Assignment 3

ML Class: CS 6375

November 13, 2024

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 50 points.
6. This will be due on Friday, December 6th. This IS A STRICT DEADLINE AND THERE WILL BE NO EXTENSIONS!!

2 Questions

1. **Naive Bayes (8 points):** Imagine you are given a dataset $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(N)}, y^{(N)}\}$, where the features $x^{(i)}$ are continuous-valued features. Answer the following questions:
   - (3 points) Explain how you will use the discrete Naive Bayes here. Please provide clear details with step-by-step procedure.
   - (5 points) Derive the expression for the continuous Naive Bayes. Describe what parameters you have, explain how you will derive the MLE estimates of those parameters, and describe how you will perform inference (i.e., given an instance, obtain the class label).
2. **Implement Logistic Regression (8 points):** Implement a regularized Logistic Regression from scratch. Pick any dataset of your choice for this (you can use the UCI Machine Learning repository: `https://archive.ics.uci.edu/ml/index.php` or you can use a dataset we used in the past). Please verify the following:
   - First, start with the regular feature set with regularization = 0. Determine if you are underfitting ot overfitting.
   - If you are not overfitting, create polynomial features the way we did in the Linear Regression demo

- Again, train a logistic regression with zero regularization and vary the polynomial degree till you overfit on the train data.
- Once you overfit on the train data, start adding regularization. What happens with regularization? Can you reduce the overfitting with regularization?

k-medoids?

3. **Clustering (8 Points):** Implement k-means and k-mediods algorithms from scratch. To test your algorithm, just use the features from the digits dataset from scikit-learn:

```
sklearn.datasets.load_digits
```

The digits dataset consists of 10 classes (digits from 0 to 9). Run your k-means and k-mediods on this dataset with $k = 10$, and visualize the obtained clustering. Compare the performance of k-means and k-mediods clustering and check if the k-means and k-mediods both reduce the clustering loss.

4. **Variance/Bias Tradeoff (8 points):** Below, you are provided two classifiers, and you need to identify the tradeoff between variance and bias in each case (i.e. for e.g. compare the second classifier to the first and ideltify if the bias/variance is lower or higher). Provide a justification as to why that is the case.

   - Logistic Regression vs Decision Tree (depth 5)
   - Decision Tree vs Random Forest vs Gradient Boosted Tree
   - Logistic Regression vs 1NN classifier
   - 5NN Classifier vs 1NN Classifier
   - Depth 5 Decision Tree vs Depth 10 Decision Tree
   - Random Forest with 100 Trees vs Random Forest with 10 Trees
   - AdaBoost with 10 Decision Stumps vs AdaBoost with 100 Decision Stumps.
   - Linear Regression vs Quadratic (Polynomial degree 2) Regression.

5. **Gradient Boosting (8 points):** In class, we studied gradient boosted decision trees with the squared L2 Loss. Specifically, $L(f(x), y) = 1/2(f(x)-y)^2$, and we compute the residual $y' = -\partial L/\partial f$. As a result, the residual in the case of the Square L2 Loss for Gradient Boosting was $y' = (y - f(x))$, where $f(x)$ is the sum of the trees learnt so far. Provide the algorithm for gradient boosting if instead of the squared L2 loss, we use the Logistic Loss and the Hinge Loss.

6. **Neural Networks (10 points):** In class, we looked at the expressions for back propogation in Neural Networks with the Sigmoid activation and the Mean Square Loss. Can you derive the expressions using a Rectified Linear Unit ($\max(x, 0)$) and using the Logistic Loss for Binary Classification.