NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 15: Lecture 1
Classification

How can we classify data into three or more categories?

DS-UA 112
Introduction to Data Science

Week 15: Lecture 1
Classification

Adapted from Nolan, Speed, Gonzalez, Lau

Announcements

- Please check Week 15 agenda on NYU Classes
  - Exam
    - Wednesday May 13
    - Gradescope
  - Project 2
    - Due on Tuesday May 12 at 11:59PM EST

Review

- The phrase true positive rate means recall

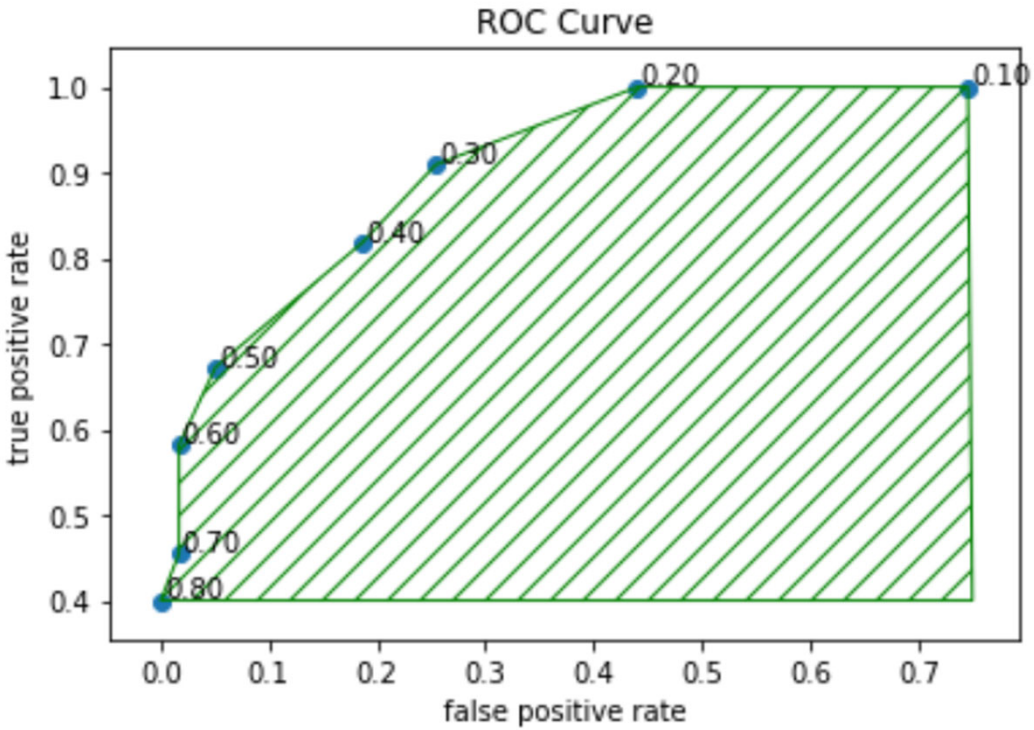
$$\text{True Positive Rate} = \frac{\#\text{True Positive}}{\#\text{True Positive} + \#\text{False Negative}}$$

- The false positive rate complements the true positive rate.

$$\text{False Positive Rate} = \frac{\#\text{False Positive}}{\#\text{True Negative} + \#\text{False Positive}}$$

