NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 15: Lecture 2
Nearest Neighbors

How can we determine a wiggly decision boundary?

DS-UA 112
Introduction to Data Science

Week 15: Lecture 2
Nearest Neighbors

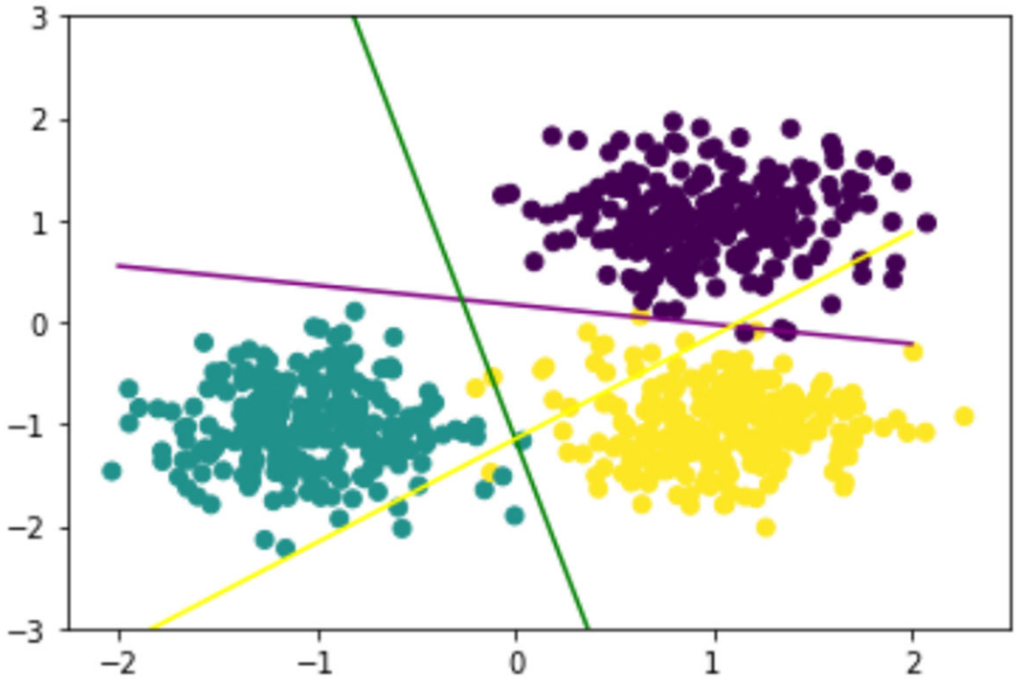
Adapted from Nolan, Speed, Gonzalez, Lau

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

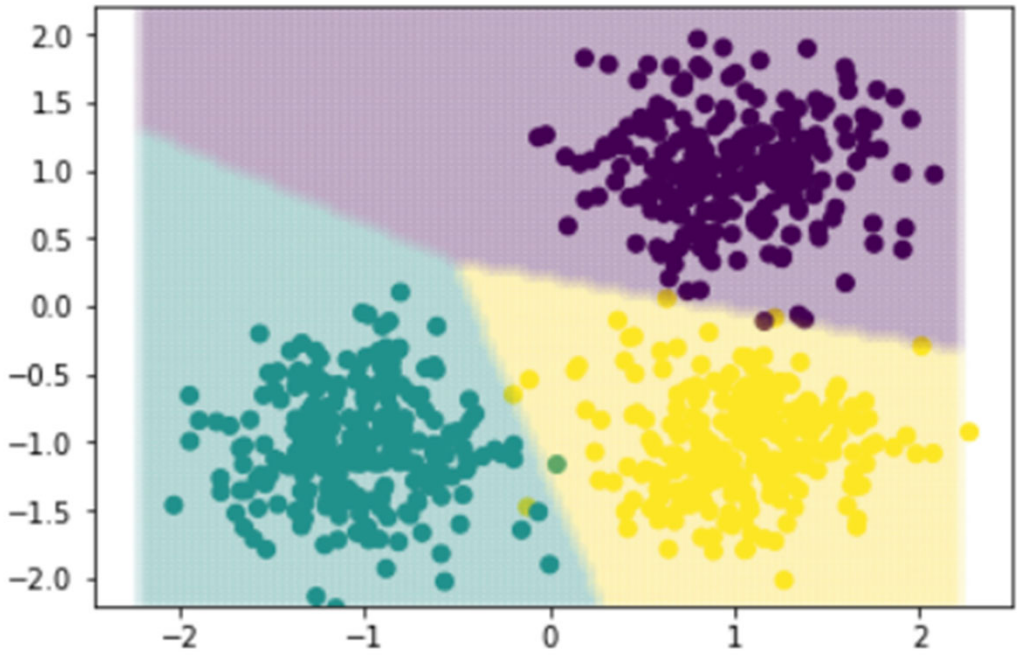
Review

- 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.



