NYU Center for Data Science

DS-UA 111
Data Science for Everyone

Week 15: Lecture 2
Classification

NYU Center for Data Science

How can we use regression to predict qualitative variables instead of quantitative variables?

DS-UA 111
Data Science for Everyone

Week 15: Lecture 2
Classification

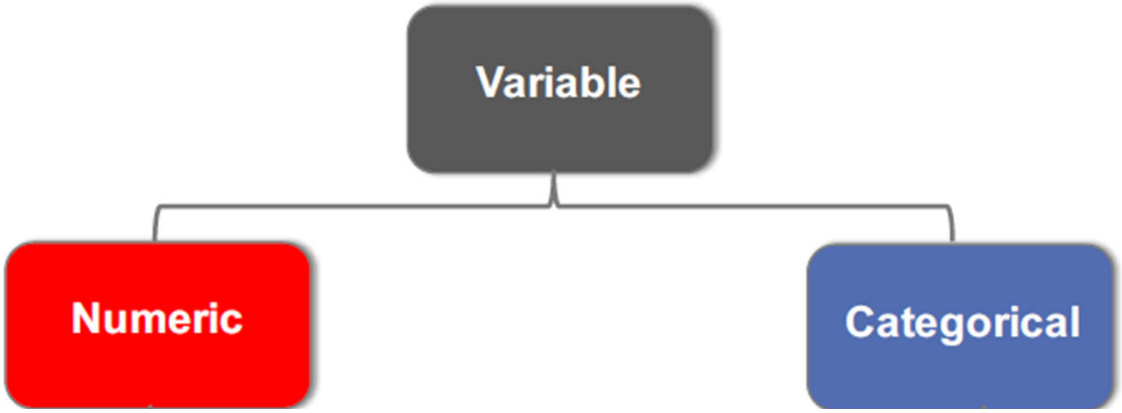
Adapted from Adhikari, DeNero, Wagner, Milner

Announcements

- Please check Week 15 agenda on NYU Classes
  - Exam
    - Monday May 11
    - Gradescope
  - Project
    - Friday May 8

Review

Statistical Data Types not Computational Data Types



- We study data with different properties. We divide these properties into two types
  - Numbers
    - We call it Quantitative Data
  - Categories
    - We call it Qualitative Data

Agenda

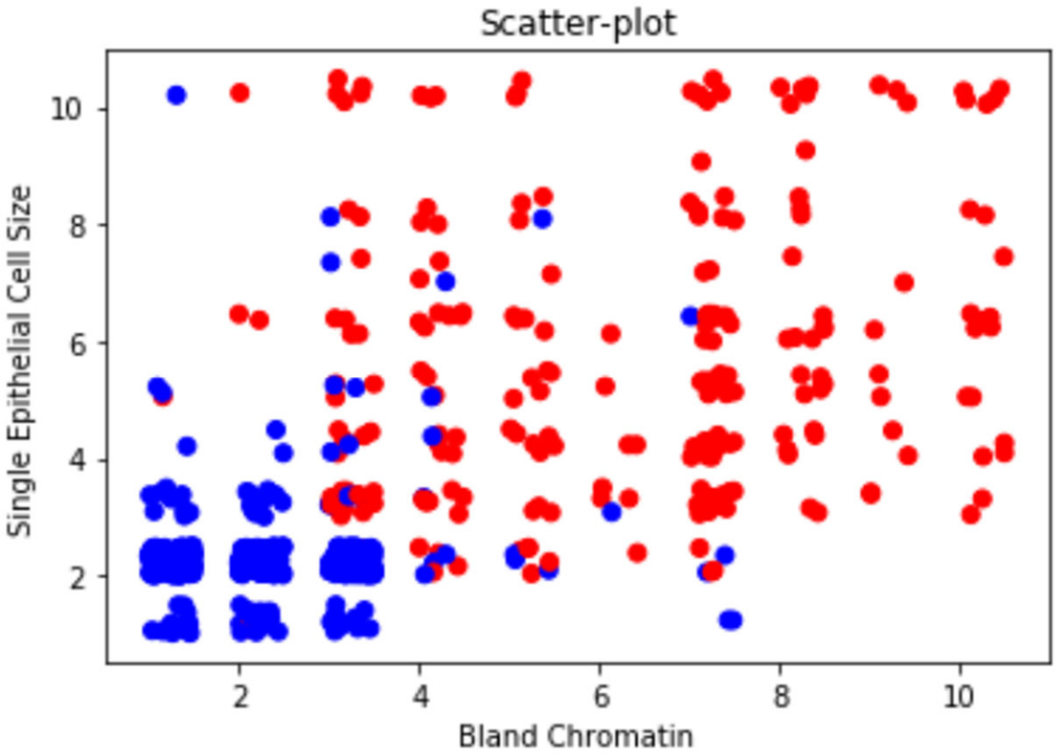
- Nearest Neighbors
- Training and Testing

