CSCI 567, Spring 2019
Prof. Victor Adamchik

Theory Assignment 1

Due: 11:59 pm, Feb 1st, 2019

Instructions

Submission: Assignment submission will be via courses.uscden.net. By the submission date, there will be a folder named 'Theory Assignment 1' set up in which you can submit your files. Please be sure to follow all directions outlined here.

You can submit multiple times, but only *the last submission* counts. That means if you finish some problems and want to submit something first and update later when you finish, that's fine. In fact you are encouraged to do this: that way, if you forget to finish the homework on time or something happens (remember Murphy's Law), you still get credit for whatever you have turned in.

Problem sets must be typewritten or neatly handwritten when submitted. In both cases, your submission must be a single PDF. It is strongly recommended that you typeset with LaTeX. There are many free integrated LaTeX editors that are convenient to use (e.g Overleaf, ShareLaTeX). Choose the one(s) you like the most. This tutorial Getting to Grips with LaTeX is a good start if you do not know how to use LaTeX yet.

Please also follow the rules below:

- The file should be named as `firstname_lastname_USCID.pdf` e.g., `Don_Quijote_de_la_Mancha_8675309045.pdf`).
- Do not have any spaces in your file name when uploading it.
- Please include your name and USCID in the header of your report as well.

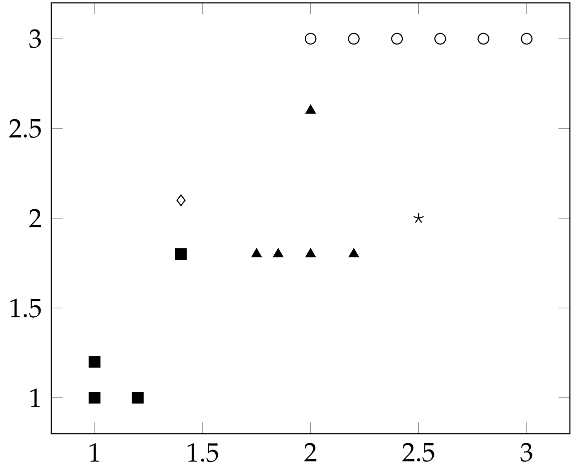
Collaboration: You may discuss with your classmates. However, you need to write your own solutions and submit separately. Also in your report, you need to list with whom you have discussed for each problem. Please consult the syllabus for what is and is not acceptable collaboration. Review the rules on academic conduct in the syllabus: a single instance of plagiarism can adversely affect you significantly more than you could stand to gain.

Notes on notation:

- Unless stated otherwise, scalars are denoted by small letter in normal font, vectors are denoted by small letters in bold font and matrices are denoted by capital letters in bold font.
- $\|.\|$ means L2-norm unless specified otherwise i.e. $\|.\| = \|.\|_2$

Problem 1 Problem 1 **(10 points)**

Using the following data to answer sub-problems I through VII , where squares, triangles, and open circles are three different classes of data in the training set and the diamond (◇) and star (*) are test points.



I Suppose I use all the training data as the full validation set (disregarding the two specially-marked test items) to classify the data using a KNN with k=1. How many of the 15 points will be*not* correctly classified?

(a) 0
(b) 2
(c) 5
(d) 10
(e) 15

Ans: 0. Trivially, all of them will be classified correctly. **(2 points)**

II For the KNN classifier with k=1, how many training data points will be misclassified with leave-one-out heuristic?

(a) 0
(b) 1
(c) 2
(d) 6
(e) 15

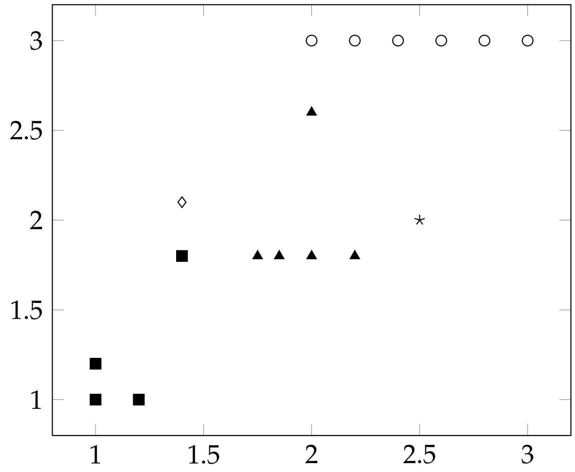
Ans: 2. The top triangle, and the upper-right box. **(2 points)**

III What is the smallest value of *k* to always classify the diamond as class square?

