CSCI 567, Fall 2019
Prof. Victor Adamchik

Theory Assignment 1

Due: 11:59 pm, Sept 20th, 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 Nearest Neighbor Classification (8 points)

In the class, we use the **Euclidean distance** as the distance metric for the nearest neighbor classification. Given data $\boldsymbol{x} \in \mathbb{R}^D$, the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as following:

$$E(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 = \sum_{d=1}^{D} (x_{id} - x_{jd})^2 \tag{1}$$

In some applications such as information retrieval and neural language processing, the **cosine distance** is widely applied. It is defined as:

$$C(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_j\|_2} = 1 - \frac{\sum_{d=1}^{D} (x_{id} \cdot x_{jd})}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_j\|_2}, \tag{2}$$

where the norm of $\boldsymbol{x}$ is defined as

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{d=1}^{D} x_d^2}. \tag{3}$$

Now you are asked to prove that for any $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ normalized to the unit norm, *i.e.* $\|\boldsymbol{x}_i\|_2 = \|\boldsymbol{x}_j\|_2 = 1$, changing the distance metric from the Euclidean distance to the cosine distance will NOT affect the nearest neighbor classification results. Specifically, for any $\boldsymbol{x}_i, \boldsymbol{x}_j$ and $\boldsymbol{x}_o$, show that, if $C(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq C(\boldsymbol{x}_i, \boldsymbol{x}_o)$, then $E(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq E(\boldsymbol{x}_i, \boldsymbol{x}_o)$, where $\|\boldsymbol{x}_i\|_2 = \|\boldsymbol{x}_j\|_2 = \|\boldsymbol{x}_o\|_2 = 1$.

Ans:

For any $\boldsymbol{x}_i$, $\boldsymbol{x}_j$, $\boldsymbol{x}_o$ that are normalized to the unit norm,

$$\begin{aligned} E(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \\ &= (\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \\ &= \boldsymbol{x}_i^T \boldsymbol{x}_i - 2\boldsymbol{x}_i^T \boldsymbol{x}_j + \boldsymbol{x}_j^T \boldsymbol{x}_j \\ &= \|\boldsymbol{x}_i\|_2^2 - 2\boldsymbol{x}_i^T \boldsymbol{x}_j + \|\boldsymbol{x}_j\|_2^2 \\ &= 2 - 2\boldsymbol{x}_i^T \boldsymbol{x}_j \end{aligned}$$

(3 points)

$$\begin{aligned} C(\boldsymbol{x}_i, \boldsymbol{x}_j) &= 1 - \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_j\|_2} \\ &= 1 - \boldsymbol{x}_i^T \boldsymbol{x}_j \\ &= E(\boldsymbol{x}_i, \boldsymbol{x}_j)/2 \end{aligned}$$

(3 points)

Thus,

$$\begin{aligned} E(\boldsymbol{x}_i, \boldsymbol{x}_j) &= 2C(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &\leq 2C(\boldsymbol{x}_i, \boldsymbol{x}_o) \\ &= E(\boldsymbol{x}_i, \boldsymbol{x}_o) \end{aligned}$$

(2 points)

Problem 2 Nearest Neighbor Classification and Decision Trees (4 points)

Assume we have a dataset, each data $\boldsymbol{x}$ is a 100 dimensional binary vector, i.e. $\boldsymbol{x} \in \{0,1\}^{100}$, and each $\boldsymbol{x}$ is assigned a label $\in \{0,1\}$.

I Can we have a decision tree to classify the dataset with zero classification error w.r.t. their labels?

II Can we specify a 1-NN over the dataset to result in exactly the same classification as our decision tree?

For both questions explain why or why not with examples. You can assume that all data points are distinct, i.e. $\forall \boldsymbol{x}_i, \boldsymbol{x}_j$ in the dataset, $\boldsymbol{x}_i \neq \boldsymbol{x}_j$. (Hint: if your model works for binary label then it will also work for any kind of labels)
Ans:

- Yes. A simple solution would be to generate all 2^{100} possible strings and use the decision tree to classify each of them. **(2 points)**
- Yes. Then let the 1-NN be the entire collection of these length 100 binary strings. **(2 points)**

Problem 3 Nearest Neighbor Classification and Decision Trees (6 points)

Assume we have a decision tree in Fig. 1 which classifies $\boldsymbol{x} \in \mathbb{C}^2$, and $\mathbb{C} = \mathbb{Z} \setminus \{A, B\}$. In other words, A and B are integers and each dimension of $\boldsymbol{x}$ is an integer excluding A and B. Can this decision tree be implemented as a 1-NN? If so, explicitly write down each of the values you use for the 1-NN (you should use the minimal number possible). If not, either explain why or provide a counter example.

