Josh Cubero
SEAS6414_HW6

Question 3 - Imbalanced Data Problem Review

Please review the following articles on the imbalanced data problem and answer the following questions:

1. What is imbalanced data in classification?

Imbalanced data in classification is a data condition that occurs when the number of samples within each class varies significantly from one class to the other. Thus, there could be one-thousand true labels, and six-hundred false labels.

2. Why is imbalanced data a problem in machine learning? There three main problems that arise when dealing with imbalanced data and its affects. The first is that imbalanced data makes it difficult to model that data and create feature correlations because there's a lower probability of selecting the lower class. Next, imbalanced data makes difficult feature class separation and identifying relevant class features. Additionally, class imbalances introduce bias into standard evaluation methods that assume data is balanced. Lastly, class imbalance causes poor generalization to unseen data as the model has not seen enough of the minority class, which leads to poor model performance.

3. What are common techniques to handle imbalanced data? The best answer to this question was found in the turintech article, which stated the best strategies at a high level. These strategies for dealing with imbalanced data are: Model-level, Evaluation-level, and Data-level. Model-level methods deal with introducing weighting into the model to penalize the majority classes, which is most commonly achieved via L1 or L2 regularization. Evaluation-level metrics deal with selecting the right evaluation metric. This entails using recall, precision, F1, and not solely accuracy. Evaluation-level metrics come with their own risks as well, since data imbalance is not addressed, these metrics are simply a work-around to the data imbalance problem. Lastly, there is Data-level methods, which deal with transforming the data itself. These methods work by employing one of two sampling methods: over or under sampling. These two strategies seek to balance the classes by altering the distribution of either the majority or minority classes.

4. How does oversampling and under sampling work? Oversampling works by synthetically creating new instances of the minority class until the imbalance is mitigated. This results in a balanced dataset, unlike evaluation and model-level methods which are merely work-arounds. Under-sampling works by removing samples from the majority class until class imbalance has dissipated. Both under and over-sampling aim to solve the imbalance problem and return a balanced dataset.

5. Give an example of a real-world use-case where extreme imbalanced data can pose a problem. Class imbalance is a prevalent problem in disease diagnosis, particularly with less common diseases such as cancer. The issue is that when handling cancer diagnosis data, negative test results are in the majority class. Thus, your model is more likely to predict a negative outcome. This becomes a problem because it's more difficult to predict the minority class, so the model will be more prone to false-negatives. A false-negative in cancer diagnosis could be extremely costly for the patient. Thus, practitioners working with cancer data must be cognizant of class imbalance.