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.
   True.

   (You may get full/partial credits if your answer is "False" and provide proper explanation: For example: over-fitting also depends on many other factors. The conclusion cannot be made only based on the size of hypothesis space.)

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

   (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.
   False. Perceptron algorithm is guaranteed to stop after making R^2/γ^2 mistakes. (There are two mistakes in the statements. Pointing out any one of them gets the full credit.)

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

   (e) If data is not linearly separable, decision tree can not reach training error zero.
   False. Decision tree is a non-linear classifier and it can reach zero training error even if the data is not linearly separable.

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

   (You may get full/partial credits if your answer is "False" and provide proper explanation: A logistic regression with non-linear mapping of input vector can reach zero training error for data not linearly separable.

   (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.

   False. To predict the probability of an event, logistic regression is preferred.

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!"
     Reject. Low training error rates can lead overfitting.
   - **accept/reject** "My model is better than yours. After tuning the parameters on the test set, my model achieves lower test error rates!"
     Reject. The parameters on the test set should not be tuned to achieve better performance.
   - **accept/reject** "My model is better than yours. After tuning the parameters using 5-fold cross validation, my model achieves lower test error rates!"
     Accept.

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.

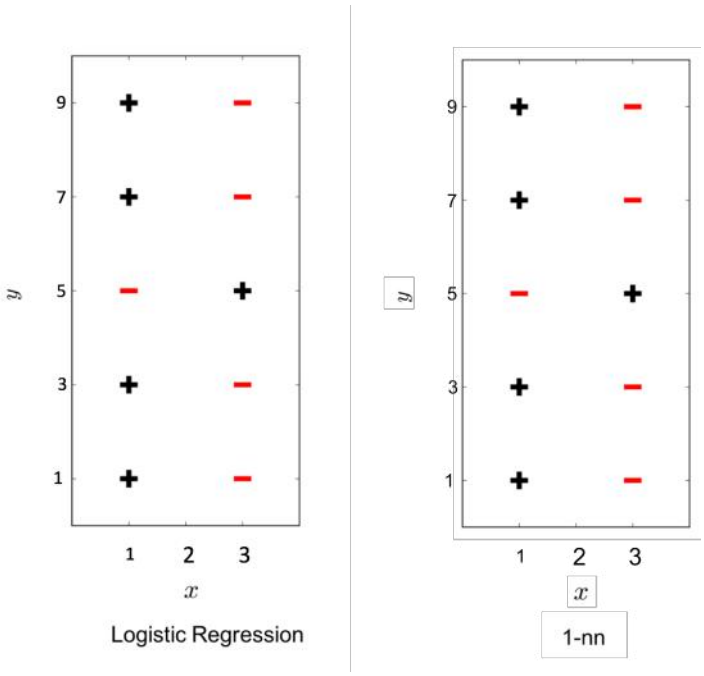
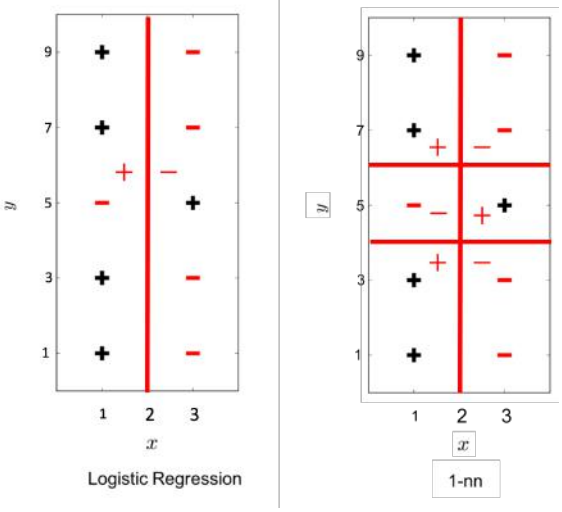


Figure 1: Example 2D dataset for question

answer: See the figure below. 3 points for showing the decision boundary of logistic regression is a line. 3 points for showing the decision boundary of 1-nn is non-linear.



i. Logistic regression

ii. 1-NN

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 __d__

   $w \leftarrow$ __l__

   if __g__

   $w \leftarrow$ __k (or m)__

   Return w

   Note: g, k, d, l or g, m, d, l are also correct.

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)

   No. Despite the data is linearly separable, without the bias term, the data cannot be separated by Eq. (1). (2 points for showing the data is linearly separable. Full credits give to students showing after augmenting data, Eq. (1) can separate the data.