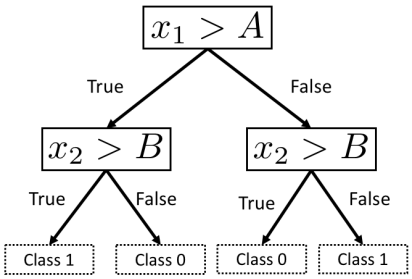

Figure 1: A decision tree example.

Ans: Yes. The following 4 points are enough to specify a 1-NN that has the exact same outcome as the decision tree: **(2 points)**

- $\{A+1, B+1\}$: class 1 **(1 points)**
- $\{A+1, B-1\}$: class 0 **(1 points)**
- $\{A-1, B+1\}$: class 0 **(1 points)**
- $\{A-1, B-1\}$: class 1 **(1 points)**

Problem 4 Decision Tree (8 points)

In this problem, you are given four 2-dimensional data points as shown in Table 1:

x_1	x_2	label y
0	0	0
0	1	1
1	0	1
1	1	0

Table 1: Four 2-dimensional data points and their labels.

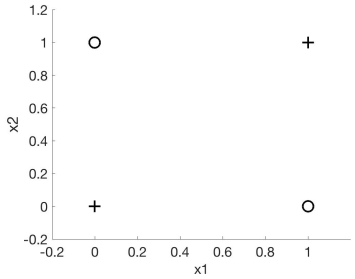

Figure 2: The plus sign means label $y = 0$ and the circle means have $y = 1$.

4.1 Fig. 3 is a decision tree of the given data with zero training error.

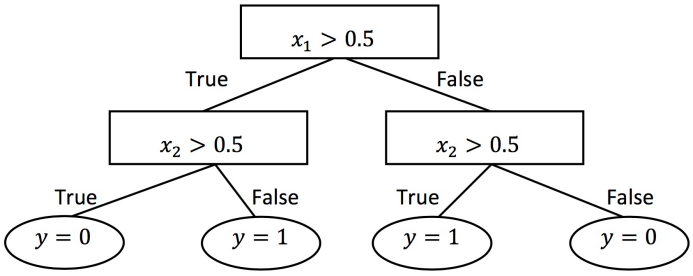

Figure 3: The decision tree with zero training error.

Suppose now you have two test data points:

x_1	x_2	label y
0.2	0.2	0
0.2	0.8	1

What would be your test error based on decision tree in Fig. 3? (Define the test error as the fraction of mis-classifications made on the testing set.) **(2 points)**

Ans: Both data will be classified correctly. So the test error is 0.

4.2 Now consider a new decision tree in Fig. 4. Note that the depth of the new decision tree is 1, and it does not have zero training error for the given data anymore.

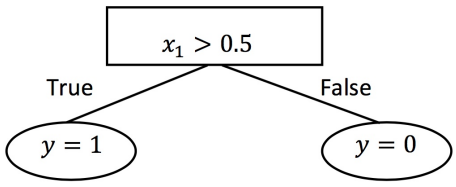

Figure 4: The decision tree with depth $= 1$.

Given the two test data points from Question 4.1, what would be the test error using the new decision tree? **(2 points)**
Ans: The second data will be mis-classified.
So the test error is 0.5.

4.3 Is the decision tree in Fig. 4 a linear or non-linear classifier in terms of (x_1, x_2) (Yes/No)? Can you classify the given data in Table 1 and get zero classification error by drawing a depth-1 decision tree similar to Fig. 4 (Yes/No)? Note that the decision rule should be based on a single variable (x_1 or x_2) in the rectangle (e.g., $x_1 > 1$ or $x_2 < 2$). **(2 points)**
Ans:

- The decision tree in Fig. 4 is a linear classifier.
- Not linearly separable.