Review

- 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

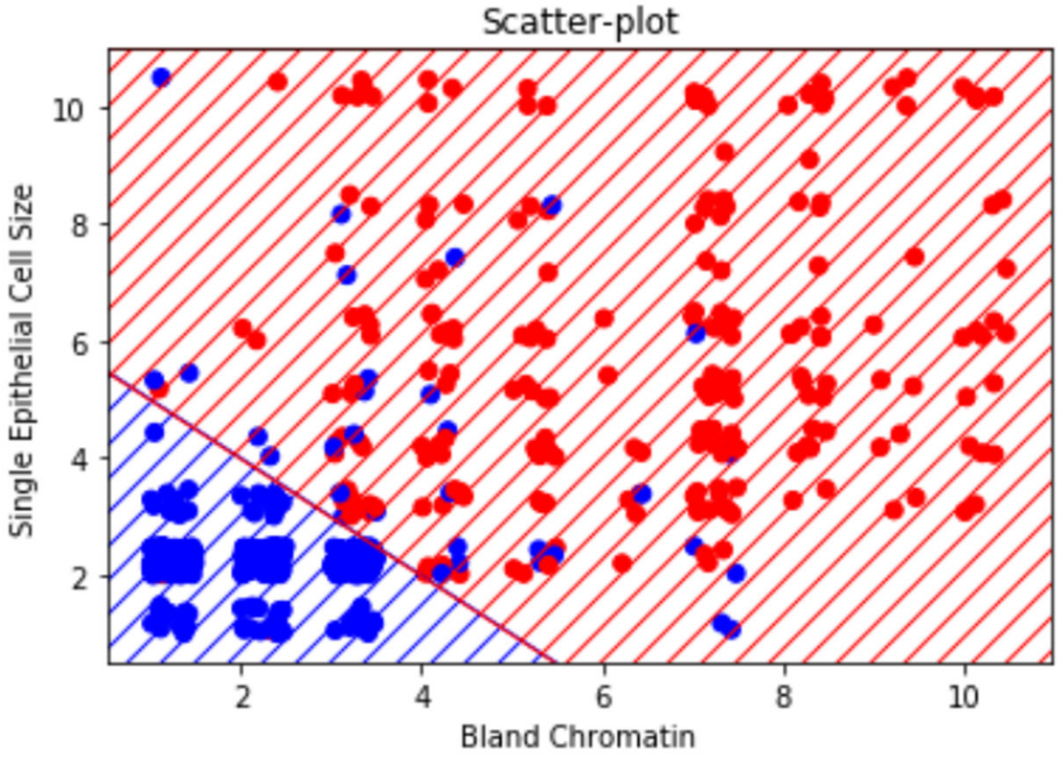


Agenda

- Nearest Neighbors
  - Comparison to Linear Regression and Logistic Regression
  - Adjusting the Number of Neighbors

