Problem Set 1
Due Monday, April 4, 2016 at 11:55pm

How to Submit

Create one .zip file (**not** .rar or something else) of your code and written answers and submit it via `ilearn.ucr.edu`. Your zip file should contain `plotdata.m`, `irissep.m`, and a file of your written answers for problems 3 and 4. Please submit your written answers in a pdf or ascii text file, not a word document.

Each file should include at the top (in comments if necessary)

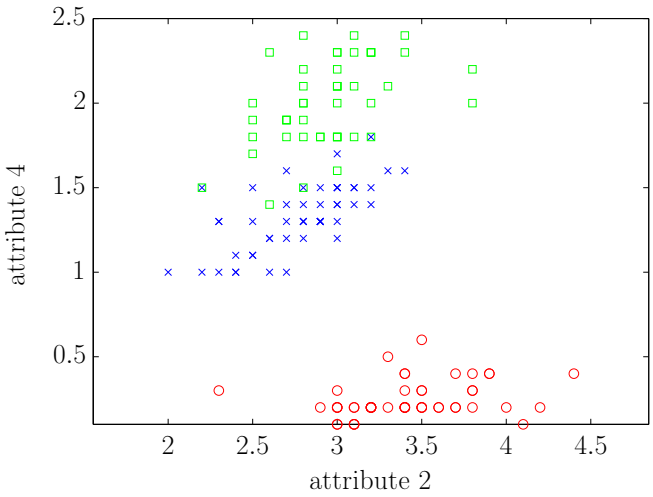
- Your name
- Your UCR student ID number
- The date
- The course (CS 171)
- The assignment number (PS 1)

Problem 1. [15 pts] The data set `iris.dat` is supplied with this problem set. Each row is an example. There are four attributes (all in centimeters), each given by one of the first four columns. In order, they are (1) the sepal length, (2) the sepal width, (3) the petal length, and (4) the petal width. The last column is the class or target (0 for Setosa, 1 for Versicolor, and 2 for Virginica).

Write a Matlab function `plotdata(fname,a1,a2)` that takes as input three arguments:

- `fname`: the filename of this data (in this case, that would always be `'iris.dat'`),
- `a1`: a feature for the horizontal axis, and
- `a2`: a feature for the vertical axis.

This function should plot (in the current figure) a scatter plot of the values of the feature number `a2` versus the values of the feature number `a1` (that is, `a1` on the horzontal axis and `a2` on the vertical axis). Each point's color and symbol should reflect its class. Setosa should be represented by red circles, Versicolor should be represented by blue xs, and Virginica should be represented by green squares. For instance, calling `plotdata('iris.dat',2,4)` should produce



Problem 2. [20 pts]

The Setosa irises seem simple to separate from the other two classes, so we will ignore them. Instead, we will assume the task is to separate the Versicolor irises from the Virginica irises. Further, we must do it based on only two attributes and by drawing a line. Our line will be parameterized by a (2-dimensional) vector w and a scalar constant b. In particular, a point x is on the line if

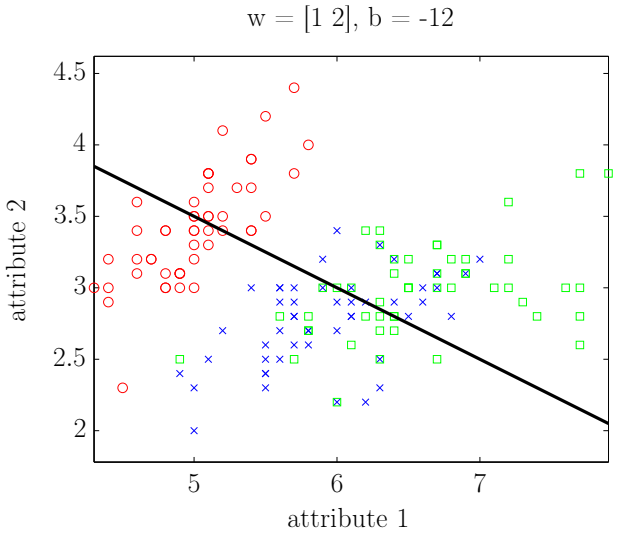
$$w^\top x + b = 0,$$

or, expanding the inner product,

$$w_1 x_1 + w_2 x_2 + b = 0 \ .$$

By hand, pick the two attributes, $a1$ and $a2$, and the line, w and b that you think best separate the two classes of irises. Write a Matlab function `irissep(fname)` that takes the name of the data file, plots these two attributes (using your `plotdata` function from problem 1), draws the line you think is best, and gives the parameters of the line in the title of the plot. You should use the supplied `drawline` Matlab function to draw the line.

For instance, a poor choice would be to use the attributes 1 and 2 and the line parameterized by $w = [1\ 2]$ and $b = -14$. However, if that were your answer, your code should produce the following figure.



Problem 3. [5 pts]

Explain why and how you chose the attributes and line in problem 2.

Problem 4. [10 pts] During your regular medical check-up, your physician orders a regular blood test (that is, a test she orders for everyone having an annual check-up) to check for "a new horrible disease" that was only discovered since your last check-up. This test has a false-positive rate of 2% (that is, if you don't have the disease, there is a 2/100 chance that the test will come back positive) and a false-negative rate of 0.1% (that is, if you do have the disease, there is a 1/1000 chance that the test will come back negative). The disease is present in 1 out of 3,000 people.

Your blood test comes back positive. What is the probability you have this disease? (Show your calculations)