STATS216v Introduction to Statistical Learning
Stanford University, Summer 2017

Problem Set 1

Due: Friday, July 7

Remember the university honor code. All work and answers must be your own.

---

1. In each of the following scenarios first explain wether it is a supervised learning problem or an unsupervised learning problem. In the case of supervised learning, specify wether it is regression or classification and provide the the number of observations and predictors.

   (a) A study in the Netherlands compares the perfomance of students from refugee families and students from Dutch families. In this study for each of 500 students the researchers have grades point average of the students, their country of origin, family's wealth and parent's education.

   (b) A group of Stanford students want to make a mobile application that suggests an entertaining actvity from 3 different options to their users based on their current location, age, gender, level of education and the time of the day. Also they have access to the 3000 reviews that people who have done those activities in the past 2 months.

   (c) Hospitals have supplies of human blood to use during emergency medical procedures. Some parts of the human blood can last at most five days before it has expired. To prevent blood shortages and wasted blood, a group of statisticians want to know how many bags of blood a hospital will need for each day. For this pupose they have monitored three hospitals for a time period of 2 years. Each day, they record the amount of blood used, the number of patients in the trauma ward, the number of scheduled surgeries, day of the week, and wheter or not it was a holiday.

   (d) A legal firm wishes to gain a better understanding of the types cases that the firm has handled in the last 10 years. A tech-savy lawyer has created a database of with the text of the legal briefs associated with each of the 6000 cases that the firm handled.

   (e) Oil excavation is a very expensive process and the oil resources are not distributed uniformly in an area, so it is important to find the best spots for for oil extraction. To do this engineers consider a very coarse grid (each edge length is of order of miles) and dig a well in the vertices of the grid and take a sample of the sand of the grid points. In one example, 28 different measurements are taken from each sand sample. An engineer has sand samples for 35 locations where they know the results of the digging (how much oil was present at that location). Additionally, the engineer has sand samples for 80 prospective well locations, and wishes to find the most promising spot to dig a future well.

(f) During the world war II when the US was doing nuclear experiments in the Nevada deserts people of the neighborhood area could feel the ground shakings as a result of the explosions and had been thinking that they were ordinary earthquakes. Today, the US wants to see which of the earthquakes in the year 2017 in South Korea are real earthquakes and which are caused by nuclear testing in North Korea. They have historical data from 100 earthquakes in South Korea before North Korea began nuclear testing, as well as measurements from the 18 nuclear tests in the Neveda desert. Each data point is a time series of earthquake intensity sampled 10 times per second for 60 seconds.

2. (a) How would you explain the difference between supervised and unsupervised learning? Can you give an example for each case that describes the differences?

   (b) How would you explain the difference between regression and classification? Can you give an example where a regression methods *could* be used for classification? An example where a classification method *could not* be used for regression.

3. You are a data science consultant! In each of the following cases decide whether you would suggest a flexible regression model or an inflexible one. Provide your reasons for your client as clearly as possible.

   (a) In the study of breast cancer, scientist are trying to find the associated genes. The total number of genes in the study is 50000 and the number of patients is 120.

   (b) The *ministry of education* in a certain country wants to identify student who need extra help. They wish to design a system which estimate student performance in the final 8th grad math exam based on their math, science and history grades in the 7th grade. To do this they want to run a regression on the data consisting of the information of all the students who have graduated from the 8th level in the last 10 years.

   (c) Kelly is a very hardworking chemistry student and she has run an experiment to find a mathematical expression that explains the speed of corrosion of iron according to the humidity and temperature of the envirnoment and the percentage of different elements in the alloy. Unfortunately the lab that she is working in was established in 1967 and the equipment has not been changed seen then. She is skeptical about the quality of her measurements of the speed of corrosion.

   (d) Kelly's advisor won the chemistry nobel prize and used the prize money to outfit the lab with the most modern equipment. Kelly has run her experiments again with the new equipments and now she can trust her numbers. But her advisor believes that she should not expect that the real relationship be linear.

4. This exercise relates to the `College` data set, which can be found in the file `College.csv`. It contains a number of variables for 777 different universities and colleges in the US. The variables are:

   - `Private` : Public/private indicator

- `Apps` : Number of applications received
- `Accept` : Number of applicants accepted
- `Enroll` : Number of new students enrolled
- `Top10perc` : New students from top 10% of high school class
- `Top25perc` : New students from top 25% of high school class
- `F.Undergrad` : Number of full-time undergraduates
- `P.Undergrad` : Number of part-time undergraduates
- `Outstate` : Out-of-state tuition
- `Room.Board` : Room and board costs
- `Books` : Estimated book costs
- `Personal` : Estimated personal spending
- `PhD` : Percent of faculty with Ph.D.’s
- `Terminal` : Percent of faculty with terminal degree
- `S.F.Ratio` : Student/faculty ratio
- `perc.alumni` : Percent of alumni who donate
- `Expend` : Instructional expenditure per student
- `Grad.Rate` : Graduation rate

Before reading the data into `R`, it can be viewed in Excel or a text editor.

