Assignment 2

ML Class: CS 6375.501

October 5, 2020

1 Assignment Policies for CS 6375

The following are the policies regarding this assignment.

1. This assignment needs 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 October 9th.

2 Questions

1. **Dual and Kernels (9 points):** This question is on computing the dual of certain loss functions and the relationship between the primal and the dual problem. Each question is 3 points.

   (a) **Ridge Regression:** Ridge Regression is L2-Regularized Least Squares. Recall that, given training data $\{(x_1, y_1), \cdots, (x_M, y_M)\}$, the optimization problem is:

   $$\min_w \sum_{i=1}^{M} (y_i - w^T x_i)^2 + \lambda ||w||^2 \tag{1}$$

   Here we assume the bias is accounted for in the weight vector. For simplicity, we can also write it in Matrix form:

   $$\min_w ||Y - Xw||^2 + \lambda ||w||^2 \tag{2}$$

   where Y is the column vector $[y_1, \cdots, y_M]^T$ and $X \in R^{M \times d}$ is a Matrix such that ith row is x_i (for $i = 1, \cdots, M$). Compute the dual of this expression. **Hint:** Introduce slack variables $\zeta_i = y_i - w^T x_i$, and then write down an optimization problem over w and ζ.

(b) **L2 Regularized Squared SVM:** Consider the L2 Regularized Square SVM:

$$\min_{w,b} \sum_{i=1}^{M} [\max(0, 1 - y_i(w^T x_i + b))]^2 + \lambda ||w||^2 \tag{3}$$

Compute the dual of this objective function. Same hint as above. I.e., introduce slack variables $\zeta_i = y_i - w^T x_i - b$, and then write down an optimization problem over w, b and ζ.

(c) **From Dual to Primal for Inference:** Recall that the Kernel SVM is:

$$\max_{\lambda \geq 0, \sum_i \lambda_i y_i = 0} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j K(x_i, x_j) + \sum_i \lambda_i \tag{4}$$

Given a solution λ^* of the dual expression above, and given a new test point x_t, how will you obtain the prediction for x_t? Give an example of the prediction if the Kernel is $K(x_i, x_j) = \exp(-||x_i - x_j||^2/2\sigma^2)$.

2. **Decision Trees (5 Points):** There are three parts to this question. The first two parts are for 2 points and third part is one point.

   (a) Part 1: In the class, we saw that finding the split which maximizes the information gain $I(X_i; Y)$ [which is also called the mutual information] was a good strategy to greedily build the decision tree. Explain why this is a good strategy with an example. What if instead of the information gain, we were to use the conditional gain $H(Y|X_i)$. Would it make sense to maximize or minimize this?

   (b) Part 2: Consider the data points shown in Figure 1. Draw the approximate decision boundry obtained by the a decision tree algorithm. Also provide an approximate solution obtained by the decision tree algorithm (provide the final tree which the DT algorithm would most likely give). I am **not** expecting you to do this programmatically. Argue based on the rough position of the data points, how the tree learning algorithm will grow. Next, assume you have no limit on the depth of the tree. How would the solution look in that case?

   (c) Part 3: Which of the following is true about 'max depth' hyperparameter in the decision tree: A) Lower is better parameter in case of same validation accuracy, B) Higher is better parameter in case of same validation accuracy, C) Increase the value of max depth may overfit the data or D) Increase the value of max depth may underfit the data. More than one may be correct.

3. **Maximum Likelihood Estimation (5 Points):** This question is on the maximum likelihood estimation of a Gaussian distribution. In the class, we derived the MLE estimator of a gaussian distribution given a one dimensional dataset $x^1, \cdots, x^N$. In particular, we obtained that $\mu_{MLE} = \sum_{i=1}^{N} x_i/N$ and $\sigma^2_{MLE} = \sum_{i=1}^{N} (x_i - \mu_{MLE})^2/N$. Next, assume that the prior distribution on the mean itself is a normal distribution with mean ν and variance β^2. Compute μ_{MAP}, i.e. the MAP estimators for the mean. Also contrast μ_{MAP} to the MLE estimators (μ_{MLE}), as $N \to \infty$.

4. **Implement a Decision Tree from Scratch (6 Points):** Implement a Decision Tree from scratch. We will implement the classification variant. Use the UCI Dataset on congressional voting records: https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records. This dataset has 435 instances, and the data is categorical. The prediction task is to predict the party (democrat/republican) given the other 16 features. Implement the simple decision tree algorithm (using the information gain for feature splitting in a greedy manner), and assume the features are categorical for simplicity. Compare your performance with sk-learn for a comparable choice of hyper-parameters (depth of the tree, number of leaf nodes etc.)

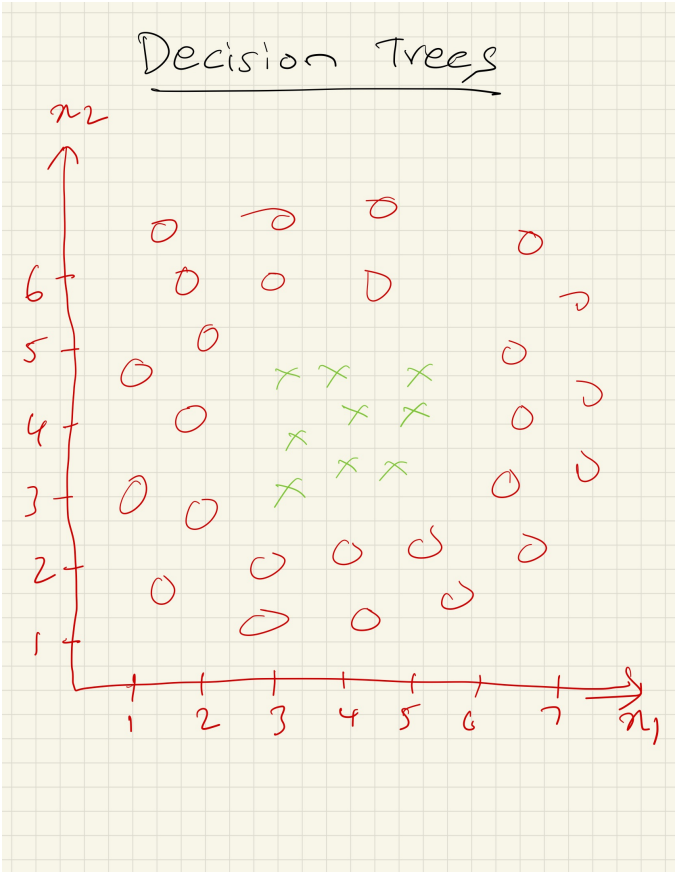

Figure 1: 2 Dimensional Data for DT Classification