Assignment 2

Due: Oct. 18

1. (10 points) Sign up for project teams in this shared document:
   https://docs.google.com/a/fordham.edu/spreadsheets/d/1FhOcHYfzBsw0FvduOIKL3XosEgEzk_arw4CGCvO4WiY/edit?usp=sharing
   Each team should have 1- 4 members. Please fill in the last two columns indicating if your team is open for more members and what programming language(s) your team is planning to use.

2. (40 points) Given 20 data points and their class labels as follows:

#	x1	x2	class
1	2.46	2.59	1
2	3.05	2.87	1
3	1.12	1.64	1
4	0.01	1.44	1
5	2.20	3.04	1
6	0.41	2.04	1
7	0.53	0.77	1
8	1.89	2.64	1
9	-0.39	0.96	1
10	-0.96	0.08	1
11	2.65	-1.33	-1
12	1.57	-1.70	-1
13	3.05	0.01	-1
14	2.66	-1.15	-1
15	4.51	-0.52	-1
16	3.06	-0.82	-1
17	3.16	-0.56	-1
18	2.05	-0.62	-1
19	0.71	-2.47	-1
20	1.63	-0.91	-1

   Suppose by solving the dual form of quadratic programming problem, we obtained the α's for each data point as follows:

   $$\alpha_7 = 0.4952, \qquad \alpha_{18} = 0.0459, \qquad \alpha_{20} = 0.4493, \qquad \text{Others} = 0$$

   (a) Point out the support vectors in the training data.
   (b) Calculate the normal vector of the hyperplane: $\mathbf{w}$
   (c) Calculate the bias (intercept) term b
   (d) Write the the learned decision boundary function $f(x) = \mathbf{w}^T\mathbf{x}+b$ (the hyperplane) by substituting the learned values of $\mathbf{w}$ and b in the formula.
   (e) Suppose there is a new data point $\mathbf{x} = (-1, 2)$, use the decision boundary to predict its class label.

3. (20 points) SVM using Weka

   For this exercise, we apply SVM with several different kernels and hyper-parameter choices to the **veh-prime.arff** file provided with the assignment. Import this file into Weka (free download from http://www.cs.waikato.ac.nz/ml/weka/) and then select the SMO classifier found under classifiers/function. Use 10 fold cross-validation. You can make kerrnel and hyper-parameter choices by clicking on "SMO ..." appearing next to Choose.

   You will make 5 runs of the algorithms. Select PolyKernel with exponent option 1, 2, and 4. Then select RBFKernel with gamma set to 0.01 and 1.0. For each run record the number of correctly and incorrectly classified instances. Explain why some of the choices do not work well.

4. (30 points) Build a Naive Bayes classifier for the given training data with **add 1 smoothing** technique covered in the lecture slides:

Instance	Education Level	Career	Years of Experience	Salary
1	High School	Management	Less than 3	Low
2	High School	Management	3 to 10	Low
3	College	Management	Less than 3	High
4	College	Service	More than 10	Low
5	High School	Service	3 to 10	Low
6	College	Service	3 to 10	High
7	College	Management	More than 10	High
8	College	Service	Less than 3	Low
9	High School	Management	More than 10	High
10	High School	Service	More than 10	Low

   Use your model to classify the following new instances:

Instance	Education Level	Career	Years of Experience
1	High School	Service	Less than 3
2	College	Retail	Less than 3
3	Graduate	Service	3 to 10