CS165B
HOMEWORK 1
Solutions

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?

Pros (2 points)

1. Faster as there is no training necessary
2. Easy to develop
3. Accurate for instances in the training set.

Cons (2 points)

1. No ability to generalize. It is similar to a hash map. When a new data is provided during test, the output won't exist.

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:
(1 point each)

San Francisco Crime Classification

(a) What is the task?
The task is to predict the category of a crime that occurred in SF given time and location.

(b) What is the experience or data used to learn a model?
SF Open Data. All crimes reported from 1/1/2003 to 5/13/2015 are included.

(c) What is (are) the primary performance measure(s)?
Multi Class Log Loss - with formula and explanation

(d) What prior assumptions are reasonable to make?
Data collected is accurate.

There will be only one crime at a given time and location.

(e) What kind(s) of learning will be done – supervised, semi-supervised, unsupervised, reinforcement?
Supervised learning as labels are provided in training set.

(f) Briefly describe the data that is made available.

The training and test set are rotated every week. There are 8 features that are used to find the category - Dates - timestamp of the crime incident, Descript - detailed description of the crime incident, DayOfWeek - the day of the week,PdDistrict - name of the Police Department District, Resolution - how the crime incident was resolved, Address - the approximate street address of the crime incident, X - Longitude and Y - Latitude

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:

(2 points each)

(a) What kind of machine learning problem is this?
Regression. It's a regression problem as we need to assign the target variable to a real-valued function of the inputs.
(b) Is it a predictive task or a descriptive task?
It's a predictive task, estimating the LDL cholesterol level.
(c) Are you likely to use a geometric model, a probabilistic model, or a logical model? (1 point for answer and 1 for justification)
Any of the models is possible here.
(d) Will your model be a grouping model or a grading model? (1 point for answer and 1 for justification)
Either grouping or grading could work. Grading would be most typical with a geometric or probabilistic model; grouping with a logical model
(e) What is the label space for this problem? (1 point for answer and 1 for justification)

R (Real positive numbers)

(f) What is the output space for this problem?(1 point for answer and 1 for justification)

R (Real positive numbers)

Problem #4 [10 points] Given two feature vectors, $x1$ = [1.7 −0.2 −3.7 3.4 1.3] $x2$ = [1.5 1.0 0.5 0.0 −0.5] what is the distance between them, measured in terms of

(2 points each)

(a) L1 distance,

Formula L1 = Sigma(|xi-yi|)

Apply

Result = 10.8

(b) L2 distance,

Formula L2 = Sigma(|xi-yi|^2)^½

Apply

Result = (0.04+1.44+17.64+11.56+3.24)^1/2=33.92^½

= 5.824

(c) L10 distance?

Formula L10 = Sigma(|xi-yi|^10)^1/10

= 0.2^10 + 1.2^10+ 4.2^10+ 3.4^10+ 1.8^10

= (1.024e-7+6.1917+1708019.8121+206437.7754+357.0467)^1/10

=1914820.8260^(1/10)

Result=4.2483 (approximate)

(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 L10 will change?

None of them

Solve the general equation

L1 = sigma|(xi+ki)-(yi+ki)| =sigma |xi+ki-yi-ki|=sigma |xi-yi| same as previous L1

L2 = (sigma|(xi+ki)-(yi+ki)|^2)^1/2=(sigma |xi+ki-yi-ki|^2)^1/2=(sigma |xi-yi|^2)½ same as previous L2

Similarly for L10

(e) If $x1$ and $x2$ are multiplied by a constant k, which (if any) of L1, L2, or L10 will change?

All of them change.

Solve the general equation
L1 = sigma|(xi*k)-(yi*k)| =sigma |xi*k-yi*k|=sigma k*|xi-yi|.
The L1 distance gets multiplied by k

L2 = (sigma|k*xi -k*yi|^2)^1/2=(sigma (k|xi-yi|)^2)^1/2=k*(sigma |xi-yi|^2)½
The L2 distance gets multiplied by k

Similarly for L10

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	**0.125**	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)?
P(g ^c ^e)/sum(g)(P(g^c^e))

	class =165B			class=basketweaving		
Grade	effort=Small	Medium	Large	effort=Small	Medium	Large
A	0	0.185	0.581	0.333	0.571	0.857
B	0.167	0.296	0.279	0.333	0.286	0.143
C	0.167	0.370	0.116	0.333	0.143	0
D	0.333	0.148	0.023	0	0	0
F	0.333	0	0	0	0	0

(b) What is the marginal probability distribution P(grade, effort)?

	Small	Medium	Large
A	**0.05**	**0.125**	**0.275**
B	**0.075**	**0.09**	**0.085**
C	**0.075**	**0.075**	**0.025**
D	**0.05**	**0.02**	**0.005**
F	**0.05**	**0**	**0**

(c) What is the marginal probability distribution P(effort)?

Small	Medium	Large
0.3	**0.31**	**0.39**

(d) What is P(grade=A | class)?
p(grade=A|class=165B)= 0.3
p(grade=A|class=basketweaving)= 0.6

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.

(2 points each)
(no need to write down the formula)

(a) Show the contingency table for this binary classification experiment. Label it clearly and fill out the table entries. (0.5 points for each of four entries. If the labels are not clear you lose points for corresponding entries)

		Actual Class		
		Face	**Not Face**	
Predicted Class	**Face**	425 (TP)	125 (FP)	550 (Est Pos)
	Not Face	75 (FN)	375 (TN)	450 (Est Neg)
		500 (P)	500(N)	1000 (Total)

We are building a face detection system. It implied the True Positives are when Face Images are classified as Face Class.
(Dont need the above justification)

(b) What is the false positive rate of the system in this experiment?
FPR = FP/N = 125/500 = 0.25

(c) What is the false negative rate?
FNR = FN/P = 75/500 = 0.15

(d) What is the error rate?
Error Rate = FP+FN/P+N = 125+75/500+500 = 0.2

(e) What is the precision?
Precision = TP/Est Pos = 425/550 =0.7727

(f) What is the accuracy?
Accuracy = TP+TN/P+N = 425+375/500+500 = 0.8
Or 1 - Error Rate = 1 -0.2 = 0.8 (this is also fine)

Problem #7 [12 points] Here is some data describing features in a machine learning problem (six instances in a three-dimensional feature space):

(a) What is the intrinsic dimensionality of the data?
The intrinsic dimensionality is 1 because

1) The data points lie along a line.
2) Dimensions F2 and F3 can be expressed through F1.

(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, ….)
The datapoints are not actually linearly separable, but if N = 1 then it will just be a point.

(c) Are the examples in this training set linearly separable
No, they are not, at least not in the one-dimensional, two-dimensional and three-dimensional spaces.