References

- Classification
  - Chapter 17.3-17.5

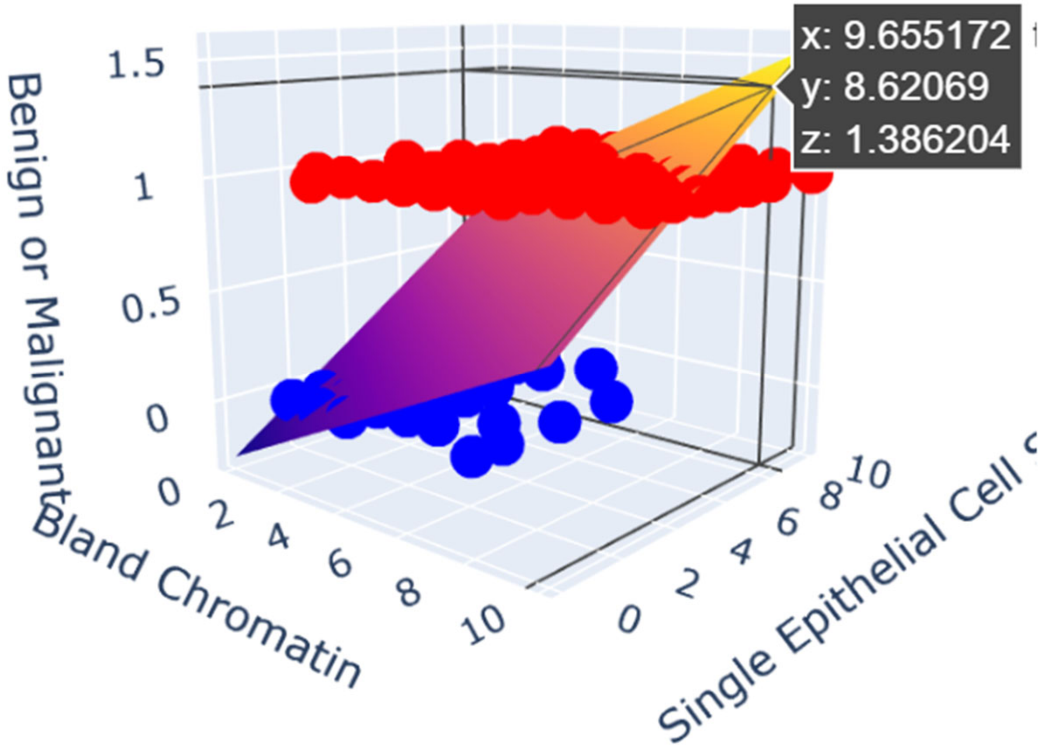
Plotting Categories

- Remember that we use scatter-plots to visualize two quantitative variables.
  - Horizontal Coordinate
  - Vertical Coordinate
- We can incorporate a qualitative variable
  - Color