(a) 1
(b) 3
(c) 5
(d) 7
(e) cannot be done

Ans: 1. **(2 points)**

The following is the same diagram from the previous page.



IV What is the smallest value of k to classify the star as open circle?

(a) 3
(b) 6
(c) 10
(d) 12
(e) cannot be done

Ans: 12 **(1 points)**

V What label will we predict for diamond with $k = 1$?

(a) Square
(b) Triangle
(c) Circle
(d) Cannot be determined from data available.

Ans: Square **(1 points)**

VI What label will we predict for star with $k = 3$?

(a) Square
(b) Triangle
(c) Circle
(d) Cannot be determined from data available.

Ans: Triangle **(1 points)**

VII What label will we predict for star with $k = 5$?

(a) Square
(b) Triangle
(c) Circle
(d) Cannot be determined from data available.

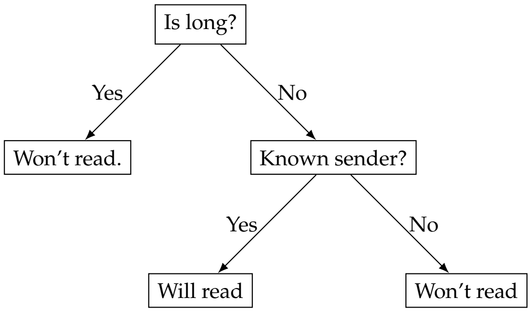
Ans: Triangle **(1 points)**

Problem 2 Decision Tree (10 points)

Suppose we would like to build a decision tree classifier to predict "whether the professor will read the email", using the following training dataset where the first 5 columns represent 5 binary features (whether the professor knows the sender, whether the email is too long, and whether it is about certain topics), and the last column is the label.

Known sender?	Is Long?	About research?	About grade?	About lottery?	Read?
no	no	yes	yes	no	no
yes	yes	no	yes	no	no
no	yes	yes	yes	yes	no
yes	yes	yes	yes	no	no
no	yes	no	no	no	no
yes	no	yes	yes	yes	yes
no	no	yes	no	no	yes
yes	no	no	no	no	yes
yes	no	yes	yes	no	yes
yes	yes	yes	yes	yes	no

Below is an example of decision tree for this task:



Define the **Information Gain** of a node n with children $\text{Children}(n)$ as

$$\text{Gain}(n) = \text{Entropy}(S_n) - \sum_{m \in \text{Children}(n)} \frac{|S_m|}{|S_n|} \text{Entropy}(S_m)$$

where S_n and S_m are the subsets of training examples that belong to the node n and one of its child node m respectively.

Compute the information gain for the node 'Known sender?' and the root ('Is long?'), respectively. Express your answer in terms of "log", that is, you do not need to calculate the value of logarithm (and therefore the base of the logarithm also does not matter).

Ans:

- 'Known sender?'
  The entropy is

$$\text{Entropy}(S) = -\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5} \qquad \text{(1 point)}$$

The entropy for the left child and right child are respectively

$$\text{Entropy}(S_1) = -\frac{3}{3}\log\frac{3}{3} - \frac{0}{3}\log\frac{0}{3} = 0 \qquad \text{(1 point)}$$

and

$$\text{Entropy}(S_2) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} \qquad \text{(1 point)}$$

The information gain is therefore

$$\text{Entropy}(S) - \frac{2}{5}\text{Entropy}(S_2) \qquad \text{(2 point)}$$

- 'Is long?'
  The entropy is

  $$\text{Entropy}(S) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} \qquad \text{(1 point)}$$

  The entropy for the left child and right child are respectively

  $$\text{Entropy}(S_1) = -\frac{0}{5}\log\frac{0}{5} - \frac{5}{5}\log\frac{5}{5} = 0 \qquad \text{(1 point)}$$

  and

  $$\text{Entropy}(S) = -\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5} \qquad \text{(1 point)}$$

  The information gain is therefore

  $$\text{Entropy}(S) - \frac{1}{2}\text{Entropy}(S_2) \qquad \text{(2 point)}$$

Note: incorrect information gain due to mistakes from previous calculations of entropy can still get the 1 point if the formula is applied correctly.

