is not met, and v.v. v.v..
- KNN should outperform LDA/logistic regression when the decision boundary is highly nonlinear. (but we don't get any predictor coefficients.

- Wait, ~~why~~ how does logistic regression assume a linear decision boundary?

~~The boundary is determined to:~~ logistic regression is modelled by

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \vec{\beta}X$$

and thus the decision boundary occurs ~~at~~ when

$$0 = \beta_0 + \vec{\beta}X \Rightarrow \vec{\beta}X = -\beta_0.$$

Solve, you get a linear subspace.

Ch. 5: Resampling Methods

- Repeatedly drawing sample from a training set and refitting a model of interest on each sample to obtain additional information about the fitted model.
- Will discuss cross-validation and the bootstrap.
- CV can be used to estimate test errors or to select model flexibility.
  - ↳ Model selection
  - ↳ Model assessment
- Bootstrap gives a measure of accuracy

5.1: Cross-Validation

Hold out a set to serve as a proxy for test data

5.1.1: The validation set approach

1) Randomly divide data into a training set and a validation set
2) Fit model on training set
3) Calculate error on the validation set. This is the ~~prox~~ estimate of the test error.

This is an alternative to looking at p-values.

Conceptually simple, but two drawbacks:

1) The validation estimate of the test error rate can be highly variable.
2) We're throwing out a lot of data, which may result in overestimating the test error.