4.4 If you can put linear combination of variables in the rectangle (e.g.: $ax_1 + bx_2 \geq c$ or $ax_1 + bx_2 < c$, where a, b, c should be numbers), can you classify the given data in Table 1 and get zero classification error by drawing a depth-1 decision tree similar to Fig. 4 (Yes/No)? Please also **briefly** justify your answer. **(2 points)**
Ans: No, because the data is not linearly separable.

Problem 5 Decision Trees (12 points)

Consider a binary dataset with 400 examples, where half of them belongs to class A and the other half belongs to class B.

Next consider two decision stumps (i.e. trees with depth 1) $\mathcal{T}_1$ and $\mathcal{T}_2$, each with two children. For $\mathcal{T}_1$, its left child has 150 examples in class A and 50 examples in class B; for $\mathcal{T}_2$, its left child has 0 example in class A and 100 examples in class B. (You should infer what are in the right child.)

5.1 For each leaf of $\mathcal{T}_1$ and $\mathcal{T}_2$, compute the corresponding classification error, entropy (base e) and Gini impurity, rounding off your answer to 2 decimal places. (Note: the value/prediction of each leaf is the majority class among all examples that belong to that leaf.) **(6 points)**

Ans: Classification error:

$$\epsilon_{1,L} = \frac{50}{150+50} = 0.25$$
$$\epsilon_{1,R} = \frac{50}{50+150} = 0.25$$
$$\epsilon_{2,L} = \frac{0}{0+100} = 0$$
$$\epsilon_{2,R} = \frac{100}{200+100} \approx 0.33$$

(2 points)

Entropy:

$$H_{1,L} = -\frac{150}{150+50}\ln(\frac{150}{150+50}) - \frac{50}{150+50}\ln(\frac{50}{150+50}) \approx 0.56$$
$$H_{1,R} = -\frac{50}{150+50}\ln(\frac{50}{150+50}) - \frac{150}{150+50}\ln(\frac{150}{150+50}) \approx 0.56$$
$$H_{2,L} = -\frac{0}{0+100}\ln(\frac{0}{0+100}) - \frac{100}{0+100}\ln(\frac{100}{0+100}) = 0$$
$$H_{2,R} = -\frac{200}{200+100}\ln(\frac{200}{200+100}) - \frac{100}{100+200}\ln(\frac{200}{200+100}) \approx 0.64$$

(2 points)

Gini impurity:

$$G_{1,L} = 1 - (\frac{150}{150+50})^2 - (\frac{50}{150+50})^2 = 0.375 \approx 0.38$$
$$G_{1,R} = 1 - (\frac{50}{150+50})^2 - (\frac{150}{150+50})^2 = 0.375 \approx 0.38$$
$$G_{2,L} = 1 - (\frac{0}{0+100})^2 - (\frac{100}{0+100})^2 = 0$$
$$G_{2,R} = 1 - (\frac{200}{200+100})^2 - (\frac{100}{200+100})^2 \approx 0.44$$

(2 points)

5.2 Compare the quality of $\mathcal{T}_1$ and $\mathcal{T}_2$ (that is, the two different splits of the root) based on classification error, conditional entropy (base e), and weighted Gini impurity respectively, rounding off your answer to 2 decimal places. **(6 points)**

Ans: Let $p_1 = \frac{150+50}{400} = 0.5$ be the fraction of examples that belong to left leaf of $\mathcal{T}_1$, and $p_2 = \frac{0+100}{400} = 0.25$ be the fraction of examples that belong to left leaf of $\mathcal{T}_2$. Then the total classification error for $\mathcal{T}_1$ and $\mathcal{T}_2$ are respectively:

$$\epsilon_1 = p_1\epsilon_{1,L} + (1-p_1)\epsilon_{1,R} = 0.25$$
$$\epsilon_2 = p_2\epsilon_{2,L} + (1-p_2)\epsilon_{2,R} = 0.25$$

(2 points)

So they are as good in terms of classification error.

The conditional entropy for $\mathcal{T}_1$ and $\mathcal{T}_2$ are respectively:

$$\epsilon_1 = p_1 H_{1,L} + (1-p_1)H_{1,R} \approx 0.56$$
$$\epsilon_2 = p_2 H_{2,L} + (1-p_2)H_{2,R} = 0.48$$

(2 points)

So $\mathcal{T}_2$ is better in terms of conditional entropy.