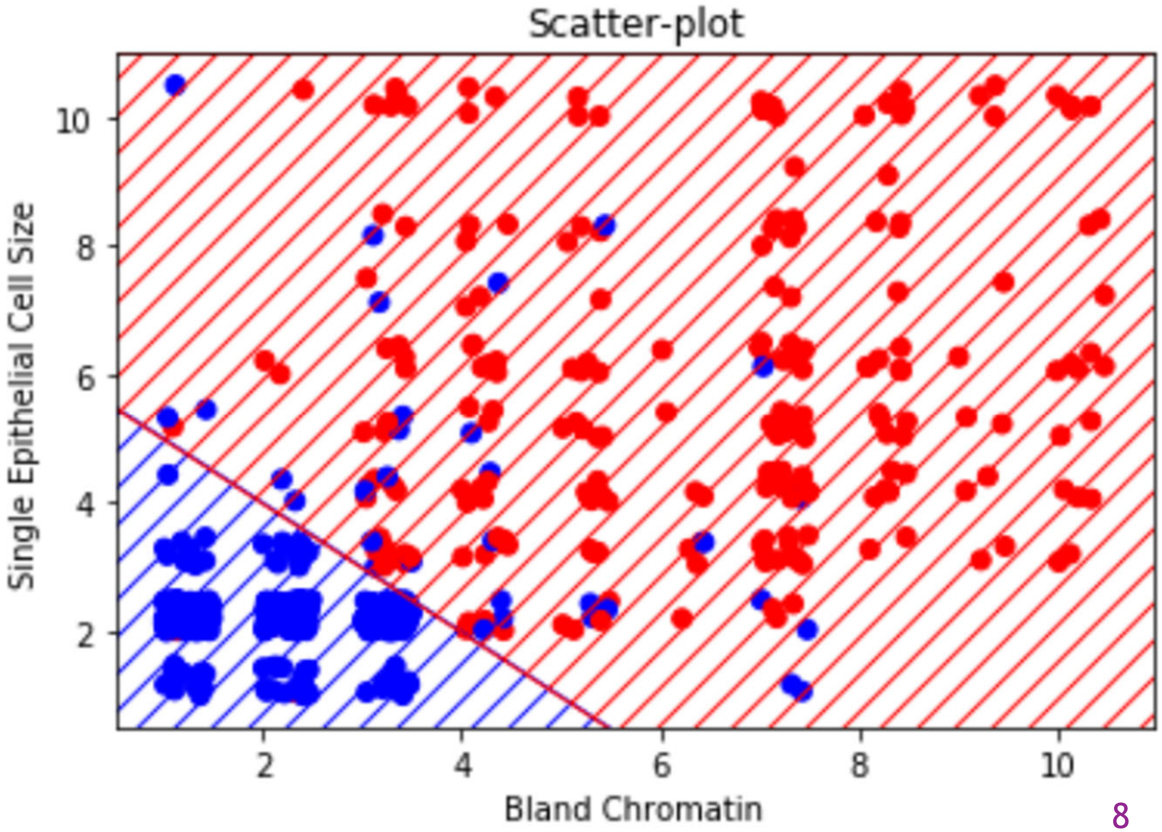
Regression

- Remember that we use regression to predict quantitative response variables from explanatory variables.
- If we pretend that the categories are quantitative data then we could try regression



