Methods for preprocessing and splitting data for multi-class classification

Nguyen Thai Hoc

Institute of Applied and Technology,

September 22, 2024

Content

1. Introduction
2. Sampling techniques and methods
3. Imbalance Data and Methods Overcome
4. Conclusion

Introduction

Institute of Applied and Technology

Machine Learning LifeCycle

- The machine learning life cycle is a periodic process. It starts with the business problem, and the last stage is monitoring and optimization.

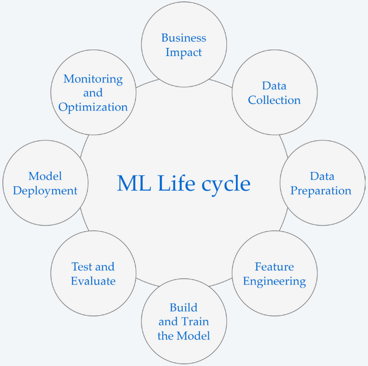


Figure: 1. ML life cycle

Data Preparation

- The data collected usually is in a raw format.
- One needs to process it so that it can be used for further analysis and model development.
- Around 70%–80% of the time goes into this stage of the ML project.
- This process of cleaning, restructuring, and standardization is known as data preparation.
- Data preparation aims to transform the raw data into a format so that EDA (Exploratory Data Analysis), can be performed efficiently to gain insights.

Challenges of Data Preparation Stage

- Missing values
- Outliers
- Disparate data format
- Data standardization
- Noise in the data
- Imbalanced data
- Lack of data

Sampling techniques and methods

Institute of Applied and Technology

What is Sampling ?

- Samling is an integral part of the ML workflow that is.
- Sampling happends in many steps of an ML project lifecycle. Sampling from a given dataset to create splits for train, validation and testing.
- First, help us avoid potential sampling biases.
- Second, help us choose the methods that improve the efficiency of the data we sample.

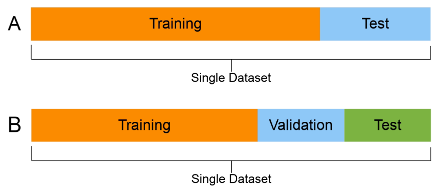


Figure: 2. Sampling of an ML project lifecycle

Nonprobability sampling

Nonprobability sampling is when the selection of data is not based on any probability.

- Convenience sampling
- Snowball sampling
- Judgment sampling

Pros:

- Nonprobability sampling can be a quick and easy way to gather initial data to get project off the round.

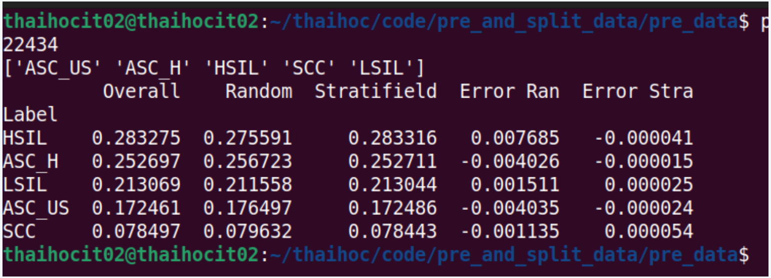
Cons:

- Criteria are not representative of the real- world data.
- Therefore are riddled with selection biases.

Probability sampling

- Simple random sampling
- Stratifield sampling
- Weight sampling

Probability sampling (cont.)



```
thaihocit02@thaihocit02:~/thaihoc/code/pre_and_split_data/pre_data$ p
22434
['ASC_US' 'ASC_H' 'HSIL' 'SCC' 'LSIL']
         Overall    Random  Stratifield  Error Ran  Error Stra
Label
HSIL    0.283275  0.275591     0.283316   0.007685   -0.000041
ASC_H   0.252697  0.256723     0.252711  -0.004026   -0.000015
LSIL    0.213069  0.211558     0.213044   0.001511    0.000025
ASC_US  0.172461  0.176497     0.172486  -0.004035   -0.000024
SCC     0.078497  0.079632     0.078443  -0.001135    0.000054
thaihocit02@thaihocit02:~/thaihoc/code/pre_and_split_data/pre_data$
```

Figure: 3. Compare between random and stratifield sampling

Imbalance Data and Methods Overcome

Institute of Applied and Technology

What is Imbalance Data ?

