

CSM148 Homework 3

Due date: Wednesday, Dec 1 at 11:59 AM PST

Instructions: All work must be completed independently, and submitted independently.

Start each problem on a new page, and be sure to clearly label where each problem and subproblem begins. All problems must be submitted in order (all of P1 before P2, etc.).

No late homeworks will be accepted. This is not out of a desire to be harsh, but rather out of fairness to all students in this large course.

1 Decision Trees

Consider the data given in the following table. Suppose we want to build a decision tree based on the given data using entropy gain to determine whether a person will decide to go for a run or not.

Training Data			
Weather	Temperature	Wind Level	Go out for running?
Sunny	High	Low	No
Sunny	Medium	Medium	Yes
Cloudy	High	Medium	Yes
Cloudy	Medium	High	Yes
Rainy	High	Low	No
Rainy	High	Medium	No
Sunny	Low	High	No

- (a) What is the overall entropy of running without considering any features? Suppose we want to build a decision tree by selecting a feature to split. What are the information gains for the three given features, and which feature you want to choose at the first step?
- (b) Based on the results from question (a), how should we construct a decision tree?
- (c) What are possible stop criterion for this process?
- (d) Should we standardize our features when using a Decision Tree?
- (e) Are decision trees robust to outliers?

2 True or False, Simple Explanations

Provide brief explanations for your answers.

- (a) (T or F) Bagging uses strong learners.
- (b) (T or F) The number of predictors to select from at each split in boosting always equal to the number of predictors to select at each split in Random Forest.
- (c) (Short answer) Describe the advantages and disadvantages to taking either a Bagging or Boosting approach to ensemble learning methods?
- (d) (Short answer) Explain how a Rectified Linear Unit (ReLU) activation function can potentially address the vanishing gradient issue in training Neural Networks?

3 Overfitting Mitigation Strategies

For each of the following strategies state whether or not it might help mitigate overfitting and why:

- 1. Using a smaller dataset
- 2. Allowing your model to train for fewer iterations
- 3. Increasing the number of parameters in your model
- 4. Randomly zeroing out half the nodes in a neural network
- 5. Training your model on a GPU or specialized chip instead of a CPU
- 6. Changing the initialization values for your models

4 Principal Component Analysis

- (a) What are some of the advantages and drawbacks of undertaking dimensionality reduction?
- (b) For each of the below situations, state whether or not PCA would work well, and briefly explain why.
 - Data that has a linear distribution (i.e. linear across different feature dimensions)
 - Data with a non-linear distribution (e.g., data lying on a hyperbolic plane)
 - Data that has been scaled
 - Data where each feature is statistically independent of all others

5 Perceptron

- (a) Assume you want to classify the following four points $(x_1, x_2) \in \mathbb{R}^2$: For this model we will use a

X_1	1	1	0	1
X_2	1	0	1	0
class	0	0	0	1

perceptron with an activation function in the form of

$$y = f_H(w_0 + \sum_{i=1}^n x_i w_i),$$

where f_H is the following threshold function

$$f_H(\alpha) = \begin{cases} 0, & \text{if } \alpha < 0 \\ 1, & \text{if } \alpha \geq 0 \end{cases}$$

Remember that the numbers x_i are the inputs of the unit.

Can a perceptron correctly classify this dataset with the proper set of parameters? If yes, provide an example that would satisfy the model, if not, explain.