CSCI 567, Summer'20
Victor Adamchik

Problem Set #1

Due: July 8, 2020 11:59 pm

Instructions

Submission: Assignment submission will be via `courses.uscden.net`. By the submission date, there will be a folder named `Written 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 online LaTeX editors that are convenient to use (e.g Overleaf). You can also use offline editor such as TeXShop.

Please follow the rules below while submitting:

- The file should be named as Firstname_Lastname_USCID.pdf e.g., Jeff_Dean_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 written 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.

Note 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.

Problem 1 Nearest Neighbor Classification 1 (15 points)

We mentioned that the Euclidean/L2 distance is often used as the *default* distance for nearest neighbor classification. It is defined as

$$E(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 = \sum_{d=1}^{D} (x_d - x'_d)^2$$

In some applications such as information retrieval, the cosine distance is widely used too. It is defined as

$$C(\boldsymbol{x}, \boldsymbol{x}') = 1 - \frac{\boldsymbol{x}^T \boldsymbol{x}'}{\|\boldsymbol{x}\|_2 \|\boldsymbol{x}'\|_2} = 1 - \frac{\sum_{d=1}^{D} (x_d \cdot x'_d)}{\|\boldsymbol{x}\|_2 \|\boldsymbol{x}'\|_2},$$

where the L2 norm of $\boldsymbol{x}$ is defined as

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

Show that, if data is normalized with unit L2 norm, that is, $\|\boldsymbol{x}\| = 1$ for all $\boldsymbol{x}$ in the training and test sets, changing the distance function from the Euclidean distance to the cosine distance will NOT affect the nearest neighbor classification results.

Problem 2 Nearest Neighbor Classification 2 (10 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 $\in \{0,1\}$.

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

2.2 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)

Problem 3 Decision Tree (40 points)

x_1	x_2	x_3	instances of c_1	instances of c_2
0	0	0	10	0
0	0	1	0	15
0	1	0	15	0
0	1	1	0	10
1	0	0	15	0
1	0	1	10	0
1	1	0	0	0
1	1	1	0	25

Table 1: Data for Problem 3

For this problem we will use data described in table 1. This is two class classification problem with input x_1, x_2 and x_3 and class c_1 and c_2. Table summarizes the counts of instances of class 1 and 2 for each input. We define the comparison measure or cost (C_T) as weighted mean of the impurity measure of a node. So if $Q(m)$ is the impurity measure of a leaf node m and N_m is number of examples at the node m, then

$$C_T = \frac{1}{\sum_m N_m} \sum_m N_m Q(m)$$

3.1 For each input variable, compute the cost of splitting at that input variable using misclassification rate, cross entropy and Gini index as impurity measure. You can leave the results in term of log or fractions. (10 points)

3.2 Using Gini Index as criteria, decide which input variable should we first split on to create a two-level decision tree. (5 points)

3.3 Repeat the splitting procedure for the children nodes of tree in previous problem to construct a two level decision tree. Describe the process and draw the decision tree. How many points are incorrectly classified? (10 points)

3.4 Create a two-level decision tree by first splitting on x_1, and choose the second split by misclassification criteria. How many points are incorrectly classified? (10 points)

3.5 Compare and comment on different decision trees from problem 3.3 and 3.4. Which decision tree performs better and why? (5 points)

Problem 4 Naive Bayes (35 points)

A group of experts would like to identify if a work of art is a genuine masterpiece from the original artist or it is simply a fake. A binary label y indicates if the art piece at question is the original ($y = 1$) or a fake ($y = 0$), where y follows Bernoulli distribution with parameter π = P($y = 1$). They apply D tests and each test comes with a binary (pass/fail) result x_j, where $j = 1..D$, that follows a Bernoulli distribution $P(x_j|y = y_k) = \theta_{jk}^{x_j}(1 - \theta_{jk})^{(1-x_j)}$, where $k = 0, 1$. These results are aggregated in a vector $\mathbf{x} = x_1, ..., x_D$.

4.1 Show that probability of the art piece being original can be expressed as: $P(y = 1|x) = \frac{1}{1+\frac{P(y=0|x)}{P(y=1|x)}}$. (10 points)

4.2 Using Naive Bayes assumption show that probability of a prediction $P(y = k|x)$ can be expressed as:

$$P(y = k|x) = \frac{1}{Z} \exp\{\ln \pi_k + \sum_{j=1}^{D}(x_j(\ln \theta_{jk} - \ln(1 - \theta_{jk})) + \ln(1 - \theta_{jk}))\} \tag{1}$$

where Z is a normalization factor,

$$Z = p(x) = \sum_{k=0,1} P(x|y = k)p(y = k)$$

$$\pi_k = \begin{cases} \pi, \text{ if } k = 1 \\ 1 - \pi, \text{ if } k = 0 \end{cases}$$

(10 points)

4.3 Finally, using the results from the previous two questions, show that $P(y = 1|x)$ can be expressed in a form:

$$P(y = 1|x) = \frac{1}{1 + exp(-w_0 + \mathbf{w}^T\mathbf{x})} \tag{2}$$

In other words, find expression of $w_0, \mathbf{w}, \mathbf{x}$ in terms of π and θ_{jk}. (15 points)