Problem 3 Linear Regression (10 points)

We consider the linear regression problem with two unknown variables $\boldsymbol{w}$ and b on a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, where we minimize the Residual Sum of Squares (RSS) over $\mathcal{D}$:

$$RSS(\boldsymbol{w}, b) = \sum_{n=1}^{N} [y_n - \boldsymbol{w}^T x_n - b]^2. \tag{1}$$

Let $\boldsymbol{w}_{\text{LMS}}$ and b_{LMS} be the least mean square (LMS) solution of the above optimization problem. We make the prediction for each point using $\boldsymbol{w}_{\text{LMS}}$ and b_{LMS},

$$\hat{y}_n = \boldsymbol{w}_{\text{LMS}}^T x_n + b_{\text{LMS}}. \tag{2}$$

Consider solving a new linear regression problem on the relabeled dataset $\mathcal{D}' = \{(x_n, y_n - \hat{y}_n)\}_{n=1}^N$. Let $\boldsymbol{v}_{\text{LMS}}$ and c_{LMS} be the LMS solution of the new linear regression that minimizes $RSS(\boldsymbol{v}, c)$ over $\mathcal{D}'$. Show that the solution is $\boldsymbol{v}_{\text{LMS}}$ = **0** and c_{LMS} = 0.

Ans:

$$\begin{aligned} RSS(\boldsymbol{v}, c) &= \sum_{n=1}^{N} [y_n - \hat{y}_n - \boldsymbol{v}^T x_n - c]^2 \textbf{ (2 points)} \\ &= \sum_{n=1}^{N} [y_n - \boldsymbol{w}_{\text{LMS}}^T x_n - b_{\text{LMS}} - \boldsymbol{v}^T x_n - c]^2 \textbf{ (1 points)} \\ &= \sum_{n=1}^{N} [y_n - (\boldsymbol{w}_{\text{LMS}} + \boldsymbol{v})^T x_n - (b_{\text{LMS}} + c)]^2 \textbf{ (2 points)} \end{aligned}$$

Consider $\boldsymbol{w}_{\text{LMS}} + \boldsymbol{v}$ and $b_{\text{LMS}} + c$ as the parameters of the original linear regression problem. Then, minimizing $RSS(\boldsymbol{v}, c)$ gives

$$\boldsymbol{w}_{\text{LMS}} + \boldsymbol{v}_{\text{LMS}} = \boldsymbol{w}_{\text{LMS}} \text{ and } b_{\text{LMS}} + c_{\text{LMS}} = b_{\text{LMS}} \textbf{ (3 points)}$$

It follows that $\boldsymbol{v}_{\text{LMS}}$ = **0** and c_{LMS} = 0. **(2 points)**

Problem 4 Naïve Bayes I (10 points)

Suppose you are given the following set of data with three Boolean input variables a, b, and c, and a single Boolean output variable K.

a	b	c	K
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

I According to the naive Bayes classifier, what is a value of $P(K = 1|a = 1, b = 1)$?

Ans:

$$\begin{aligned}
P(a = 1, b = 1) &= P(a = 1, b = 1|K = 0)P(K = 0) + P(a = 1, b = 1|K = 1)P(K = 1) \\
&= (2/4)(2/4)(1/2) + (2/4)(1/4)(1/2) \\
&= 1/8 + 1/16 \\
&= 3/16 \textbf{ (2 points)} \\
P(K = 1|a = 1, b = 1) &= \frac{P(a = 1|K = 1)P(b = 1|K = 1)P(K = 1)}{P(a = 1, b = 1)} \textbf{ (1 points)} \\
&= (2/4)(1/4)(1/2)/(3/16) = 1/3 \textbf{ (3 points)}
\end{aligned}$$

II In an unrelated example, there are three variables denoted as X, Y, and Z. If you are told that

$$P(Z|X) = 0.7$$
$$P(Z|Y) = 0.4$$

Can you compute $P(Z|X, Y)$? Compute the value of $P(Z|X, Y)$ if so or write 'not enough info' otherwise. Ans: Not enough info. **(4 points)**

Problem 5 Naïve Bayes II (10 points)

Let X, Y, and Z be random variables taking values in $\{0, 1\}$. The following table lists the probability of each possible assignment of 0 and 1 to the variables X, Y, and Z:

	Z=0		Z=1	
	X=0	X=1	X=0	X=1
Y=0	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{4}{15}$	$\frac{2}{15}$
Y=1	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{8}{45}$	$\frac{4}{45}$

I Is X independent of Y? Why or why not?

Ans:

No.

$P(X = 0) = 11/18, P(Y = 0) = 8/15, P(X = 0|Y = 0) = 5/8.$ **(1 points)**

Since $P(X = 0)$ does not equal $P(X = 0|Y = 0)$, X is not independent of Y. **(1 points)**

II Is X conditionally independent of Y given Z? Why or why not?

