CSCI 567, Summer 2022
Prof. Victor Adamchik

Theory Assignment 1

Due: 11:59 pm, May 26, 2022

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 Understanding Entropy (36 points)

The goal of this problem is to gain an intuitive understanding of entropy and why it is a good criteria for growing decision trees. Recall from lecture that the entropy of a discrete distribution P over C classes is defined as:

$$H(P) = -\sum_{k=1}^{C} P(Y = k) \log P(Y = k)$$

1.1 Consider a node where all samples belong to the same class i. We can write the probability distribution as follows,

$$P(Y = k) = \begin{cases} 1 & k = i \\ 0 & k \neq i \end{cases}$$

We will prove that P is a minimum entropy distribution.

(a) Prove that for any distribution P, the minimum possible entropy is 0. In other words, $H(P) \geq 0$. **(6 points)**

(b) Show that a node where all samples belong to the same class achieves minimum entropy. **(6 points)**

(c) Explain why this is desirable for a decision tree leaf. **(4 points)**

1.2 Next consider a node with an equal number of samples from all classes. Here the probability of each label is the same.

$$P(Y = k) = \frac{1}{C}$$

We'll prove that P is a maximum entropy distribution for the 2-class case ($C = 2$). Clarification: For the remainder of the problem, consider an arbitrary 2-class distribution P.

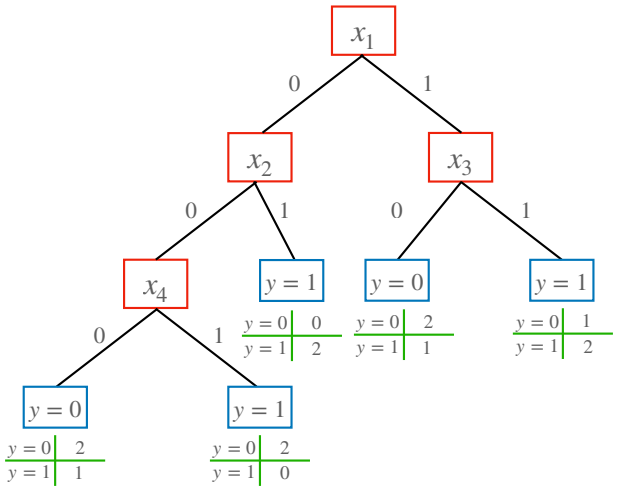
(a) Define $P_1 = P(Y = 1)$ and $P_2 = P(Y = 2)$. Rewrite $H(P)$ as a function of just P_1. **(6 points)**

(b) We want to find P such that $H(P)$ is maximized. In this case we can just find a local maximum of $H(P)$. Calculate the values of P_1 and P_2 that maximize entropy. **(10 points)**

(c) Explain why nodes with high entropy are not desirable for decision trees. **(4 points)**

Problem 2 Decision Tree Pruning (32 points)

Consider the following decision tree with features x_1, x_2, x_3 taking on values $\{0, 1\}$ (labelled in red) and predictions y at each leaf node (labelled in blue). There are 13 validation examples. The number of validation examples within each class are noted in the table below each leaf (labelled in green).



2.1 Calculate the classification error over the validation set. Show your work. **(6 points)**

2.2 Calculate the classification error if you were to prune the x_2 node. Show your work. **(7 points)**

2.3 Calculate the classification error if you were to prune the x_3 node. Show your work. **(7 points)**

2.4 Calculate the classification error if you were to prune the x_4 node. Show your work. **(7 points)**

2.5 Based on classification error, should this tree be pruned? If so, which node should be pruned? **(5 points)**

Problem 3 Nearest Neighbor Classification (20 points)

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

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

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

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

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. **(10 points)**

3.2 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 $\in \{0,1\}$. You can assume that all data points are distinct, i.e. $\forall \boldsymbol{x}_i, \boldsymbol{x}_j$ in the dataset, $\boldsymbol{x}_i \neq \boldsymbol{x}_j$.

(a) Can we have a decision tree to classify the dataset with zero classification error w.r.t. their labels? Explain your answer. **(5 points)**

(b) Can we specify a 1-NN over the dataset to result in exactly the same classification as our decision tree? Explain your answer. **(5 points)**

Problem 4 Multicollinearity in Ridge Regression (12 points)

Ridge Regression: $L2$ regularization is often used to prevent overfitting of models. We discussed in lectures that mean squared loss (MSE) is used to solve regression problems. $L2$ regularization is used with MSE loss to keep a check on the magnitude of coefficients that we learn through regression.
Multicollinearity happens when independent variables in the regression model are highly correlated to each other. We will investigate how it affects our regression coefficients.

(a) We are given a dataset **X** that has only 1 feature. $X = [X_1]$ i.e. X_{nxp} where $p = 1$. Write the ridge regression loss function (MSE + regularization) for this problem. What is the value of β_1 in terms of **X** and **Y** ? **(1 point)**

(b) We expand the feature set by adding $k - 1$ copies of this feature to our dataset as new columns to get a new dataset $X = [X_1 X_2 X_3 ... X_k]$ i.e. X_{nxp} where $p = k$ now. We fit ridge regression again. Find the new coefficients β_2 where β_2 is a $kx1$ vector. Write the loss function and show your work. Summarize how the magnitude compares with β_1 in (a). **(4 points)**

(c) We will try to carry over our understanding to higher dimensional vectors. Consider a general dataset X_{nxp} where $p > 1$. Write the loss function and the value of β using the matrix notation as discussed in class. **(1 point)**

(d) We augment the features in X with a copy of all features in the dataset. $X' = [XX]$ such that X' has dimension $nx2p$, how will the value of β change? Write the new loss function and calculate new coefficients. Show your work. **(4 points)**

(e) **Lasso and Multicollinearity :** Consider a scenario that we were using LASSO instead of ridge regression. Can you predict how the new β will be different from old β when multicollinearity is introduced? Give a brief summary. No proof or calculation required. **(2 points)**