Kurt Matthew Amodia
CMSC 197 - Machine Learning
Hw-4

1. **What is accuracy and when is it a misleading metric in machine learning problems? Can you provide an example of a misleading case or situation?**

   Accuracy is a metric used in machine learning that measures the ratio of correct predictions to the total number of predictions. If we put into into a mathematical form,

   $$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

   While accuracy can be a good metric to use in balanced datasets, it can cause problems in skewed datasets like the example given in the book.

   When dealing with imbalanced datasets, a high accuracy score might not reflect the true performance of the model. This is because a model can achieve a high accuracy simply by predicting the majority class most of the time, even if it completely fails to identify the minority class.

   Like the example in the book. The binary classification model where 90% of the data points belong to Class Not 5s (the majority class) and only 10% belong to Class 5s (the minority class). If a classifier always predicts "Class Not 5s" (even without learning anything), it will still have 90% accuracy because 90% of the data is in Class Not 5s. However, this classifier is practically useless because it is failing to identify any instances of Class 5s, which may be the more important class in some contexts.

2. **Give examples of at least 1 problem in which the machine learning classifier is better evaluated with the following metrics:**

   a. Precision

   $$Precision = \frac{TP}{TP + FP}$$

   An example would be a spam detection model. Precision is more recommended here because we want to lessen the number of false positives in our model, meaning that we want to minimize the number of non-spam emails being incorrectly marked as spam. High precision ensures that when an email is classified spam, it is very likely to be spam.

   b. Recall

   $$Recall = \frac{TP}{TP + FN}$$

   An example would be fraud detection classification models. Recall is more recommended here because we want to minimize the possibility of incorrectly classifying a positive fraud case as a non-fraud case. It is much more acceptable to have false positives than false negatives in this case because by having false negatives, it can result in significant financial losses. The priority is to capture all potential fraud cases to minimize losses.

3. **Increasing precision reduces recall, and vice versa. This is called the precision recall tradeoff. Discuss in at most 3 sentences why/how this tradeoff occurs.**

   The tradeoff occurs because as we increase the decision threshold, the number of positive instances decreases. This leads to an increase in precision because the number of false positives also decreases, but the number of false negatives increases, resulting in a decrease in recall. If we decrease the decision threshold, the number of positive instances increases, which raises the chance of false positives, thereby decreasing precision, but it reduces the number of false negatives, thereby increasing recall.

4. **What is logistic regression and how does it differ from linear regression? Discuss in at most 2 sentences.**

   Logistic regression, also known as logit regression, is a classification model used to estimate the probability that a given input belongs to a particular class. Unlike linear regression, which produces the result directly, logistic regression applies the sigmoid function to the result to output a number between 0 and 1, classifying inputs into the positive class if the probability is ≥ 0.5, or the negative class otherwise.

5. **Why is logistic regression suitable for binary classification? How can it be extended to multi-class classification?**

   Logistic regression is suitable for binary classification because it estimates the probability of an input belonging to one of two classes. The model assigns one class if the probability is greater than 0.5, and the other class if it is not, making it effective for distinguishing between two categories.

   Logistic regression is extended to multiclass classification using Softmax Regression (or Multinomial Logistic Regression). This method computes a score for each class and applies the softmax function to estimate the probability distribution across all classes. The predicted class is the one with the highest probability.

6. **What is the cost function in logistic regression and why can't we use meansquared error (MSE) like in linear classification?**

   In logistic regression, the cost function used is called the log loss function given

$$Cost(\Theta) = -y \log(\hat{p}) - (1-y)\log(1-\hat{p})$$

Where y is the true label (either 0 or 1).

$\hat{p}$ is the estimated probability that the instance belongs to 1 (using the sigmoid function).

We want lower log loss score because it means that there is bigger gap between the classes.

We cannot use Mean Squared Error (MSE) because in logistic regression, our goal is not to minimize the squared distances between predicted and actual values. In logistic regression, we are focused on estimating probabilities and getting it as close as possible to the true labels. If y = 1, we want the probabilities to be close to 1. If y = 0, we want the probabilities to be close to 0.