CS 165B – Machine Learning, Spring 2016

Assignment #1
Due Friday, April 8 by 4:30pm

Notes:

- *This assignment is to be done individually. You may discuss the problems at a general level with others in the class (e.g., about the concepts underlying the question, or what lecture or reading material may be relevant), but the work you turn in must be solely your own.*
- Justify every answer you give – **show the work** that achieves the answer or **explain** your response.
- Be sure to re-read the “Policy on Academic Integrity” on the course syllabus.
- Be aware of the late policy in the course syllabus – i.e., *late submissions will not be accepted*, so turn in what you have by the due time.
- Any updates or corrections will be posted on the Assignments page (of the course web site), so check there occasionally.
- There are two options for turning in your assignment:
  - Deliver a hardcopy to the homework box in HFH 2108
  - Submit a typeset PDF version to Gauchospace – NO scanned or photographed submissions accepted!

Problem #1 [4 points]
One possible approach in a machine learning problem might be to create a lookup table that saves (input, output) pairs so that, when an input value is obtained, the output can be determined directly from the table. What are the main pros and cons of this approach?

Problem #2 [6 points]
Choose one of the active data science competitions at Kaggle (www.kaggle.com) and describe the core machine learning problem by answering these questions:

(a) What is the task?
(b) What is the experience or data used to learn a model?
(c) What is (are) the primary performance measure(s)?
(d) What prior assumptions are reasonable to make?
(e) What kind(s) of learning will be done – supervised, semi-supervised, unsupervised, reinforcement?
(f) Briefly describe the data that is made available.

Problem #3 [12 points]
You are asked to build a machine learning system to estimate someone's LDL cholesterol level based on the following inputs: the patient's sex, age, weight, average grams of fat consumed per day, number of servings of red meat per week, servings of fruits and vegetables per day, mg of cholesterol consumed per day. You are given a training data set of values for all of these variables and the LDL cholesterol level for 10,000 patients.

Answer (and explain) the following questions:

(a) What kind of machine learning problem is this?
(b) Is it a predictive task or a descriptive task?
(c) Are you likely to use a geometric model, a probabilistic model, or a logical model?
(d) Will your model be a grouping model or a grading model?
(e) What is the label space for this problem?
(f) What is the output space for this problem?

Problem #4 [10 points]
Given two feature vectors,

$$x1 = \begin{bmatrix} 1.7 \\ -0.2 \\ -3.7 \\ 3.4 \\ 1.3 \end{bmatrix} \qquad x2 = \begin{bmatrix} 1.5 \\ 1.0 \\ 0.5 \\ 0.0 \\ -0.5 \end{bmatrix}$$

what is the distance between them, measured in terms of (a) L1 distance, (b) L2 distance, (c) L_{10} distance?
(d) If a constant vector v = [1 0 1 2 -1]^T is added to both $x1$ and $x2$, which (if any) of L1, L2, or L_{10} will change?
(e) If $x1$ and $x2$ are multiplied by a constant k, which (if any) of L1, L2, or L_{10} will change?

Problem #5 [12 points]
The joint probability distribution of three variables, *class*, *grade* and *effort* is given in the following table:

	class = 165B			**class** = basketweaving		
grade	**effort**=Small	Medium	Large	**effort**=Small	Medium	Large
A	0	0.025	???	0.05	0.1	0.15
B	0.025	0.04	0.06	0.05	0.05	0.025
C	0.025	0.05	0.025	0.05	0.025	0
D	0.05	0.02	0.005	0	0	0
F	0.05	0	0	0	0	0

(a) What is the conditional probability distribution P(*grade* | *class*, *effort*)?
(b) What is the marginal probability distribution P(*grade*, *effort*)?
(c) What is the marginal probability distribution P(*effort*)?
(d) What is P(*grade*=A | *class*)?

Problem #6 [12 points]
There are 10,000 images used to train a face detection system – 2,000 of them are of faces and the rest do not contain faces. To test the system, you have 1000 images – 500 faces and 500 non-faces – in your test set.

The results of the test are as follows: 75 of the face images are classified as non-face, and the rest are classified as faces; 125 of the non-face images are classified as faces, and the rest are classified as non-faces.

(a) Show the contingency table for this binary classification experiment. Label it clearly and fill out the table entries.
(b) What is the false positive rate of the system in this experiment?
(c) What is the false negative rate?
(d) What is the error rate?
(e) What is the precision?
(f) What is the accuracy?

Problem #7 [12 points]

Here are some data describing features in a machine learning problem (six instances in a three-dimensional feature space):

#	F1	F2	F3	Label
1	1.8	4	5.6	A
2	4.8	7	8.6	B
3	1.2	3.4	5.0	A
4	4.5	6.7	8.3	B
5	1.5	3.7	5.3	B
6	4.2	6.4	8	B

(a) What is the intrinsic dimensionality of the data?

(b) After dimensionality reduction is applied to the data to produce a transformed feature space, and assuming that these examples are linearly separable into two classes, what is the geometric form of the discriminating function that separates the data? (E.g., a point, a line, a plane, a sphere, a non-linear contour, ….)

(c) Are the examples in this training set linearly separable?