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.

   We need to augment the data and the weight vector with one additional dimension for the bias term. There are several ways to extend the data. The following is one example.

Instance	1	2	3	4	5	6	7	8
Label y	+1	-1	+1	+1	+1	-1	-1	+1
Data (x_1, x_2, x_3)	(2, 0, 1)	(2, 4, 1)	(-1, 1, 1)	(1, -1, 1)	(-1, -1, 1)	(4, 0, 1)	(2, 2, 1)	(0, 2, 1)

   Rubrics:

   - Full points are given if the answer meets one of the following: a) The student describes how to extend x (like [x1,x2,1]) and how to extend w (like [w1,w2,w3]). b) The student provides a correct weight vector (e.g. w = [-1,-1,3]).
   - 2 points are given if the answer meets one of the following: a) The student describes dimension extension methods on either w or x (but missing another one). b) The student provides an incorrect 3-d weight vector (e.g. w = [1,1,-3]) with some explanation about how to extend w or x.
   - 0 points are given if the answer meets one of the following: a) The student provides a 2-dim weight vector b) The student provides an incorrect 3-d weight vector with no explanation.

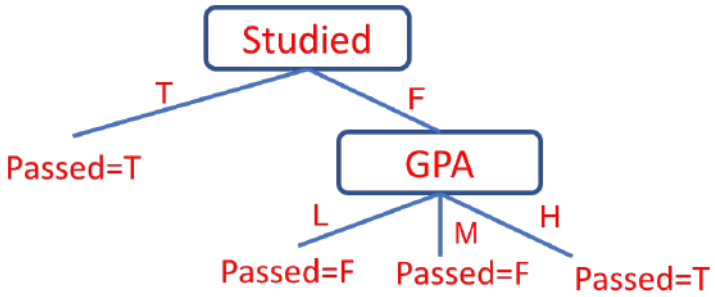
Decision Tree [18 points]

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)?
   $-\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6} = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \approx 0.92$ (or $\frac{14}{15}$) (you don't have to simplify it).
2. [4 points] What is the entropy G(Passed, GPA)?
   $H(Passed) - \frac{1}{3}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) - \frac{1}{3}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}) - \frac{1}{3} \times 0 = 0.92 - 0.66 = 0.26$ (or $\frac{4}{15}$)(Full credits also give to answer without simplification.)
3. [4 points] What is the entropy G(Passed, Studied)?
   $H(Passed) - \frac{1}{2}(-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}) - \frac{1}{2} \times 0 = 0.92 - 0.46 = 0.46$ (or $\frac{7}{15}$, or $\frac{1}{2}$H(Passed)) (Full credits also give to answer without simplification.)
4. [6 points] Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations.
   The following one is the one learned by the ID3 algorithm.



Linear Regression [20 points]

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

   Any application. 2 points for input, output, and features, respectively.

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.

   There is a typo in the Eq. (2); however it doesn't affect the model. Showing answer to solve Eq. (2) or

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

   both get full credits. We show the answer of the latter case.

   Closed-form solution: 6 points for correct formulation. 4 points for stating there is a closed-form solution.

   SGD/GD: 6 points for correct procedure and gradient. 4 points for correct procedure or correct gradient.

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.)

To prove Eq. (2) has a global optimum, we have to show the function is convex (2 points). To prove the function is convex, we need to demonstrate the Hessian matrix (or the second derivatives) is positive semi-definite (2 points).

The Hessian of Eq. (2) is

$$H = \begin{bmatrix} \sum (x_1^{(i)})^2 & \sum x_1^{(i)} x_2^{(i)} \\ \sum x_1^{(i)} x_2^{(i)} & \sum (x_2^{(i)})^2 \end{bmatrix} \tag{3}$$

(2 points)

To show H is positive semi-definite, we have to prove for every vector $z \neq 0$, $z^T H z \geq 0$. This can be done by the following equations:

$$\begin{aligned} z^T H z &= \sum (x_1^{(i)})^2 z_1^2 + 2 \sum x_1^{(i)} x_2^{(i)} z_1 z_2 z_1 z_2 + \sum (x_1^{(i)})^2 z_2^2 \\ &= \sum (z_1 x_1^{(i)} + z_2 x_2^{(i)})^2 \\ &\geq 0 \end{aligned} \tag{4}$$

(2 points)