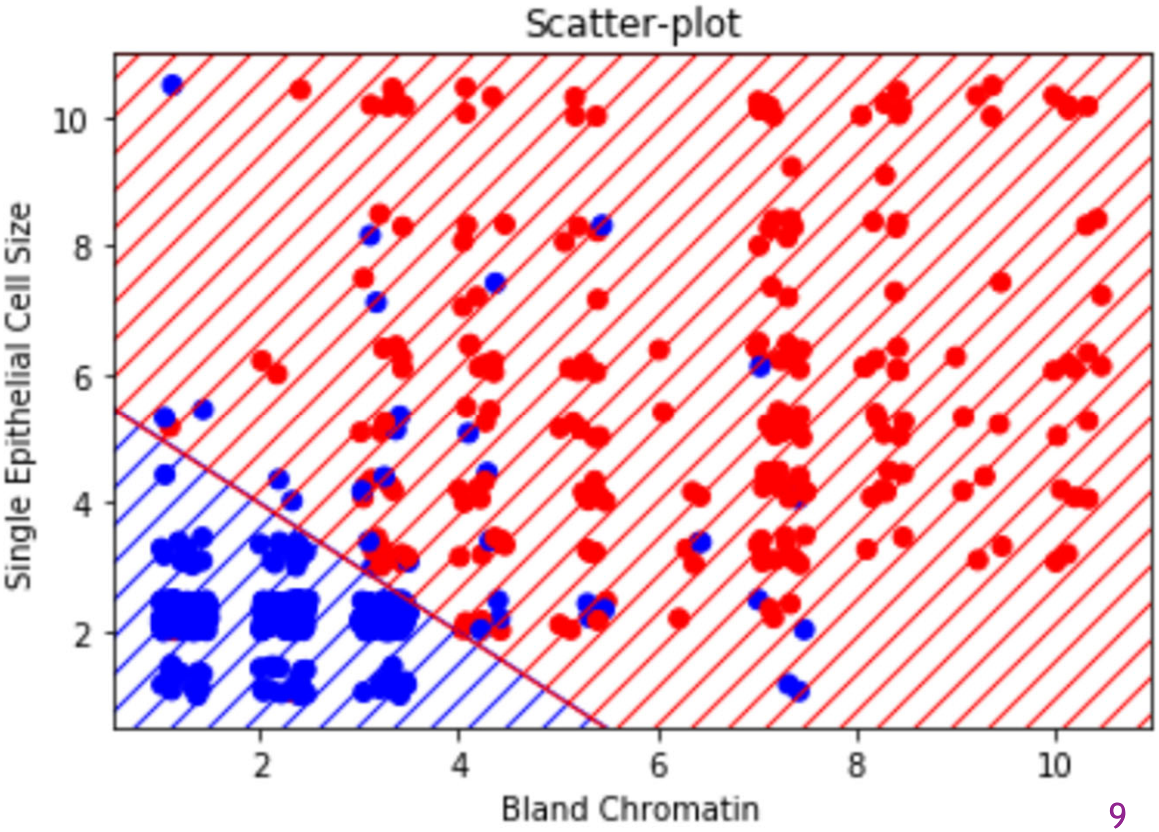
Classification

- We use classification to predict qualitative response variables from explanatory variables.
- Based on the explanatory variables, we separate the data into two regions corresponding to the two categories