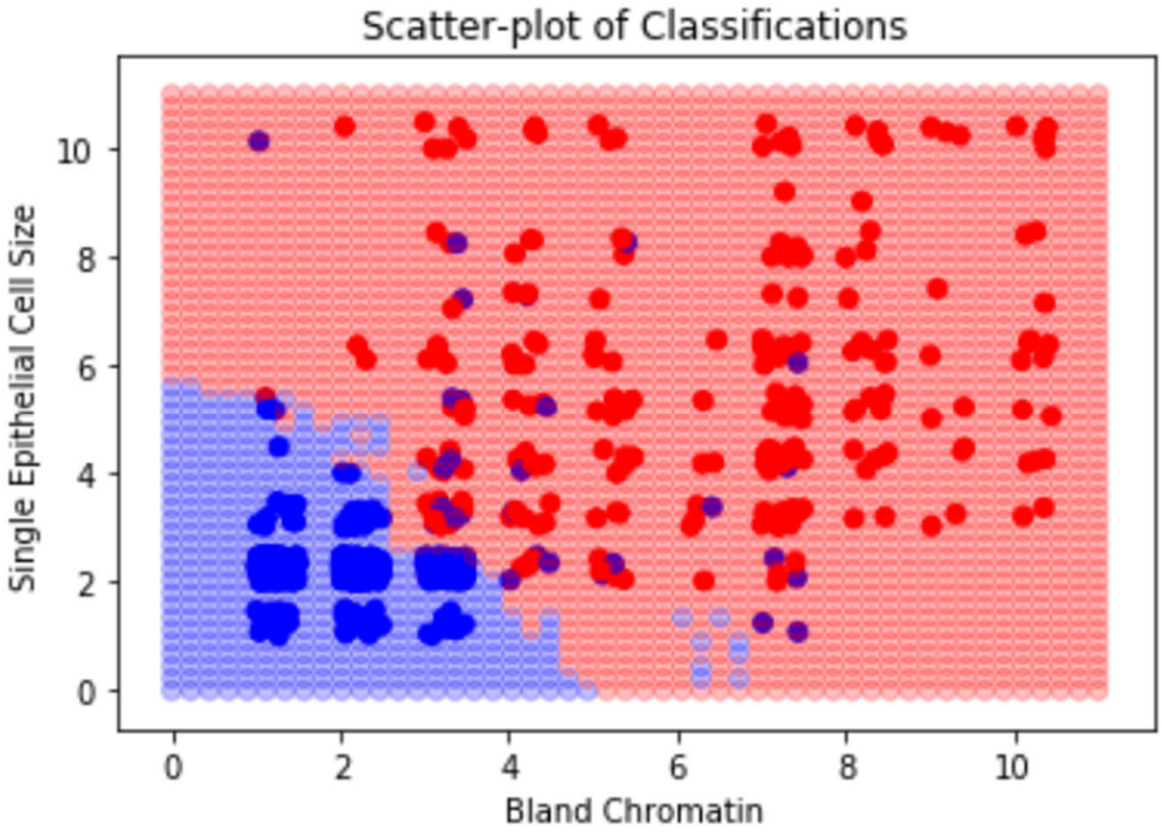
Decision Boundaries

- 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



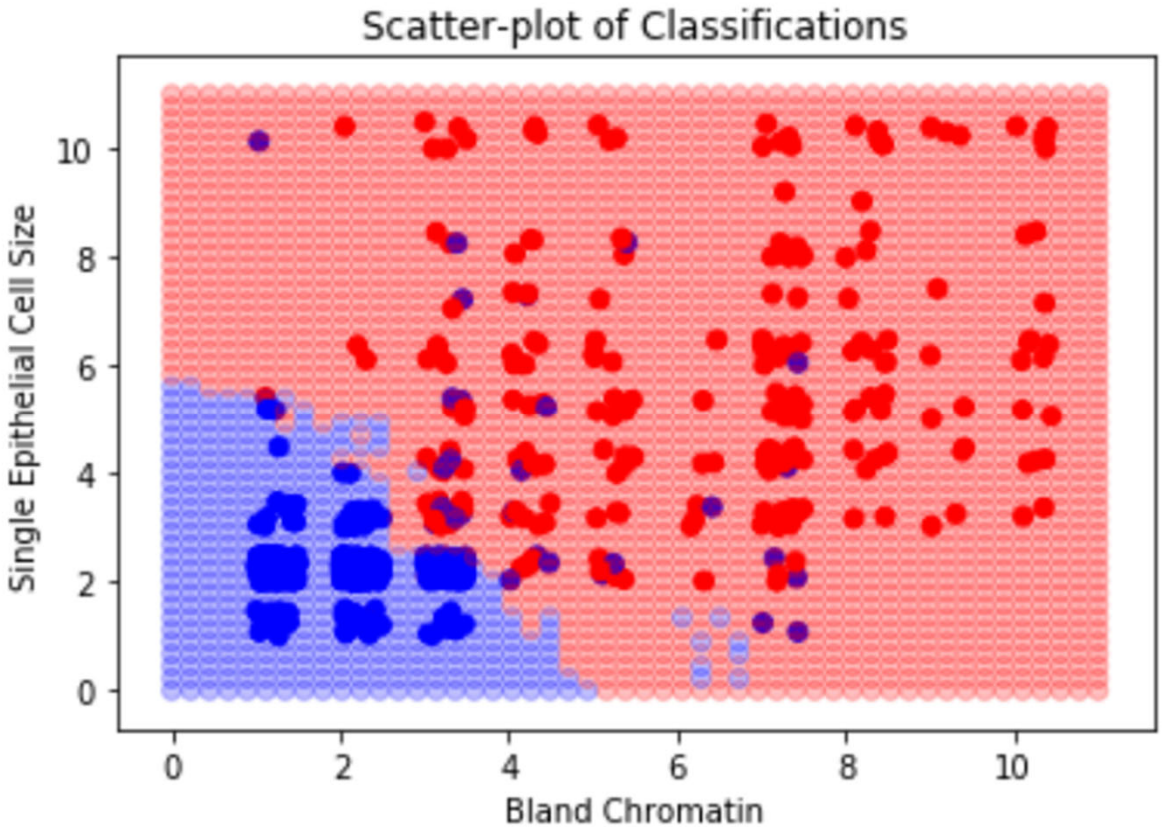
Nearest Neighbors

- Each record in the dataset has a label for the two categories.
- If we have an unlabeled record, then we can compare values for its explanatory variables to values of the explanatory variables for the labeled records.



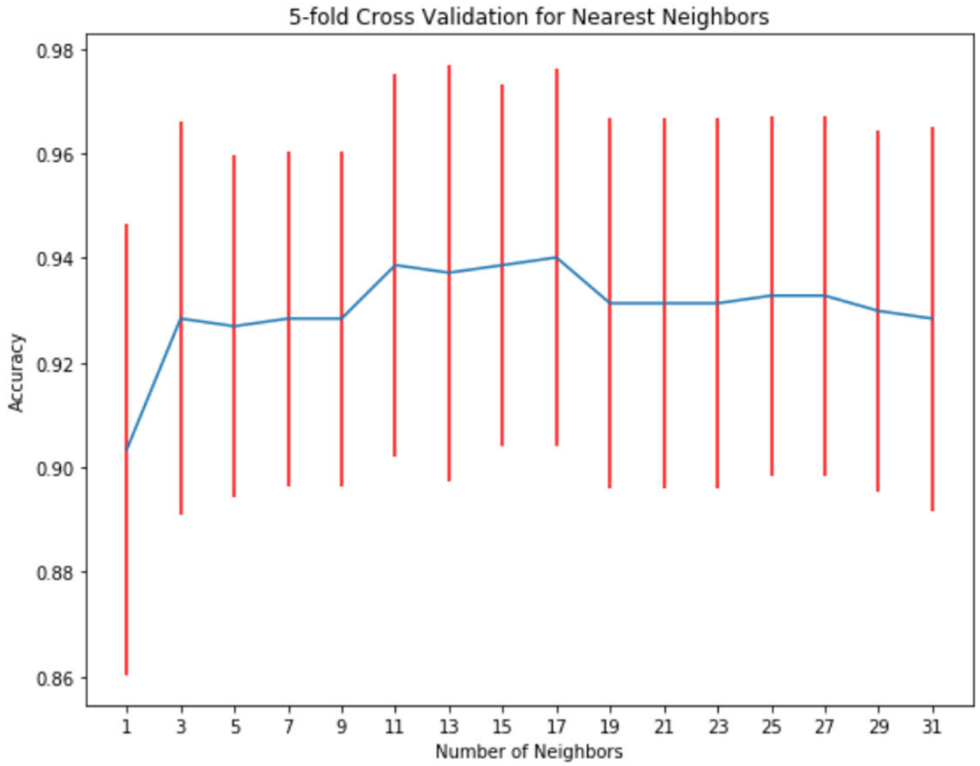
Nearest Neighbors

- We determine the category of the unlabeled record from the categories of the nearest labeled records.
- If we predict categories for many unlabeled records then we can determine the boundary



Number of Neighbors

- If we have few neighbors then we might have overfitting
- If we have too many neighbors then we might have underfitting
- We can use cross validation to determine the number of neighbors



Summary

- Nearest Neighbors
  - Comparison to Linear Regression and Logistic Regression
  - Adjusting the Number of Neighbors

Goals

- Implement nearest neighbor model for classification
- Use cross validation to determine the number of neighbors