Assignment 7

Christina Lam

What is/are the requirement(s) of LDA?

- **Normality:** assumes predictor variables/features within each class are normally distributed
- **Homoscedasticity:** variance of predictor variables is the same for all classes
- **Independence:** assumes predictor variables are independent of each other within each class
- **Equal Covariance Matrices:** assumes covariance matrices of predictor variables are equal across all classes
- **Large Sample Size:** not strictly a requirement, but tends to perform better with large sample size relative to number of predictor variables
- **Balanced Classes:** works best when classes in dataset are balanced so there is an equal number of observations for each class
- **Continuous Predictors:** designed for continuous predictor variables

How LDA is different from Logistic Regression?

- **Objective Function**
  - LDA: maximizes likelihood of observing data given class labels
  - Logistic Regression: maximizes likelihood of observing data given class labels by minimizing logistic loss function/equivalently maximizing the log likelihood of data
- **Output Type**
  - LDA: produces posterior of probability for each class and assigns the observation to class with highest probability
  - Logistic Regression: directly models probability of belonging to class using logistic function
- **Assumption about the Data**
  - LDA: assumes predictor variables within each class are normally distributed and have the same covariance matrix
  - Logistic Regression: makes no assumptions about distribution of predictor variables

What is ROC?

- **Receiver Operating Characteristic**
  - ROC curve is a graphical representation of the performance of a binary classification model across various threshold settings
  - Plots the true positive rate against the false positive rate for different threshold values
  - Area under the ROC curve quantifies the overall performance of a binary classification model (represents the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance)
  - Important tool for evaluating the performance of binary classification models and comparing different classifiers
  - Provides insights into the trade-off between sensitivity and specificity across various threshold settings

Sensitivity vs Specificity

- **Sensitivity**
  - More important when the cost of missing positive cases (false negatives) is high
  - Ex. Failing to diagnose a disease can have serious consequences

- **Specificity**
  - More important when the cost of false positives is high
  - Ex. False positives in screening tests may lead to unnecessary follow-up tests/procedures

Which is more critical for the purpose of prediction?

- **True Class (Actual Outcome)**
  - Represents the ground truth or actual outcome of the event being predicted
  - Essential for evaluating accuracy/effectiveness of predictive model
  - Benchmark for assessing model's performance (prediction accuracy measured based on how well the model's predictions match the true classes
- **Predicted Class**
  - Outcome predicted by model based on input features
  - Indicates model's beliefs/estimation of the most likely outome for a given input
  - Predicted class guides decision-making processes (determines the action/strategy to be taken based on model's predictions
- **Both are critical to determine the success of the predictive model!**

Calculate the prediction error from the following:

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

252 + 23 / 10,000 = <u>0.0275</u>