OMNES CYPRIAN

BSc-CS

DATA IMBALANCE

Data imbalance; this refers to as a situation in which the number of data points available for different the classes is different. For most machine learning techniques, little imbalance is not a problem compared to high imbalance situation

WHY DATA IMBALANCE HAPPENS

This is because we need to determine the reflection of an unequal distribution of the classes within a dataset.

WHAT DATA ARE IMBALANCED?

Data which are imbalanced refers to as a data that does not show an unequal distribution of the classes within a dataset.

THE METHODS USED TO IMBALANCE DATA

A. Use the right evaluation metrics

Applying inappropriate evaluation metrics for model generated using imbalanced data can be dangerous. Imagine our training data is the one illustrated in graph above.

B. Resample the training set

Apart from using different evaluation criteria, one can also work on getting different dataset two approaches to make a balanced dataset out of an imbalanced one are under-sampling and oversampling.

i. Under-sampling

Under-sampling balances the dataset by reducing the size of the abundant class. This method is used when quantity of data is sufficient. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for further modelling.

ii. Over-sampling

On the contrary, oversampling is used when the quantity of data is insufficient. It tries to balance dataset by increasing the size of rare samples. Rather than getting rid of abundant samples, new rare samples are generated by using e.g. repetition, bootstrapping or SMOTE (Synthetic Minority OverSampling Technique) [1].

Note that there is no absolute advantage of one resampling method over another. Application of these two methods depends on the use case it applies to and the dataset itself. A combination of over- and under-sampling is often successful as well.

C. Use K-fold Cross-Validation in the right way

It is noteworthy that cross-validation should be applied properly while using over-sampling method to address imbalance problems

Keep in mind that over-sampling takes observed rare samples and applies bootstrapping to generate new random data based on a distribution function. If cross-validation is applied after over-sampling, basically what we are doing is overfitting our model to a specific artificial bootstrapping result. That is why cross-validation should always be done before over-sampling the data, just as how feature selection should be implemented. Only by resampling the data repeatedly, randomness can be introduced into the dataset to make sure that there won't be an overfitting problem.

Introduced into the dataset to make sure that there won't be an overfitting problem

D. Ensemble different resampled datasets

The easiest way to successfully generalize a model is by using more data. The problem is that out-ofthe-box classifiers like logistic regression or random forest tend to generalize by discarding the rare class. One easy best practice is building n models that use all the samples of the rare class and ndiffering samples of the abundant class. Given that you want to ensemble 10 models, you would keep
Eg.. the 1.000 cases of the rare class and randomly sample 10.000 cases of the abundant class. then you just split the 10.000 cases in 10 chunks and train 10 different models.
This approach is simple and perfectly horizontally scalable if you have a lot of data, since you can just train and run your models on different cluster nodes. Ensemble models also tend to generalize better, which makes this approach easy to handle.

E. Resample with different ratios

The previous approach can be fine-tuned by playing with the ratio between the rare and the abundant class. The best ratio heavily depends on the data and the models that are used. But instead of training all models with the same ratio in the ensemble, it is worth trying to ensemble different ratios. So if 10 models are trained, it might make sense to have a model that has a ratio of 1:1 (rare:abundant) and another one with 1:3, or even 2:1. Depending on the model used this can influence the weight that one class gets.

F. Cluster the abundant class

An elegant approach was proposed by Sergey on Quora [2]. Instead of relying on random samples to cover the variety of the training samples, he suggests clustering the abundant class in r groups, with r being the number of cases in r. For each group, only the medoid (centre of cluster) is kept. The model is then trained with the rare class and the medoids only.

G. Design your own models

All the previous methods focus on the data and keep the models as a fixed component. But in fact, there is no need to resample the data if the model is suited for imbalanced data. The famous XGBoost is already a good starting point if the classes are not skewed too much, because it internally takes care that the bags it trains on are not imbalanced. But then again, the data is resampled, it is just happening secretly. By designing a cost function that is penalizing wrong classification of the rare class more than wrong classifications of the abundant class, it is possible to design many models that naturally generalize in favour of the rare class. For example, tweaking an SVM to penalize wrong classifications of the rare class by the same ratio that this class is underrepresented.