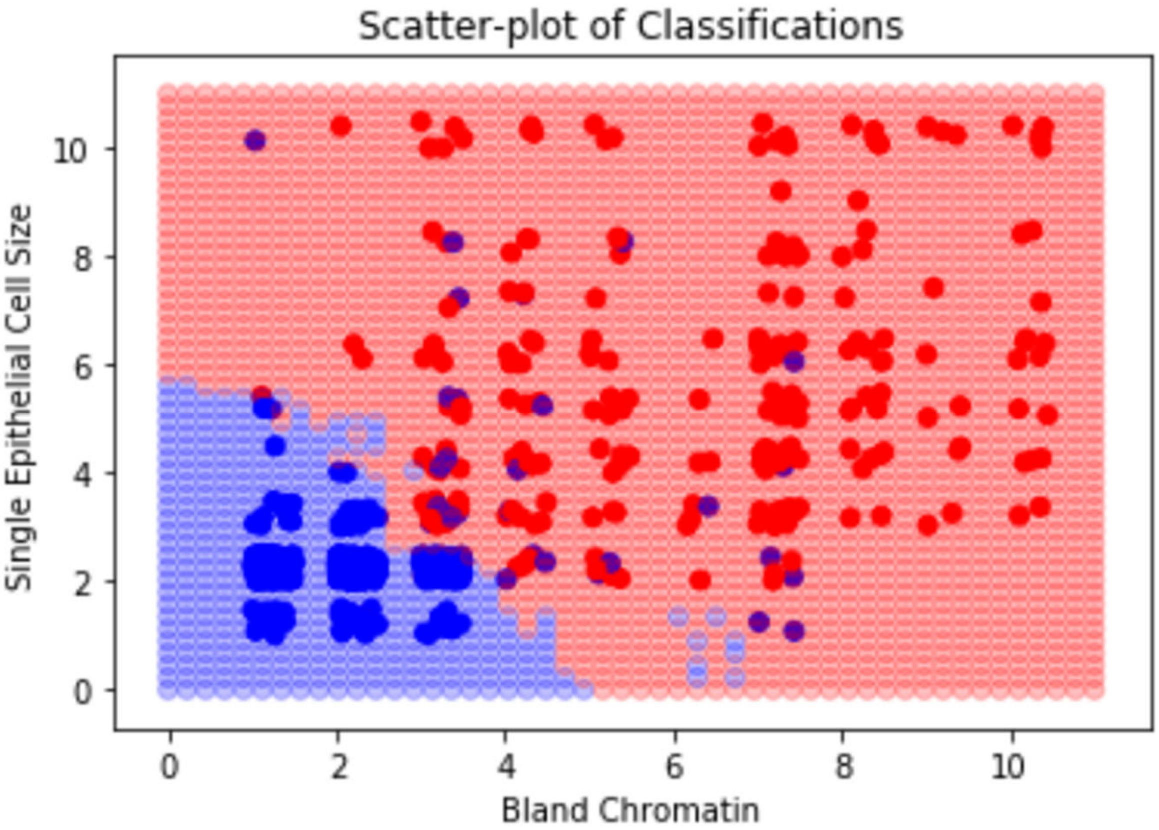
Classification

- We need to use the data to determine a boundary that separates the regions.
- We can compare determining the boundary in classification to fitting the line in regression



