Assignment 3

ML Class: CS 4375

February 28, 2025

1 Assignment Policies for CS 4375

The following are the policies regarding this assignment.

1. This assignment needs to be done individually by everyone.
2. You are expected to work on the assignments on your own. If I find the assignments of a group of (two or more) students very similar, the group will get zero points towards this assignment. You may possibly also be reported to the judiciary committee.
3. Please use Python for writing code. You can submit the code as a Jupyter notebook.
4. For the theory questions, please use Latex.
5. This Assignment is for 25 points.
6. This will be due on March 14th.
7. Late policy: We will deduct two points per day the assignment is submitted late.

2 Questions

1. **Linear Models (5 Points):** This question focuses on linear models for regression.
   (a) **(2 points) Linear Regression Assumptions:** Explain the key assumptions of linear regression and discuss how violating these assumptions can affect model performance.
   (b) **(3 points) Overfitting and Underfitting:** Suppose you implement a Linear Regression Model. How will you determine if the model is overfitting and underfitting. Provide atleast one way you can fix underfitting if the model is underfitting and provide one way to fix overfitting.
2. **Decision Trees (6 Points):** There are three parts to this question. The first two are for one point each and the third part if for four points.
   (a) Part 1: Describe the concept of Gini Split and how it is used in decision trees. Why is it important for determining the best feature split?
   (b) Part 2: Decision trees select features at each node based on a chosen metric. How does the tree determine which feature to split on at each step? Why might some features never be selected?
   (c) Part 3: Implement a decision tree classifier from scratch using entropy and gini split as the splitting criteria. Compare their performance with sklearn's implementation on a dataset of your choice. Implement stopping criteria (min-samples-leaf and max-depth).

3. **Nearest Neighbor Methods (6 Points):** This question is on Nearest Neighbor classification.

   (a) **(2 points) Curse of Dimensionality:** Explain how the curse of dimensionality (the performance degradation with higher dimensional features) affects the performance of nearest neighbor classifiers.

   (b) **(2 points) Feature Normalization:** Why is feature scaling important for nearest neighbor methods? Demonstrate with an example and provide examples of feature normalization methods.

   (c) **(2 points) Implementation:** Implement a k-Nearest Neighbor classifier from scratch and compare its performance with sklearn's implementation on a dataset.

4. **MLE and MAP (8 Points):** Let us consider the Dice Roll problem.

   - Define the parameters of a dice roll process and specify the total number of parameters involved.
   - Given N rolls, where each outcome i appears α_i times, write the likelihood function for the multinomial model.
   - Provide the Maximum Likelihood Estimation (MLE) for the parameters.
   - Assume a Dirichlet prior for the dice probabilities. Derive the posterior distribution and compute the Maximum A Posteriori (MAP) estimate.