MIS 382N: Advanced Predictive Modelling

Assignment #2

Due: Wed, Oct 6, 2014 3.30 pm; Total points: 40

Upload instructions: Your homework should be a pdf file containing all the answers, plots/pictures and code. Name the pdf file as <your eid>.pdf You may however insert equations by hand if you wish. The included code should be straightforward to run by just copying it into the interpretrator or the command line without any changes. So add any include statements for libraries etc. in the code as well. Homeworks are due at the beginning of class on the due date. If the choice of a programming language is specified in the problem, please don't use any other language.

Note: All the additional files (shiny1.png, shiny2.png, katarina.csv, congress.csv) needed for this homework are in a zip file named hw2files.zip on canvas.

1. **(4+4=8 pts) Shiny app using R**

   In this problem, you'll build a Shiny application. Shiny is an R package which lets you publish web applications from R easily. For more information on Shiny, see http://shiny.rstudio.com. The problem statement is as follows:

   In 'Katrina.csv' (available on canvas) you have data on 673 businesses in downtown New Orleans. Your goal is to build a Shiny app which allows the user explore the following features present in the dataset: *log_medinc*, $flood\_depth$. The requirements are as follows:

   (a) You will give the user the option to choose between *log_medinc* and $flood\_depth$. The default option should be $flood\_depth$.

   Hint: You can make use of radio buttons.

   (b) Plot a histogram of the feature chosen by the user. You will also give the user the ability to configure the number of bins in the histogram. You should ensure that the number of bins is in the range [6, 18].

   Hint: You can use a slider bar, and the app will immediately respond to the input.

   We have made available sample screenshots of our Shiny app that supports the above requirements, namely *shiny*1*.png* and *shiny*2*.png* (available on canvas). Your interface should look similar to the screenshots.

   The tutorials listed below should provide you the needed background to solve this problem:

   (a) http://shiny.rstudio.com/tutorial/lesson1

   (b) http://shiny.rstudio.com/gallery/faithful.html

   (c) http://www.inside-r.org/packages/cran/shiny/docs/radioButtons

2. **(5 pts) Principal Component Analysis**

   Note: Please use python to implement the coding questions in this problem.You can use `matplotlib` for visualizations.

   Load the data in 'congress.csv' (available on canvas). The rows are members of the 109th U.S. Congress. The first column gives the names of the Congress members.The next 1000 columns are phrases uttered during floor speeches. Entry (i, j) in the matrix is the number of times member i uttered phrase j. The last two columns have information about the party membership and the chamber membership of these members. Our data matrix is obtained by removing the first and the last two columns.

In this question, you will explore an application of PCA, to see if dimensionality reduction on the speaker-phrase matrix can help us in determining the house/party-affiliation using only two principal components. You can use the function `truncatedSVD` to do PCA and project the data onto top-k components. You can look up the details online, but the following commands should be helpful.

i Import the library: `from sklearn.decomposition import TruncatedSVD`

ii Initialize to project onto top-k components: `pcaobject = TruncatedSVD(n_components=np.int(k))`

iii Project data matrix `X` onto the top-k: `Xt = pcaobject.fit_transform(X)`

(a) First center the data matrix: Find the mean vector, and subtract from the data matrix.

(b) We would like to analyze what fraction of the total data variance is explained by each principal component. We will do this via a scree plot. Plot the cumulative variance explained by all k principal components as compared to the total variance. This is called the explained variance ratio. On the same plot, also plot the proportion of variance explained by the k^{th} principal component. Repeat this for the $k = 1, 2, 5, 10, 20, 50, 100, 200$. To get the ratio of the explained variance, you can use: `explained_variances = np.var(Xt, axis=0) / np.var(X, axis=0).sum()`

(c) Now project the data onto the top-2 dimensions. Visualize by labeling each member with his or her party membership. You can do this by doing a scatter plot and using different colors for party membership. Is there any separation in the data?