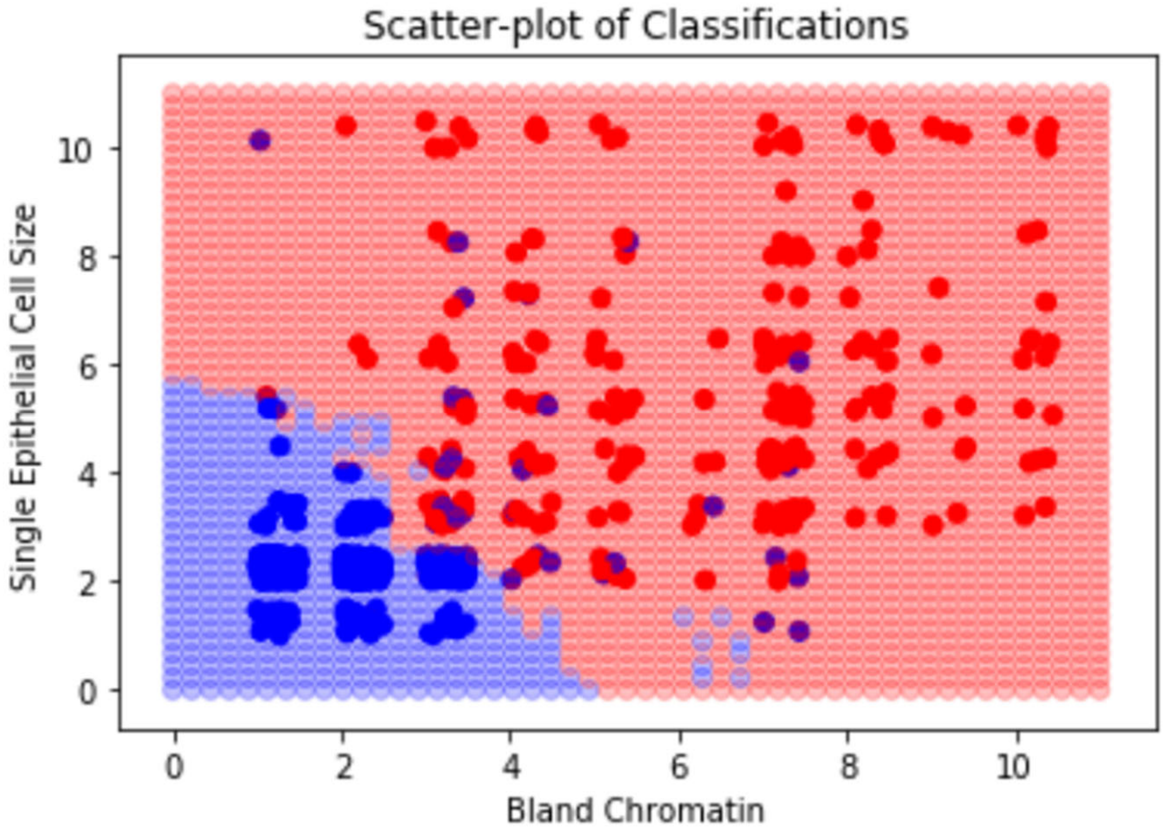
Nearest Neighbors

- 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.



