Machine Learning, Fall 2015, Midterm

- This is an INDIVIDUAL exam. You cannot work in groups.
- The exam must be submitted on chalk before 11:59 pm on Friday Oct 30. This deadline is the same for all sections. You should submit a pdf document on chalk.
- Ask coding questions on Piazza. Do not reveal answers when formulating questions.
- When answering questions, provide plots and supporting analysis. Label plots and axes.
- Be concise.
- Provide supporting R code to help us understand your answers.

1 Question

In the event of a car accident, there may be limited resources available for dealing with the ensuing poperty damage and injuries.

In particular, there may be a limited number of people available with the ability to deliver high end medical attention if serious injuries have resulted from the accident.

It would be useful if we could predict whether or not a serious injury resulted from the accident at the time the accident is reported. This could help us decide what kind of medical personel should be sent out initially. To this end, 42,183 observations have been collected on automobile accidents. The data is in the file Accidents.csv and there are variable definitions in the file Accidents.xls.

For example, MAX_SEV_IR is a categorical variable with the three levels: 0=no injury, 1=non-fatal inj., 2=fatal inj. It would be useful if we could predict MAX_SEV_IR (or some target variable related to it) using variables available at the time the accident is reported.

1.1

Choose a target variable to predict. Explain how predicting this variable might be useful in practice. For example you might just try to predict the three level variable MAX_SEV_IR or you could collapse it down to a binary variable.

1.2

Develop a predictive model for the chosen target. Try to use predictor variables that would be known at the time the decision is made about what kind of team is sent to the accident site.

For example, let's assume that when the accident is reported, you do not know if alcohol is involved.

Justify your predictive model based on its out of sample performance.

1.3

Now suppose estimating the effect of "alcohol involvement" is a major goal of the study.

How can you assess the effect of alcohol on the severity of the accident?

2 Question

In class, we looked at the tabloid data set which had 15,000 observations with 4 predictors and a binary dependent variable indicating whether or not a purchase was made in response to a promotion.

The complete data set has 20,000 observations and 9 predictors.
(variable descriptions are on the next page).

The data is in the file tabdatp9n20.csv.

2.1

Develop a predictive model for the variable purchase which indicates whether the promotion resulted in a purchase (1 if purchase, 0 else).

2.2

In the notes we just used the variable ”nTab”, ”moCbook”, ”iRecMer1”, ”llDol”.

Do the addtional 5 variable help in prediction?

What are the important variables?

2.3

In the notes we tried to evaluate our model by looking at the profit resulting from a strategy that only targets customers who are likely to respond to the promotion (as opposed to exposing each customer to the promotion).

How much profit can be realized from your predictive model by using it to decide the optimal choice of customers to target?

```
variables:

"purchase" : 1 if made purchase from mailed tabloid, 0 else

 "nTab" : number of past tabloid orders

"moCbook"  : months since last tabloid order

"iRecMer1" : 1/ months since last order in merchandise category 1

"propSpec" : proportion of orders which are special orders

"recW4"   : months since last order in Women’s, division 4

"moShoe"   : months since last shoe order

"nWoApp"  : number of purchases in Women’s apparel

"nMen"  : number of purchases in Men’s

"llDol"  : log of the dollar value of past purchases

The data are from a major (unnamed) retailor.
There are 20,000 rows.
```

3 Question

We would like to target some subset of the huge number of visitors to our main retail web page with a new special offer. Instead of the normal early May special offer of a discounted flower bouquet for Mom, weve decided to offer select customers a 30electric razor purchase from our stock.

3.1

Show how computing expected value provides a framework for thinking about what models need to be built for this problem.

3.2

Specify what models you would build.

3.3

Do you expect to have the data necessary to build these models? If so, from where, if not, what do you propose to do about it?