- Imbalance Data typically refers to a problem in classification tasks where there is a substantial difference in the number of samples in each class of the training data.
- For instance, in a dataset for the task of detecting lung cancer form X-ray images, 99% of the X-rays might be of normal lungs, and only 1% might contain cancerours cells.

Challenges of Class Imbalance

ML, especially deep leanring, works well in when the data distribution is more balanced, and usually not so well when the classes are heavily imbalanced.

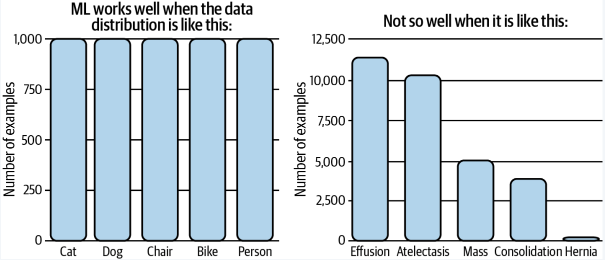


Figure: 3. ML works well in situations where the classes are balanced

Handling Class Imbalance

- Using the right evaluation metrics
- Data Augumentation
- Data-level methods: Resampling

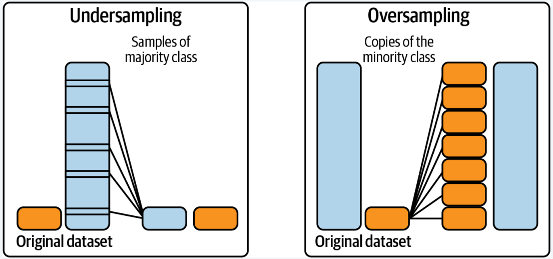


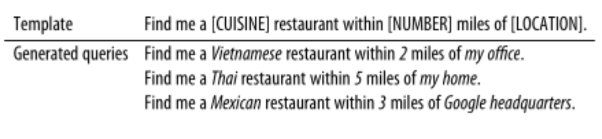
Figure: 4. Illustrations of how undersampling and oversampling work

Data Augumentation

- Data augmentation is a family of techniques that are used to increase the amount of training data.
- Traditionally, these techniques are used for tasks that have limited training data, such as in medical imaging.
- However, in the last few years, they have shown to be useful even when we have a lot of data–augmented data can make our models more robust to noise and even adversarial attacks.

Data Augumentation Methods

- Simple Label-Preserving Transformations: in computer vision, the simplest data augmentation technique is to randomly modify an image while preserving its label. You can modify the image by cropping, flipping, rotating, inverting (horizontally or vertically), erasing part of the image, and so on.
- Perturbation: Perturbation is also a label-preserving operation, but because sometimes it's used to trick models into making wrong predictions.
- Data Synthesis: Combine some training data to boost a model is performance.



Template	Find me a [CUISINE] restaurant within [NUMBER] miles of [LOCATION].
Generated queries	Find me a *Vietnamese* restaurant within *2* miles of *my office*.
	Find me a *Thai* restaurant within *5* miles of *my home*.
	Find me a *Mexican* restaurant within *3* miles of *Google headquarters*.

Figure: 5. Three sentences generated from a template

Conclusion

Institute of Applied and Technology

Highlights

- Training data still forms the foundation of modern ML algorithms. No matter how clever algorithms might be, if training data is bad, algorithms won’t be able to perform well.
- Discuss and propose methods for sampling, data augmentation, and data imbalance overcome.

Thank you everyone for your concern !

Nguyen Thai Hoc

Institute of Applied and Technology,

September 22, 2024