香港城市大學
City University of Hong Kong

CS4296 Cloud Computing

Assignment 3

Full name: **Luka Moderc**

SID number: **55594415**

Question 1:

Logistic Regression

Logistic regression is a supervised learning algorithm used in classification tasks. It can be applied to predict the probability of an event, outcome, or observation. Logistic regression analyzes the relationship between independent variables and classifies data into discreet classes. It falls on the side of easier-to-implement algorithms and it is very effective when applied to linearly separable data. It is a discriminative model, meaning that the model is tasked to learn $P(y\,|X)$. Logistic regression utilizes the sigmoid function and is represented using formula:

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$

Where b_0 is called the intercept term and b_1 the coefficient. Therefore, logistic regression takes as input a real value and outputs a value between 0 and 1. To then perform binary classification we define a threshold value. Dependent on whether the output is greater than the threshold value, the model predicts if the instance belongs to the appropriate class. This approach can be abstracted to be applied to multiclass classification problems, such as the MNIST classification problem, and is known as softmax regression.

Random Forest

Random forest is a commonly used algorithm for solving supervised learning classification tasks. It is based on the decision tree model.

Decision trees are trained using divide and conquer strategy. The learning follows a greedy algorithm that searches for the best way to split data points according to some metric. The commonly used metrics are Gini impurity, means square error (MSE) and information gain. However, decision trees are prone to bias and overfitting problems. This is where random forest model comes into play. If we assemble multiple decision trees into a forest and aggregate their results, we can predict more accurate results, especially if the individual trees are uncorrelated among each other.

Random forest comprises of a collection of decision trees where each tree is trained on a subset of the training set (the data points are sampled randomly with repetition). Additionally, we perform feature bagging, meaning that each tree is trained on a randomly sampled subset of features. To

perform classification, the random forest will take a majority vote, the most frequently predicted class among the decision trees is picked. Random forest can handle both continuous and categorical variables as an input and can be used in both classification and regression tasks, although, it performs better for classification tasks.

Question 2 & 3:

I. Logistic Regression

I have experimented with different values of maxReg and elasticNetParam parameters. maxReg parameter controls the amount of regularization applied to the model, whereas elasticNetParam determines the balance between L1 and L2 regularization in elastic net regularization. The accuracy on training data can be find in table below:

maxReg / **elasticNetParam**	0.01	0.1	1.0
0.0	**0.90**	0.89	0.86
0.5	0.88	0.71	0.11
1.0	0.87	0.31	0.11

Training accuracy: **0.90**

Test accuracy: **0.90**

II. Random forest

I have experimented with different values of numTrees and maxDepth parameters. numTrees parameter specifies the number of decision trees to be included in the model, whereas maxDepth parameter controls the maximum depth of trees in the model. The accuracy on training data can be find in table below:

numTrees / maxDepth	10	20	30
3	0.65	0.72	0.73
5	0.8	0.84	0.84
10	0.92	0.93	**0.94**

Training accuracy: **0.96**

Test accuracy**: 0.94**

Question 4

Overall, random forest model performed better in classifying MNIST dataset. Logistic regression constructs only linear boundaries, hence, its accuracy peaked at 0.90. Random forest model outperformed this result, however, we can notice slight difference in accuracy achieved on training and test datasets which could be associated with slight overfitting. Additionally, it should be noted that training random forest model was considerably more time and memory consuming. Finally, I have included some additional advantages and disadvantages for each model in the following section.

Some advantages of logistic regression are:

1. model is easy to implement, train and interpret.
2. no assumptions are made about distribution of classes in feature space.
3. very fast at classifying unknown records.

Some disadvantages of logistic regression are:

1. constructs only linear boundaries.
2. bad at learning complex relationships among data.

Some advantages of random forest are:

1. robustness to noise in data.
2. handles both categorical and numerical data.
3. non-parametric nature.

Some disadvantages of random forest are:

1. high computational complexity.
2. lack of interpretability.
3. high memory usage.