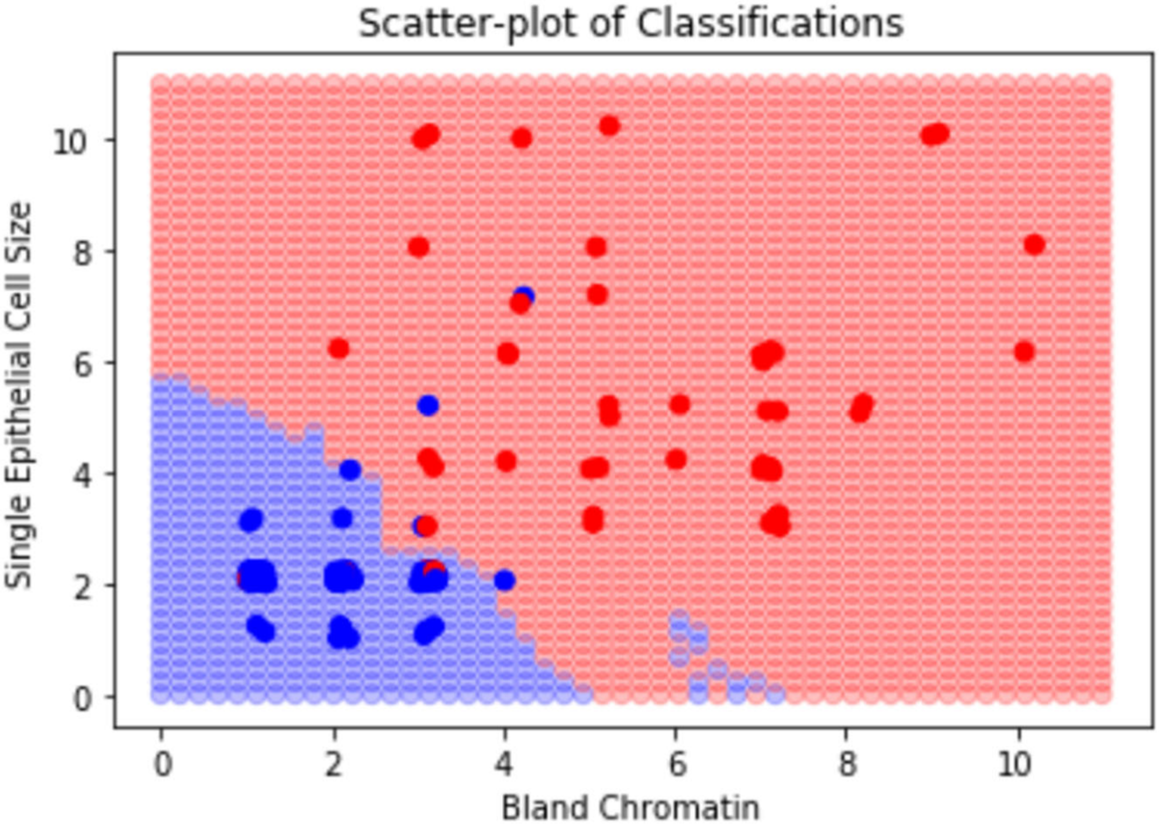
Nearest Neighbors

- 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



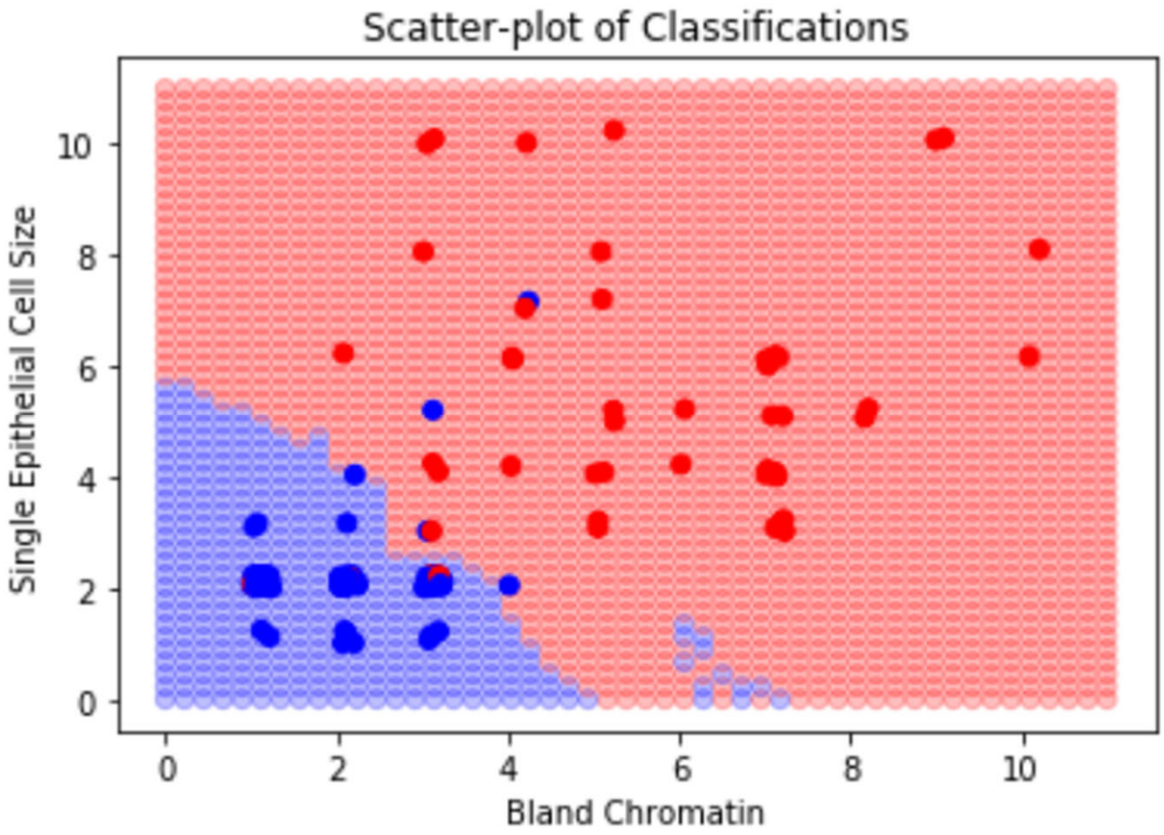
Training and Testing

- Accuracy measures the number of correct predictions
- For evaluating the accuracy, we should randomly split the dataset into 80% training set and 20% testing set



Training and Testing

- ▶ We determine the boundary on the training set
- ▶ We calculate the accuracy on the testing set.
- ▶ We should contrast *in-sample* accuracy and *out-of-sample* accuracy



Summary

- Nearest Neighbors
- Training and Testing

Goals

- Understand the nearest neighbors approach to classification into two categories.
- Randomly split a dataset into a training set and testing set