DSC 10, Spring 2018
Lecture 27

Accuracy of Classifier, Multiple Regression

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Announcements

- Project 10 due Saturday
  - Please re-click download link to get updated tests
  - See Piazza to fix known issue with questions 3.1, 3.2
- Please fill out CAPE
  - This is a new major and your feedback matters!
- Friday's lecture will review for the final

Measuring Accuracy

Accuracy of Classifier

- What fraction of individuals does it classify correctly?

- Need to compare:
  - Classifier's predictions
  - True classes of individuals

- For this, need to know the true classes. But we only know those for the training set. So now what?

The Test Set

- Split original training set at random into two sets
- Use one of the sets for training:
  - Explore as much as you want
  - Develop classifier
- Use the other set (test set) to compare the classifier's predictions and the true classes

(Demo)

Discussion Question

```
def evaluate_accuracy(training, test, k):
    test_attributes = test.drop('Class')
    def classify_testrow(row):
        return classify(training, row, k)
    c = test_attributes.apply(classify_testrow)
    return count_equal(c, test.column('Class')) / test.num_rows
```

What is the type of the `test_attribute` variable?

A. Number
B. Array
C. Table
D. Row
E. List

Discussion Question

```
def evaluate_accuracy(training, test, k):
    test_attributes = test.drop('Class')
    def classify_testrow(row):
        return classify(training, row, k)
    c = test_attributes.apply(classify_testrow)
    return count_equal(c, test.column('Class')) / test.num_rows
```

What is the purpose and return type of the `classify_testrow` function?

A. Predicts a class for one row, returns a number
B. Predicts a class for the table, returns an array
C. Predicts a class for one row, returns an array
D. Predicts a class for the table, returns a number
E. None of the above

Discussion Question

```
def evaluate_accuracy(training, test, k):
    test_attributes = test.drop('Class')
    def classify_testrow(row):
        return classify(training, row, k)
    c = test_attributes.apply(classify_testrow)
    return count_equal(c, test.column('Class')) / test.num_rows
```

What is the type of the variable named `c`?

A. Number
B. Array
C. Table
D. Row
E. None of the above

Discussion Question

```
def evaluate_accuracy(training, test, k):
    test_attributes = test.drop('Class')
    def classify_testrow(row):
        return classify(training, row, k)
    c = test_attributes.apply(classify_testrow)
    return count_equal(c, test.column('Class')) / test.num_rows
```

How does this function measure accuracy?

A. The number of 1's in the column('Class')
B. The number of 0's in the column('Class')
C. The number of rows where actual and predicted values are the same
D. The proportion of rows where actual and predicted values are the same
E. None of the above

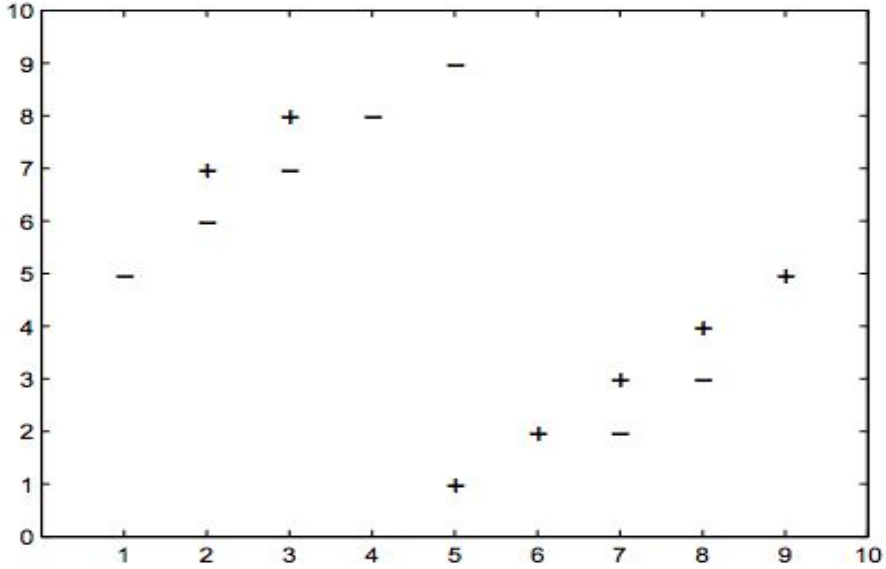
(Demo)

Discussion Question

Suppose you want to test your classifier using the training set. One point becomes a *test* point and everything else is *training*. Then you repeat until each point has been the unlabeled test point once.

What value of k will give us the largest error (number of misclassified labels)?

A. 0
B. 1
C. 5
D. 13



Discussion Question

When we run a computer program, we'd like it to run as fast as possible. k-NN algorithm has two stages: *training* and *testing*.

Which stage will take longer to run: training or testing?

A. Training
B. Testing
C. Same time for both
D. Depends on the problem

Multiple Regression

Prediction Problems

- Classification: predicted variable is categorical
- Regression: predicted variable is quantitative

Regression Methods:

- Simple linear regression
  - Predict using only one quantitative attribute
  - Find a slope and intercept to minimize squared error
- Multiple (linear) regression
  - Predict using multiple quantitative attributes
  - Find many slopes to minimize squared error

Regression Slopes

To predict a house price from its size (X_1) and age (X_2)
A multiple (linear) regression prediction has this form:

$(A_1$ \$/sqft) * (X_1 sqft of space) + (A_2 \$/year) * (X_2 years old) + B

- A_1 and A_2 are both slopes of a line in 3-D
- Each slope describes how much the mean house price increases/decreases for each increase of 1 in an attribute
- The slopes (and intercept) can be chosen together to minimize squared error on training examples

(Demo)

Multiple Regression Cautions

- Slopes are hard to interpret because of correlation between predictor variables
- Very different slopes can give nearly the same error
- For estimates to be reliable across the range of possible inputs, we require a linear association in the population
- There's a lot more to learn about multiple regression

Discussion Question

When we run a computer program, we'd like it to run as fast as possible. Multiple regression has two stages: *training* and *testing*.

Which stage will take longer to run: training or testing?

A. Training
B. Testing
C. Same time for both
D. Depends on the problem