The weighted Gini impurity for $\mathcal{T}_1$ and $\mathcal{T}_2$ are respectively:

$$\epsilon_1 = p_1 G_{1,L} + (1-p_1)G_{1,R} \approx 0.38$$
$$\epsilon_2 = p_2 G_{2,L} + (1-p_2)G_{2,R} \approx 0.33$$

(2 points)

So $\mathcal{T}_2$ is also better in terms of weighted Gini impurity.

Problem 6 Naive Bayes (12 points)

In this problem, we will try to predict whether it is suitable for playing tennis or not based on the weather condition, the emotion and the amount of homework, using Naive Bayes Classifier. You can think 'PlayTennis' is a label and 'PlayTennis $=$ *Yes*' means it is suitable for playing tennis. We assume the probability $P(\text{Weather, Emotion, Homework}|\text{PlayTennis})$ can be factorized into the product form such that

$$P(\text{Weather, Emotion, Homework}|\text{PlayTennis}) =$$
$$P(\text{Weather}|\text{PlayTennis}) \times P(\text{Emotion}|\text{PlayTennis}) \times P(\text{Homework}|\text{PlayTennis}).$$

The training data is as following. Each data point has three attributes (Weather, Emotion, Homework), where Weather $\in \{Sunny, Cloudy\}$, Emotion $\in \{Happy, Normal, Unhappy\}$, Homework $\in \{Much, Little\}$.

Weather	Emotion	Homework	PlayTennis
Sunny	Happy	Little	Yes
Sunny	Normal	Little	Yes
Cloudy	Happy	Much	Yes
Cloudy	Unhappy	Little	Yes
Sunny	Unhappy	Little	No
Cloudy	Normal	Much	No

6.1 What are the probabilities of $P(\text{PlayTennis} = Yes)$ and $P(\text{PlayTennis} = No)$? Each of your answer should be an irreducible fraction. **(2 points)**

Ans: $P(\text{PlayTennis} = Yes) = \frac{2}{3}$ and $P(\text{PlayTennis} = No) = \frac{1}{3}$

6.2 Write down the following conditional probabilities. Each of your answer should be an irreducible fraction. **(6 points)**

- $P(\text{Weather} = Sunny|\text{PlayTennis} = Yes) =?$
- $P(\text{Emotion} = Normal|\text{PlayTennis} = Yes) =?$
- $P(\text{Homework} = Much|\text{PlayTennis} = Yes) =?$

Ans:

- $P(\text{Weather} = Sunny|\text{PlayTennis} = Yes) = \frac{1}{2}$ **(2 points)**
- $P(\text{Emotion} = Normal|\text{PlayTennis} = Yes) = \frac{1}{4}$ **(2 points)**
- $P(\text{Homework} = Much|\text{PlayTennis} = Yes) = \frac{1}{4}$ **(2 points)**

6.3 Given the new data instance $\boldsymbol{x} = (\text{Wheather} = \textit{Sunny}, \text{Emotion} = \textit{Normal}, \text{Homework} = \textit{Much})$, which of the following has larger value: $P(\text{PlayTennis} = \textit{Yes}|\boldsymbol{x})$ or $P(\text{PlayTennis} = \textit{No}|\boldsymbol{x})$? Each of your answer should be an irreducible fraction. **(4 points)**

Ans: $P(\text{PlayTennis} = \textit{No}|\boldsymbol{x})$ is larger.

Denote PlayTennis $= \textit{Yes}$ as Y, and PlayTennis $= \textit{No}$ as N.

$$\begin{aligned} P(Y|\boldsymbol{x}) &= \frac{P(\boldsymbol{x}|Y)P(Y)}{P(\boldsymbol{x})} \\ &= \frac{P(\boldsymbol{x}|Y)P(Y)}{P(\boldsymbol{x}|Y)P(Y) + P(\boldsymbol{x}|N)P(N)} \\ &= \frac{\frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} \times \frac{2}{3}}{\frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} \times \frac{2}{3} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3}} \\ &= \frac{1}{3} \end{aligned}$$

$$\begin{aligned} P(N|\boldsymbol{x}) &= \frac{P(\boldsymbol{x}|N)P(N)}{P(\boldsymbol{x})} \\ &= \frac{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{4} \times \frac{1}{4} \times \frac{2}{3} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3}} \\ &= \frac{2}{3} \end{aligned}$$

So $P(\text{PlayTennis} = \textit{No}|\boldsymbol{x})$ is larger.