CM146: Introduction to Machine Learning	**Winter 2018**

Midterm

Feb. 13^{th}, 2018

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam booklet contains **four** problems.
- You have 90 minutes to earn a total of 100 points.
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**

Good Luck!

Name and ID: (2 Point)

Name		/2
Short Questions		/40
Perceptron		/20
Decision Tree		/18
Regression		/20
Total		/100

Short Questions [40 points]

1. [21 points] True/False Questions (Add 1 sentence to justify your answer if the answer is "False".)

   (a) When the hypothesis space is richer, over-fitting is more likely.

   (b) Nearest neighbors is more efficient at training time than logistic regression.

   (c) Perceptron algorithms can always stop after seeing γ^2/R^2 number of examples if the data is linearly separable, where γ is the size of the margin and R is the size of the largest instance.

   (d) Instead of maximizing a likelihood function, we can minimize the corresponding negative log-likelihood function.

   (e) If data is not linearly separable, decision tree can not reach training error zero.

   (f) If data is not linearly separable, logistic regression can not reach training error zero.

   (g) To predict the probability of an event, one would prefer a linear regression model trained with squared error to a classifier trained with logistic regression.

2. [9 points] You are a reviewer for the International Conference on Machine Learning, and you read papers with the following claims. Would you accept or reject each paper? Provide a one sentence justification if your answer is "reject".

   - **accept/reject**] "My model is better than yours. Look at the training error rates!"

   - **accept/reject** "My model is better than yours. After tuning the parameters on the test set, my model achieves lower test error rates!"

   - **accept/reject** "My model is better than yours. After tuning the parameters using 5-fold cross validation, my model achieves lower test error rates!"

3. [10 points] On the 2D dataset of Fig. 1, draw the decision boundaries learned by logistic regression and 1-NN (using two features x and y). Be sure to mark which regions are labeled positive or negative, and assume that ties are broken arbitrarily.

   i. Logistic regression
   ii. 1-NN

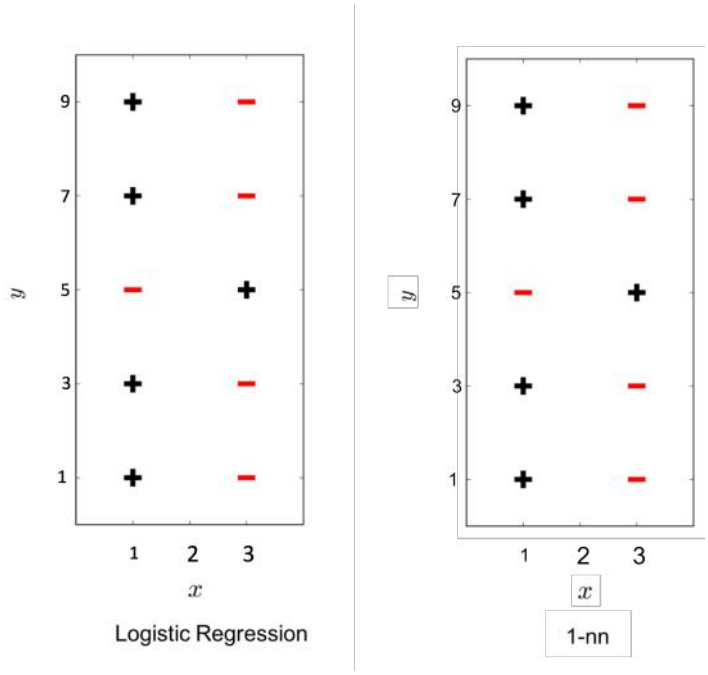

Figure 1: Example 2D dataset for question

Perceptron [20 points]

Recall that the Perceptron algorithm makes an updates when the model makes a mistake. Assume now our model makes prediction using the following formulation:

$$y = \begin{cases} 1 & \text{if} \quad w^T x \geq 1, \\ -1 & \text{if} \quad w^T x < 1. \end{cases} \tag{1}$$

1. [12 points] Finish the following Perceptron algorithm by choosing from the following options.

(a) $w^T x_i \geq 0$	(b) $y_i = 1$	(c) $w^T x \geq 1$ and $y_i = 1$	(d) $w^T x \geq 1$ and $y_i = -1$
(e) $w^T x_i < 0$	(f) $y_i = -1$	(g) $w^T x < 1$ and $y_i = 1$	(h) $w^T x < 1$ and $y_i = -1$
(i) x_i	(j) - x_i	(k) $w + x_i$	(l) $w - x_i$
(m) $y_i(w + x_i)$	(n) $-y_i(w + x_i)$	(o) $w^T x_i$	(p) $-w^T x_i$

Given a training set $D = \{x_i, y_i\}_{i=1}^m$

Initialize $w \leftarrow 0$.

For $(x_i, y_i) \in D$:

if _____

$w \leftarrow$ _____

if _____

$w \leftarrow$ _____

Return w

2. [4 points] Let w to be a two dimensional vector. Given the following dataset, can the function described in (1) separate the dataset?

Instance	1	2	3	4	5	6	7	8
Label y	+1	-1	+1	+1	+1	-1	-1	+1
Data (x_1, x_2)	(2, 0)	(2, 4)	(-1, 1)	(1, -1)	(-1, -1)	(4, 0)	(2, 2)	(0, 2)

Instance	1	2	3	4	5	6	7	8
Label y	+1	-1	+1	+1	+1	-1	-1	+1
Data (x_1, x_2)	(2, 0)	(2, 4)	(-1, 1)	(1, -1)	(-1, -1)	(4, 0)	(2, 2)	(0, 2)

3. [4 points] If your answer to the previous question is "no", please describe how to extend w and data points x into 3-dimensional vectors, such that the data can be separable. If your answer to the previous question is "yes", write down the w that can separate the data.

Decision Tree [18 points]

We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied.

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

For this problem, you can write your answers using $\log_2$, but it may be helpful to note that $\log_2 3 \approx 1.6$ and entropy $H(S) = -\sum_{v=1}^{K} P(S = v) \log_2 P(S = v)$. The information gain of an attribute A is $G(S, A) = H(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} H(S_v)$, where S_v is the subset of S for which A has value v.

1. [4 points] What is the entropy H(Passed)?

2. [4 points] What is the entropy G(Passed, GPA)?

3. [4 points] What is the entropy G(Passed, Studied)?

4. [6 points] Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations.

Linear Regression [20 points]

1. [6 points] Describe one application of linear regression. Please define clearly what are your input, output, and features.

2. [6 points] Given a dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^{M}$ in a two dimensional space. The objective function of linear regression with square loss is

$$J(w_1, w_2) = \frac{1}{2}\sum_{i=1}^{M}(y_i - (w_1 x_1^{(i)} - w_2 x_2^{(i)}))^2, \tag{2}$$

   where w_1 and w_2 are feature weight to be learned. Write down one optimization procedure that can learn w_1 and w_2 from data. Please be as explicit as possible.

3. [8 points] Prove that Eq. (2) has a global optimal solution. (Full points if the proof is mathematically correct. 4 points if you can describe the procedure for proving the claim.)