(a) Use the `read.csv()` function to read the data into `R`. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data.

(b) To acces documentation within `R`, you can use the `?` operator. As an example, try `? read.csv`. Based on this documentation, if our file did not have a header row how should we modify the call to `read.csv` in part a?

(c) Look at the data using the `View()` function. You should notice that the first column is just the name of each university. We don’t really want `R` to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
> rownames(college)=college[,1]
> fix(college)
```

You should see that there is now a `row.names` column with the name of each university recorded. This means that `R` has given each row a name corresponding to the appropriate university. `R` will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```
> college=college[,-1]
> View(college)
```

Now you should see that the first data column is `Private`. Note that another column labeled `row.names` now appears before the `Private` column. However, this is not a data column but rather the name that `R` is giving to each row.

(d) Use the `summary()` function to produce a numerical summary of the variables in the data set.

(e) Use the `plot()` function to produce a scatterplot of the column `PhD` versus the column `Grad.Rate`.

(f) Use the `which()` and `length()` function to see how many of these colleges are private.

(g) Create a new qualitative variable, called `Elite`, by *binning* the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

```
> Elite=rep("No",nrow(college))
> Elite[college$Top10perc >50]="Yes"
> Elite=as.factor(Elite)
> college=data.frame(college,Elite)
```

Use the `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

(h) Use the `hist()` function to produce 4 histograms with differing numbers of bins for 4 different quantitative variables. Use the command `par(mfrow=c(2,2))` to divide the print window into four regions so that four plots can be made simultaneously.

Note: modifying the arguments to the `par(mfrow=c())` command can be used divide the screen in other ways, which will be useful in future assignments.

5. For this problem, we will consider the `ALS` dataset. ALS (amyotrophic lateral sclerosis), or Lou Gehrig's disease, is a fatal neurodegenerative disease with no known cure and few known causes. In July of 2012, Prize4Life launched a challenge to most accurately predict the rate of progression of ALS in patients.[1] The ALS Prediction Prize4Life Challenge featured a subset of the PRO-ACT database, the largest compilation of ALS clinical trial data ever assembled. The `ALS` dataset is drawn from this challenge, and is worth getting acquainted with: it will be used in the four problem sets in this course. Our goal will be to predict the rate of progression of ALS in patients.

   The `ALS` dataset is composed of four objects: the recorded values in `train.y` and `test.y` represent the rate of change of that patient's ALS Functional Rating Score (a 40-point measure of a person's ability to carry out everyday tasks like walking,

[1] Stanford statistics professor Lester Mackey, working with Lilly Fang, was awarded first prize in the challenge. This exercise and the subsequent ones using this dataset are based on his work. You can read more about it at http://stanford.edu/~lmackey/alsprize4life.html

speaking, swallowing, dressing, etc.) over the final 9 months of the trial. The values in `train.X` and `test.X` contain the predictors we will be using — there are 323 of them.

(a) First, load the `als.RData` file using the `load` command. Make sure that you have the directory set to the correct location for the data.

(b) Each entry in `train.y` or `test.y` corresponds to an ALS patient in a 12-month clinical trial. What are the lengths of each object? Use the `length()` function.

(c) Look at the information in `summary(train.y)`, and produce a histogram of `train.y`, and use the parameter `breaks=40` to ensure you have enough bins. What are some striking features of the distribution of the target values in `train.y`?

(d) Investigate the names of the first few columns in `train.X` and look at some pairwise plots by trying

```
> colnames(train.X)[1:20]
> pairs(train.X[,21:25])
```

(e) Fit a linear regression to your model using only the `train.X` data via the `lm()` function, and call it `lm.model`. It is ok if you get a warning. Since there are too many predictors, look at the first few coefficients fitted and the R^2 statistic using the following commands:

```
> coef(summary(lm.model))[1:20, 1:4]
> summary(lm.model)$r.squared
```

(f) Find the *test* root mean squared error (RMSE) of your model. The `predict()` function will probably be useful.

(g) As we shall later see, the error rate produced by using a simple linear regression on this data is much too high. What could account for this? Try to relate your answer to the bias-variance trade-off.