hw3: KNN

Rob McCulloch

February 1, 2018

Contents

Fitting kNN to the Cars Data **1**

Using Cross Validation **1**

kNN, Cars Data with Mileage and Year **2**

Choice of Kernel **2**

Illustrate Bias-Variance Tradeoff with a Simulation **2**

Fitting kNN to the Cars Data

Get the susedcars.csv data set from the webpage. Plot x=mileage versus y=price. (price is the price of a used car.)

Does the relationship between mileage and price make sense?

Add the fit from a linear regression to the plot. Add the fit from kNN for various values of k to the plot.

For what value of k does the plot look nice?

Using your "nice" value of k, what is the predicted price of a car with 100,000 miles on it?

What is the prediction from a linear fit?

Using Cross Validation

We are going to use the used cars data again.

Previously, we used the "eye-ball" method to choose k for a kNN fit for mileage predicting price.

Use 5-fold cross-validation to choose k. How does your fit compare with the eyeball method?

Plot the data and then add the fit using the k you chose using cross-validation and the k you choose by eye-ball.

Use kNN with the k you chose using cross-validation to get a prediction for a used car with 100,000 miles on it. Use all the observations as training data to get your prediction (given your choice of k).

---

kNN, Cars Data with Mileage and Year

Use kNN to get a prediction for a 2008 car with 75,000 miles on it!

Remember:

- Use cross-validation to choose k.
- Scale your x's !!

Is your predictive accuracy better using (mileage,year) than it was with just mileage?

Choice of Kernel

In our class we examples we used kernel="rectangular" when calling kknn.

Have a look at the help for kknn (?kknn).

The rectangular option simple averages the y values over the neighbors.

The other kernel options allow you to use a weighted average. The idea is that closer neighbors should get more weight.

Using the used cars data and predictors (features!!) (mileage,year) see if the optimal kernel option gives different (better?) results than the rectanglular option.

Illustrate Bias-Variance Tradeoff with a Simulation

In class we looked at the script bias-variance-illustration.R.

In that script we learnt about the bias variance tradeoff by repeately subsampling the Boston housing data.

Alternatively we could simulate data from the model

$$Y_i = f(x_i) + \epsilon_i$$

Write a new version of bias-variance-illustration.R which uses simulated data instead of real data.

You will have to make several choices !! Sample size, function f ....

Use your script to explain the bias-variance tradeoff to a friend - *everyone needs to know this !!!*