Review

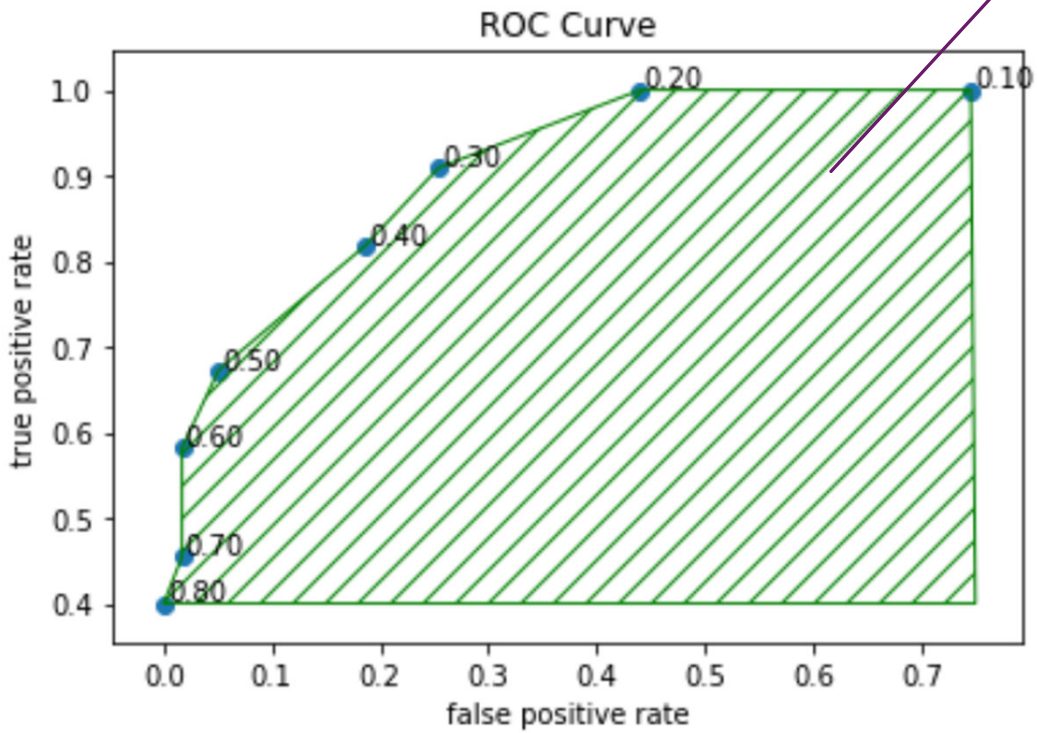
- A ROC curve plots the true positive rate and the false positive rate
- The acronym ROC stands for Receiver Operating Characteristic.
- We can summarize the ROC curve with the area under the curve. We abbreviate the area under the curve as AUC.



Review

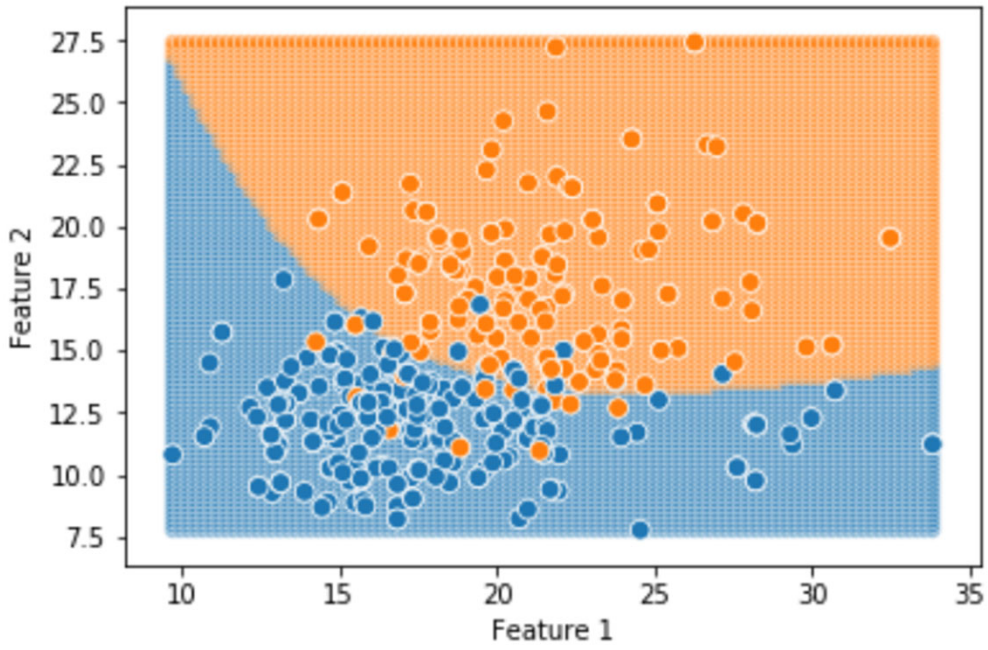
- A ROC curve plots the true positive rate and the false positive rate
- The acronym ROC stands for Receiver Operating Characteristic.
- We can summarize the ROC curve with the area under the curve. We abbreviate the area under the curve as AUC.