Ans:

For all pairs y, z 0, 1, we need to check that $P(X = 0|Y = y, Z = z) = P(X = 0|Z = z)$. That the other probabilities are equal follows from the law of total probability. First we have

$$P(X = 0|Y = 0, Z = 0) = \frac{1/15}{1/15 + 1/15} = 1/2 \ \textbf{(0.5 points)}$$
$$P(X = 0|Y = 1, Z = 0) = \frac{1/10}{1/10 + 1/10} = 1/2 \ \textbf{(0.5 points)}$$
$$P(X = 0|Y = 0, Z = 1) = \frac{4/15}{4/15 + 2/15} = 2/3 \ \textbf{(0.5 points)}$$
$$P(X = 0|Y = 1, Z = 1) = \frac{8/45}{8/45 + 4/45} = 2/3 \ \textbf{(0.5 points)}$$

Then we have

$$P(X = 0|Z = 0) = \frac{1/15 + 1/10}{1/15 + 1/15 + 1/10 + 1/10} = 1/2 \ \textbf{(2 points)}$$
$$P(X = 0|Z = 1) = \frac{4/15 + 8/45}{4/15 + 2/15 + 8/45 + 4/45} = 2/3 \ \textbf{(2 points)}$$

This shows that X is independent of Y given Z.

III Calculate $P(X = 0|X + Y > 0)$.

Ans:

$$P(X = 0|X + Y > 0) = \frac{1/10 + 8/45}{1/15 + 1/10 + 1/10 + 2/15 + 4/45 + 8/45} = 5/12.$$

(2 points)

Problem 6 Naïve Bayes III (10 points)

Consider a Naïve Bayes model with a class variable $Y \in \{0, 1\}$ and input feature variable $X \in \mathbb{R}^d$. By the assumption of Naïve Bayes, the value of a particular dimension of X is independent of the value of any other dimensions, given the class variable Y.

I Denote the conditional probability for each dimension of X given Y as $P(X_i|Y)$, $i = 1, \cdots, d$, and the class prior of Y as $P(Y)$. Write down the joint probability $P(X, Y)$ and the conditional probability $P(X|Y)$.

Ans:
$P(X|Y) = \prod_{i=1}^d P(X_i|Y)$ **(1 points)**
$P(X, Y) = \prod_{i=1}^d P(X_i|Y)P(Y)$ **(1 points)**

II Now the conditional probability for each dimension of X given Y as $P(X_i|Y)$ is defined as:

$$P(X_i = x_i|Y = y) = h_i(x_i)exp(\eta_{iy} f_i(x_i))$$

and the prior of Y is a Bernoulli distribution parameterized by $\theta \in [0, 1]$:$P(Y = 1) = \theta$. Please write down the class probability $P(Y = 1|X = x)$ and $P(Y = 0|X = x)$, and simply them into the form like sigmoid function:

$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

Ans:

$$\begin{aligned}
P(Y = y|X = x) &\propto P(X = x, Y = y) \\
&= \prod_{i=1}^d P(X_i|Y)P(Y) \\
&= \prod_{i=1}^d h_i(x_i)exp(\eta_{iy} f_i(x_i))\theta^y(1-\theta)^{1-y} \\
&= (\prod_{i=1}^d h_i(x_i))exp(\sum_{i=1}^d(\eta_{iy} f_i(x_i)))\theta^y(1-\theta)^{1-y} \textbf{ (2 points)}
\end{aligned}$$

Since $Y \in \{0, 1\}$, $P(Y = 1|X = x) + P(Y = 0|X = x) = 1$. We compute $P(Y = 1|X = x)$ as follows:

$$\begin{aligned}
P(Y = 1|X = x) &= \frac{(\prod_{i=1}^d h_i(x_i))exp(\sum_{i=1}^d(\eta_{i1} f_i(x_i)))\theta}{(\prod_{i=1}^d h_i(x_i))exp(\sum_{i=1}^d(\eta_{i1} f_i(x_i)))\theta + (\prod_{i=1}^d h_i(x_i))exp(\sum_{i=1}^d(\eta_{i0} f_i(x_i)))(1-\theta)} \\
&= \frac{1}{1 + \frac{1-\theta}{\theta} exp(\sum_{i=1}^d(\eta_{i0} - \eta_{i1}) f_i(x_i))} \textbf{ (3 points)} \\
P(Y = 0|X = x) &= \frac{1}{1 + \frac{\theta}{1-\theta} exp(\sum_{i=1}^d(\eta_{i1} - \eta_{i0}) f_i(x_i))} \textbf{ (3 points)}
\end{aligned}$$

Note that if the probability is not simplified, deduct 0.5 point.