Assignment 1

September 21, 2020

All files referred to in this homework can be found on CourseWorks.
Your hand-in should be made on Gradescope. You must submit *three* files:

1. A pdf with your answers to written questions.
2. An `.ipynb` file (i.e. jupyter notebook file) with your data analysis for answering questions
3. A pdf file with your jupyter notebook output. Please use the workflow described here for generating this pdf.

1 Lab 3.6 from ISLR

Go through the lab exercise in Section 3.6 of ISLR. The book is written to use the programming language *R* for these exercises. If you like, you can use R to complete these exercises (in this case, I highly recommend using the IDE `rstudio`). Alternatively, since we will later use python for other assignments, I have provided a corresponding set of python commands that you can use. These are given in the form of a Jupyter notebook. I recommend that you open this notebook for reference (run the command `jupyter notebook` in a terminal from the directory containing the notebook, and open the notebook from there), but type all the commands into a new notebook of your own, to make sure that you pay attention to what each cell is doing.

If you use python, you will need the various packages listed at the top of the notebook (these packages are extremely common at companies, becoming proficient with `pandas, numpy, sklearn` is highly recommended). The easiest way to get these is to install the `anaconda` package from here:
`https://www.anaconda.com/products/individual`.
You do not need to turn in your code. The goal of this exercise is to practice your data skills.

Questions

1. Compare the plots of the residuals vs. the fitted values for the regression `medv ~ lstat + np.square(lstat)` and the regression using only `lstat` as a predictor. What's the qualitative difference?
2. Does the fifth-order polynomial from your python regression correspond to the one from the ISLR book? If not, why might this occur?

No. Because poly(lstat, 5) works differently in R.

2 EDA with the Spam Filtering Data Set

The csv file `spam.csv` contains a data set for emails that were categorized as spam or not spam. The documentation for this data set is in the file `spam-info.pdf`.

1. Look at the documentation. What is the variable of interest, i.e. the dependent variable?

'Spam'

2. For each of the independent variables, report something about it. Specifically, you should report on each variable's relationship with the response, i.e dependent, variable. Pay special attention to variable type (binary, ordinal, real) when doing this. Your comments should contain at least some tables and graphs.

3. Investigate the variable 'spampct'.

   (a) How many missing values does it have? 1353

   (b) Compare graphically the distribution for time.of.day for the cases where spampct is missing against the distribution of time.of.day when spampct is present. Do you see any differences?

   (c) Plot a scatter plot of time of day vs. spampct. How many unique points (x,y coordinates) are plotted? Explain a technique you might use to deal with the overplotting.

3 Exploring the Relationship Between Overfitting and Noise

Do exercise 13 from Section 3.7 of ISLR. The example codes are for R, but below I provide a table of translations to python. You will need to use the numpy documentation to look up how to use the various commands. Make sure you look up the documentation for your version of numpy.

R command	python command
`set.seed(1)`	`np.random.seed(1)`
`rnorm()`	`np.random.randn()`
`rnorm()`	`np.random.randn()`

10 Parts in total.
(a),(b) — coding only

4 Naive Bayes and Spam Filtering

1. Use the spam data from Question 2 and Naive Bayes to build a classifier that distinguishes spam from non-spam. You can use Naive Bayes from `sklearn` for this. Your code should split the data into training and test sets and then estimate the generalization error of your classifier.

2. Randomly assign 80% of your data to the training set, 20% to the test set and now estimate the test error, E_{test}, of your classifier. Repeat this 10 times. How much variability do you see in E_{test}? What conclusions can you draw from this?

3. There are two types of error that a spam classifier can make. Should these errors be treated equally when constructing a classifier. Can we adapt our naive Bayes classifier to reflect this?

5 Least Squares Linear Regression is MLE for Gaussian noise

Consider the linear regression model

$$Y = X^T\beta + \epsilon,$$

where $\beta, X \in \mathbb{R}^d$, are fixed, and the error $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is distributed according to a Gaussian distribution.

In class we saw how to derive the least squares estimator. In this exercise, you just must prove that the least squares estimator is also the maximum-likelihood estimator, given that the error is Gaussian.

6 k Nearest Neighbors and the Curse of Dimensionality

Solve exercise 4 from Section 4.7 of ISLR.

5 Parts
(a) to (e)