If AUC is close to 1, then we have high true positive rate and low false positive rate



Review

- Remember that we can transform the features in a linear regression model to fit data with a nonlinear shape
- Similarly we can transform the features in a logistic regression model to obtain a curved decision boundary.
- Sometimes we want the decision boundary to bend around the regions containing the two categories

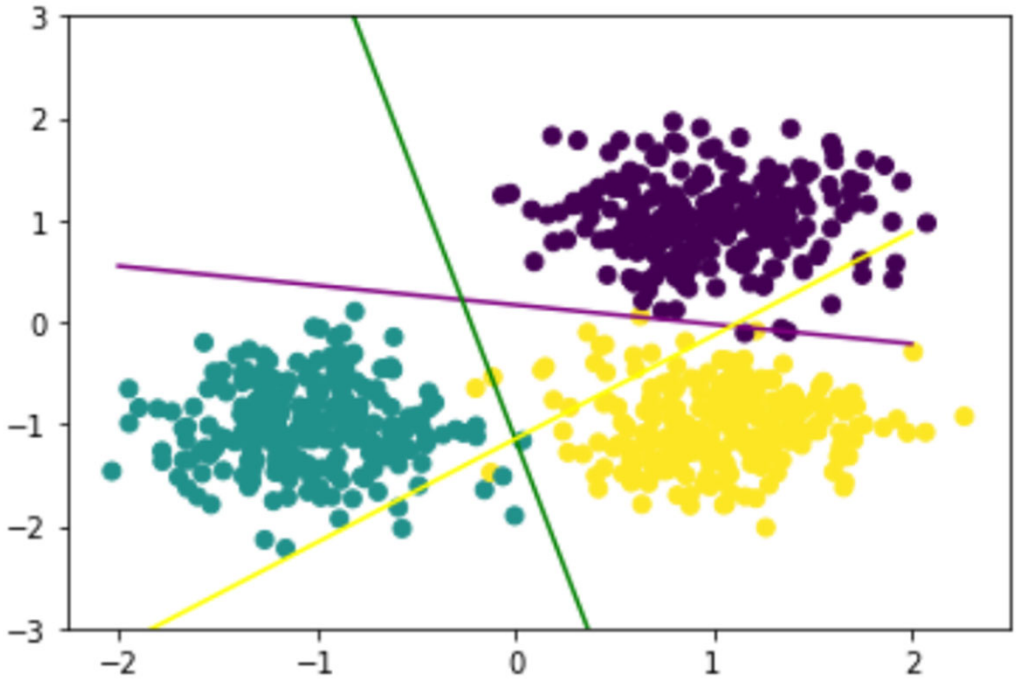


Agenda

- Multiple Categories
- Nearest Neighbors

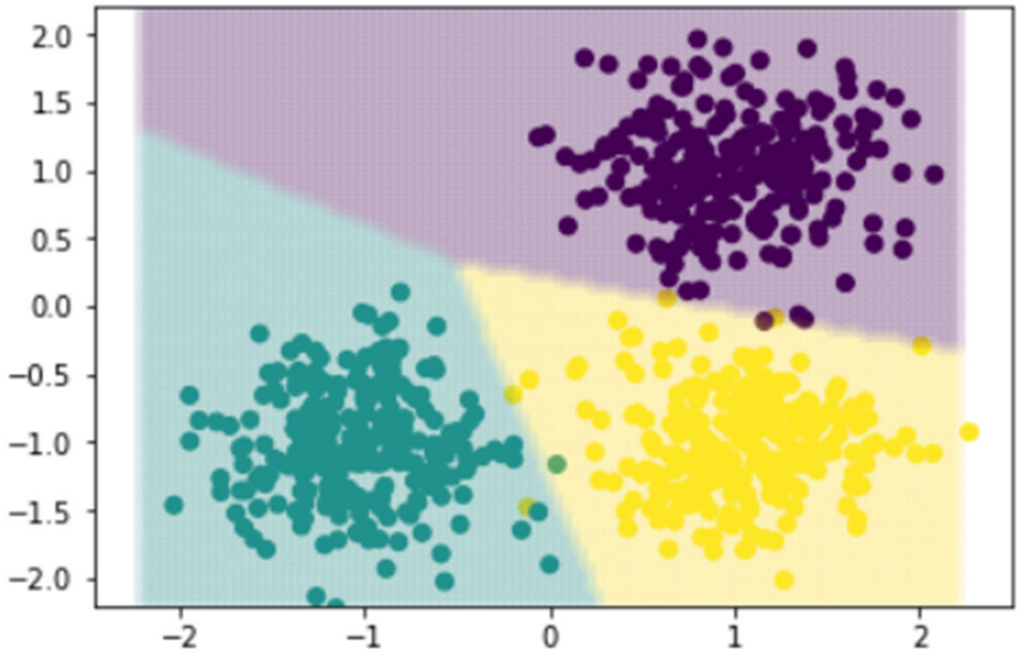
Multiple Categories

- If we have three or more categories, then we can split the classification problem into multiple problems with two categories.
- Each problem try to classify one category versus the other categories. We call the approach One-versus-Rest.



Multiple Categories

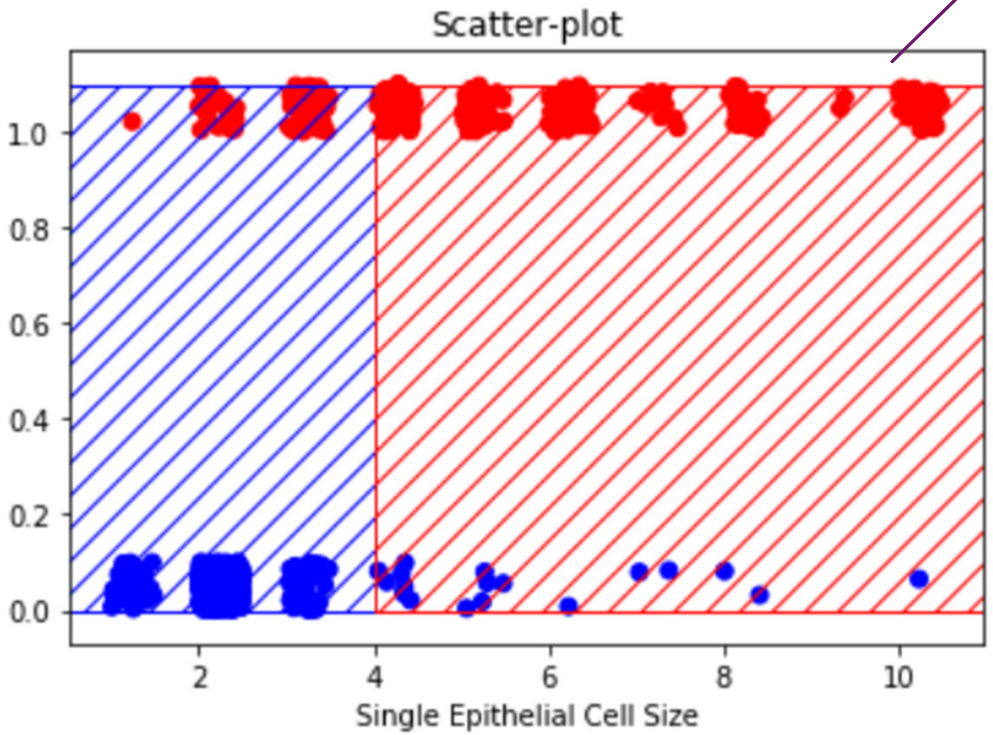
- Alternatively we can extend logistic regression to treat multiple categories. We need to allow for more parameters in the model. However, we do not need to fit a model for each category.
- The One-vs-Rest approach can sometimes neglect categories with fewer points leading to lower accuracy



Decision Boundaries

- With regression we predict a quantitative response variable from explanatory variables
- With classification we predict a qualitative response variable from explanatory variables
- We should compare fitting a line to the data in regression to determining a decision boundary in classification

How can we determine wiggly decision boundaries?



Summary

- Multiple Categories
- Nearest Neighbors

Goals

- Extend logistic regression to allow for multiple categories
- Classify an unlabeled point with nearby labeled points