HW1

0.1 Visualizing the height/weight data (Matlab)

For the questions below, turn in a printout (hardcopy) of your code and figures/ results. Do not use BLT.

1. Load the data from the file `heightWeightData.txt` using the matlab command `dlmread`. The first column represents class (1=male, 2=female), the second column represents height in inches, the third column represents weight in pounds of some US college students.

2. Make a scatter plot of the data, color coding the points by their class. Label your axes. The result should look something like Figure 1(a). Hint: the following commands may be helpful: `scatter, plot, hold on, find, title, xlabel, ylabel`.

3. Plot histograms of all the male heights, female heights, male weights, and female weights. Ensure the area represented by the histogram sums to one by using

$$\hat{p}(x) = \frac{n_k}{nh} I(x \in B_k) \tag{1}$$

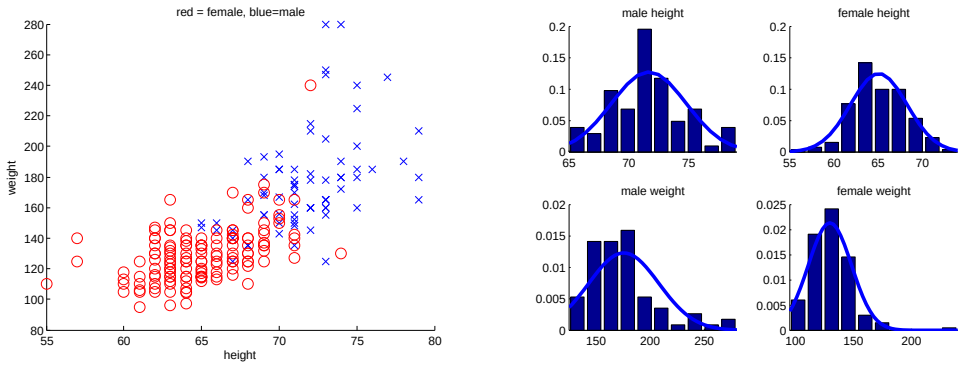
   where B_k is bin k, n_k is the number of points in bin k, h is the width of the bins, and n is the total number of data points. The result should look something like Figure 1(b), but without the superimposed Gaussian curves. Hint: the following commands may be helpful: `subplot, hist, bar`.

4. Compute the empirical mean and standard deviation of the the male heights, female heights, male weights, and female weights. What values of μ and σ do you get? Hint: the following commands may be helpful: `mean, std, var`.

5. Plot Gaussians with the specified μ and σ on top of the histograms. The result should look something like Figure 1(b). Hint: the following commands may be helpful: `normpdf, linspace`.

0.2 Bayes rule

(Source: Koller)
After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of tetsing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)



(a) (b)

Figure 1: (a) Scatter plot of height/weight data. Blue cross = male, red circle = female. (b) Histogram of marginals, with fitted Gaussians superimposed

0.3 Conditional independence

(Source: Koller)

1. Let $H \in \{1, \ldots, K\}$ be a discrete random variable, and let e_1 and e_2 be the observed values of two other random variables E_1 and E_2. Suppose we wish to calculate the vector

$$\vec{P}(H|e_1, e_2) = (P(H = 1|e_1, e_2), \ldots, P(H = K|e_1, e_2))$$

   Which of the following sets of numbers are sufficient for the calculation?

   (a) $P(e_1, e_2)$, $P(H)$, $P(e_1|H)$, $P(e_2|H)$
   (b) $P(e_1, e_2)$, $P(H)$, $P(e_1, e_2|H)$
   (c) $P(e_1|H)$, $P(e_2|H)$, $P(H)$

2. Now suppose we now assume $E_1 \perp E_2|H$ (i.e., E_1 and E_2 are conditionally independent given H). Which of the above 3 sets are sufficent now?

Show your calculations as well as giving the final result. Hint: recall Bayes rule

$$P(H|\vec{e}) = \frac{P(\vec{e}|H)P(H)}{P(\vec{e})}$$

0.4 The Monty Hall problem

(Source: Mackay)
On a game show, a contestant is told the rules as follows:

> There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will *not* be opened. Instead, the gameshow host will open one of the other two doors, and *he will do so in such a way as not to reveal the prize.* For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed.
>
> At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference? You may assume that initially, the prize is equally likely to be behind any of the 3 doors. Hint: use Bayes rule.