ASSIGNMENT 2B

These questions are meant to hit a high point of basic machine learning. I want particularly to introduce matrix calculus.

Problem 1. Conditional, marginals, and joint.

In this problem we will consider generating and plotting data from a high school where all 3rd and 4th year students take the SAT. For the past several decades, the school has kept a record of the rounded SAT scores achieved by its juniors and seniors. The possible values of the rounded SAT scores are \{1000, 1100, 1200, 1300, 1400, 1500, 1600\}. Analysis of the records indicates that the distribution of scores for 3rd years is multinomial with parameters [.23, .31, .23, .16, .07, 0, 0]. For 4th years, the parameter values are [0, 0, .15, .23, .31, .23, .08]. We will use these parameter values to generate samples of what the distribution of scores would look like if 40 3rd years and 60 4th years take the SAT this year. Each data point will consist of a pair (SAT Score, year). We will use the `mnrnd` command to sample from the multinomial distributions representing the distribution of SAT Scores in past years. To generate the sample vector of SAT scores for the 40 3rd year students, call `mnrnd` with:

```matlab
n3rd=40; p3rd=[.23 .31 .23 .16 .07 0 0];
```

Likewise, generate the sample vector of SAT scores for the 60 4th years by calling `mnrnd` with:

```matlab
n4th=60; p4th=[0 0 .15 .23 .31 .23 .08];
```

From these samples, make your data points. We will now make plots of the joint distribution \( p(\text{SAT Score}, \text{year}) \), of the marginal distributions \( p(\text{SAT Score}) \) and \( p(\text{Year}) \), and of the conditional distribution \( p(\text{SAT Score}|\text{Year} = 4) \). Our end result will resemble Figure 1.11 in the Bishop book.

We will be using `scatter` to plot the data points, but for the purposes of display we will need to jitter the points. Generate the jitter values using:

```matlab
score_jitter=50*(rand(size(data,1),1)-.5);
year_jitter=.5*(rand(size(data,1),1)-.5);
```

In subplot(2,2,1), plot the jittered data points using `scatter`. Modify the scatter plot with the following commands:

```matlab
axis([950 1650 2.5 4.5])
grid on
set(gca,'XColor',[1 0 0])
set(gca,'YColor',[1 0 0])
set(gca,'GridLineStyle','-')
```

COG 118A Assignment 2b 1
set(gca,'LineWidth',4)
set(gca,'XTick',(9.5:16.5)*100)
set(gca,'XTickLabel',[])
set(gca,'YTick',(2.5:4.5))
set(gca,'YTickLabel',[])
ylabel('Year','Color',[0 0 0])
xlabel('SAT Score','Color',[0 0 0])

In subplot(2,2,2), use the \texttt{hist} command to plot the marginal distribution \(p(\text{Year})\). Specify the bin values to be \{3,4\}. Modify the resulting plot with the following commands:

set(gca,'CameraUpVector',[-1 0 0])
set(gca,'XDir','reverse')
set(gca,'YAxisLocation','right')
axis([2.3 4.7 0 80])

In subplot(2,2,3), use the \texttt{hist} command to plot the marginal distribution \(p(\text{SAT Score})\). Specify the bin values to match the possible SAT Scores: \{1000,1100,1200,1300,1400,1500,1600\}. Set the axis values with

axis([950 1650 0 40])

In subplot(2,2,4), use the \texttt{hist} command to plot the conditional distribution \(p(\text{SAT Score}|\text{Year} = 4)\). Specify the bin values to match the possible SAT Scores. Set the axis values with

axis([950 1650 0 40])

The \texttt{find} command may be useful for selecting the SAT Scores for which Year=4.

Title each subplot according to the distribution plotted therein.

Are SAT Score and Year independent? Your answer should reference \(p(\text{SAT Score})\) and \(p(\text{SAT Score}|\text{Year} = 4)\)

\textbf{Problem 2.} Getting familiar with the beta distribution