3. **(3+3+4=10 pts) MLR using Scikit-Learn**

The "Boston Housing" data set, part of the sklearn.datasets, records properties of 506 housing zones in the Greater Boston area. For a description of the data, see http://archive.ics.uci.edu/ml/datasets/Housing Typically one is interested in predicting MEDV (median home value) based on other attributes.

(a) Generate box-plots of the LSTAT (% of lower status in the population) and MEDV (median home value) attributes and identify the cutoff values for outliers.

(b) Generate a scatter plot of LSTAT and MEDV. Plot two models on the same graph (obtained respectively, from the original dataset and the derived one after removal of outliers) to visualize the effect of outliers on the model. (Hint: Such effects may be easier to visualize if the outliers are a different color or symbol than the other data.)

(c) Let us try to fit an MLR to this dataset, with MEDV as the dependent variable. MEDV has a somewhat longish tail and is not so Gaussian-like, so we will take a log transform, and then predict LMDEV instead. (You should convince yourself that this is a better idea by looking at the histograms and quantile plots to assess normality; however no need to submit such plots). Keep the first 350 records as a training set (call it Bostrain) which you will use to fit the model; the remaining 156 will be used as a test set (Bostest). Use only the following variables (put in "R" form for convenience) in your model: $LMEDV \sim LSTAT + RM + CRIM + ZN + CHAS$.

i. Report the MSE obtained on Bostrain. How much does this increase when you score your model on Bostest?

ii. Report the coefficients obtained by learning the regression model.

iii. Do you think your MLR model is reasonable for this problem? You may look at the distribution of residuals to provide an informed answer.

Note: Use the function *linear_model.LinearRegression* present in the package *sklearn* to build the model.

4. **(3+2+2+3=10) Multicollinearity** In this question, we'll use the dataset *mtcars* which is available in R to explore handling multicollinearity using *Ridge* and *Lasso*.

Load the dataset using `data(mtcars)`. Construct the feature matrices `X` and `X2`, and the response vector `Y` as :

```
X = data.frame(disp = mtcars$disp, hp = mtcars$hp, wt = mtcars$wt, drat = mtcars$drat)
X2 = X
X2$alsohp = X$hp
Y = mtcars$mpg
```

Note that `X2` has two columns which have correlation 1. Standardize the data using the function `scale`.

(a) Fit least squares regression, `Ridge`, and `Lasso` on `Y` $\sim$ `X` and `Y`$\sim$ `X2`. Set the regularization parameter to 1. Report the coefficients from all of the above regression experiments (6 of them). Submit the `R` code you wrote.

You can use the commands `lm`, `lm.ridge` and `glmnet`. You might need to install packages `MASS` (for `lm.ridge`) and `glmnet` (for `glmnet`)

(b) For `Y`$\sim$`X`, fit ridge regression by using different values of the reguralization parameter: 0,1,10,100,1000. Report the reconstruction error, and the 2-norm of the coefficient vector for each case. Comment on any pattern that emerges for the error and the norm.

(c) Consider the least squares fit of `Y`$\sim$`X2` and `Y`$\sim$`X`. Recall the closed form formula for least sqaure fit, and use it to explain the learnt coefficients for these two experiments.

(d) `Ridge` and `Lasso` add regularization to handle the multicollinearity observed in part (b). Recall the bias-variance tradeoff. By adding regularization, do we increase or decrease bias? What about variance? Explain.

5. **(2+2+3=7 pts) Bias-variance tradeoff**

(a) State briefly what you understand by the bias-variance tradeoff.

(b) For a given model and problem, what happens to these two quantities when the amount of training data available decreases?

(c) Suppose you want to approximate the pdf of a continuous random variable X, that takes on values over the interval (a,b), as follows: Get N i.i.d samples of X; bin the interval into k equi-spaced bins, and construct a histogram, which you then normalize so that total area under the histogram is 1. This normalized histogram will be an approximation of the true pdf. Clearly the histogram will change if you repeat this experiment using another N samples; hence you can consider the quality of the solution in term of the "mean" histogram (bias) and the variations among the histograms (variance). Qualitatively explain how you would expect the bias-variance tradeoff to be reflected in this situation, as a function of "k".