

# Key

## EE435 Spring 2012 PS02 (Problem Set 02)

Due: Wednesday 2/1/2012

Download the file from the course website called PS01-MidData.xlsx. This is a compilation of all the data I received from you provided from problem set 1, as it was sent to me. Using this data, answer the following questions:

1. Why will some values of PRT run time cause a problem? What can be done about it?

Time was supposed to be entered in secs, but someone made some minute : second entries. These entries must be converted to seconds in order to use as this feature.

3. Why will some entries for "Overall PRT Score" cause a problem? What can be done about it?

They were left blank. For those particular people, that data is unavailable. That feature cannot be used for those individuals.

4. Create and printout a well-labeled histogram for males and females for "Run Time", after fixing any issues in the data. Who runs faster? Where would you set a threshold to decide male or female for a new sample that you don't know if it was a male or female? Using this threshold, how many errors would you have?

plot attached - I chose a threshold of 600, and it resulted in 16 male errors and 22 female errors total 38 errors. Yours may vary

5. Create and printout a well-labeled scatter plot for "# Situps" (feature #1) vs. "# Pushups" (feature # 2). Choose a suitable linear boundary as a decision boundary, and add this line to your plot (using MATLAB code). Using this boundary, write MATLAB code that will automatically determine the number of errors you have. Record your decision rule, based on your boundary, below.

$$f_1 = \text{situps}, f_2 = \text{pushups}$$

If  $f_2 + \frac{1}{2}f_1 - 140 \geq 0$  choose male (class 1)

Decision rule:  
If  $f_2 + \frac{1}{2}f_1 - 140 < 0$  choose female (class 2)

Results in 23 errors

---

Note: Some of the recorded data is inaccurate, if you look at the scatter plot. Someone recorded 6.5 situps, and 10 situps in the data turned in. Overall, this and the answers to questions 1 and 2 give you exposure to problems with using real data that you may run into.

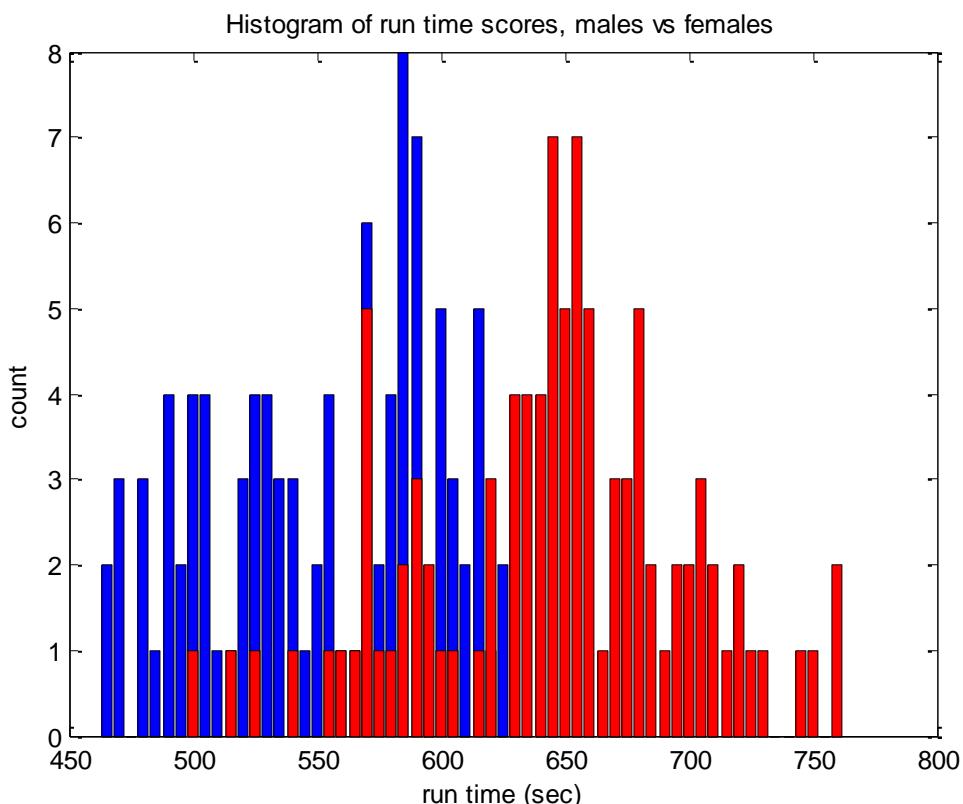
my decision boundary has the equation

$$f_2 = -\frac{1}{2}f_1 + 140$$

#### Problem # 4

```
% plot histogram of run times
% for the following code, "gt" = ground truth, "rt" = run time
edges=min(rt):5:max(rt);
h1=histc(rt(find(gt==1)),edges);
h2=histc(rt(find(gt==2)),edges);
figure(1)
bar(edges,h1,'b'),hold on, bar(edges,h2,'r')
xlabel('run time (sec)')
ylabel('count')
title('Histogram of run time scores, males vs females')
hold off

thresh=600; % the threshold I chose for recognition. Males looked to be faster
% so their times would be lower
index=find(rt(find(gt==1)) > 600);, male_errors = length(index)
index=find(rt(find(gt==2)) <= 600); female_errors=length(index)
```



**Problem # 5**

```
% for the following code, "gt" = ground truth, "s" = situps, "p"=pushups
figure(2)
plot(s(find(gt==1)),p(find(gt==1)), 'ro', s(find(gt==2)),p(find(gt==2)), 'bx')
xlabel('# situps'), ylabel('# pushups'), legend('males', 'females'), grid on
title('Scatter Plot for EE435, PS02')
hold on

% now add the decision boundary to the plot
f1=40:120;
f2=-0.5*f1+130;
plot(f1,f2, 'k', 'linewidth', 2)
hold off

% now apply decision rule using our situp and pushup data
numerrors=0;
for k=1:length(gt)
    class=1;
    test=p(k)+0.5*s(k)-140;
    if test < 0
        class=2;
    end
    if class ~= gt(k)
        numerrors = numerrors+1;
    end
end
disp(sprintf('Number of errors using situps/pushups = %d\n', numerrors))
```

