

**Applied Machine Learning**

**Assignment II**

Alireza Nasoodi

401131059

# Why and how

## 1- Question:

What are the challenges associated with imbalanced datasets, and how can they impact the performance of machine learning models? Additionally, what preprocessing techniques can be employed to address these challenges and improve the model's performance?

## Answer:

Imbalanced datasets are a common problem in machine learning, where the distribution of classes in the training data is not uniform. This can cause several challenges for machine learning models, including:

1. Bias towards majority class: If the majority class is much larger than the minority class, the model may become biased towards the majority class, resulting in poor performance on the minority class.
2. Overfitting: Overfitting can occur when the model is trained on an imbalanced dataset, leading to a model that performs well on the training set but poorly on the test set.
3. Misclassification of minority class: Since the minority class is underrepresented in the training data, the model may not have enough information to learn the patterns associated with the minority class, leading to misclassification.

To handle these challenges, we can use several techniques including:

1. Resampling: This involves either under-sampling the majority class or oversampling the minority class to balance the distribution of classes in the training data. Under-sampling involves randomly removing samples from the majority class, while oversampling involves duplicating samples from the minority class or generating new synthetic samples using techniques like SMOTE (Synthetic Minority Over-sampling Technique).
2. Cost-sensitive learning: This involves assigning higher costs to misclassifications of the minority class to ensure that the model prioritizes correctly classifying the minority class.
3. Ensemble methods: Ensemble methods like bagging and boosting can be used to improve the performance of imbalanced datasets by combining the predictions of multiple models.
4. Feature selection: Feature selection techniques can be used to identify the most important features associated with the minority class, allowing the model to focus on those features during training.

Overall, the choice of preprocessing technique depends on the specific characteristics of the dataset and the machine learning algorithm being used. It is important to evaluate the performance of the model on both the majority and minority classes to ensure that the model is not biased towards either class.

## 1-2-1- Question:

1. How does data normalization impact the performance of machine learning models, and what are some common techniques for scaling data to a similar range? how can the choice of normalization technique affect the results of a machine learning model, and what factors should be considered when selecting a normalization technique for a particular dataset?

## 1-2-2- Answer:

1. Data normalization can have a huge impact on the performance of machine learning models. Normalization helps to put the data on a similar scale, which makes it easier for the algorithm to learn and make accurate predictions.
2. Some common techniques for scaling data to a similar range are:
3. 1- Min-Max scaling:

It scales the data to a range between 0 and 1.

1. 2- Z-score normalization :

It normalizes scales the data to have a mean of 0 and standard deviation of 1

1. 3- log transformation:

scales the data by taking the logarithm of each value.

4- Robust Scaling:

This technique scales the data to a similar range like Min-Max scaling, but it uses the median and interquartile range instead of the minimum and maximum values. This technique is more robust to outliers than Min-Max scaling.

* + 1. The choice of normalization technique can affect the results of a machine learning model, and it depends on the characteristics of the dataset.
    2. When selecting a technique, we have to consider some factors like:
    3. 1- The distribution of the data
    4. 2- The presence of outliers
    5. 3- Specific requirements of the machine learning algorithm being used

1. 4- The computational complexity
2. In general, it is important to experiment with different normalization techniques and evaluate the impact on the performance of the machine learning model before selecting a final normalization method.

## 1-3-1- Question:

How can preprocessing techniques be used to transform categorical variables into numerical features, and what are the implications of different encoding methods on the accuracy of machine learning models? Additionally, how can the choice of encoding technique affect the computational complexity and interpretability of a machine learning model, and what factors should be considered when selecting an appropriate encoding technique for a particular dataset?

## 1-3-2- Answer:

We can use preprocessing techniques to transform categorical variables into numerical features to use them in machine learning algorithms. There are many techniques for encoding these variables, each of them with its own implications for model accuracy, computational complexity, and interpretability. For example:

1. One-hot encoding is a commonly used for converting categorical variables into binaural numerical features. In this method, each category is assigned a binary, and a new column is created for each category. The column corresponding to the category of a particular instance is assigned a value of 1, while all other columns are assigned a value of 0. One-hot encoding will give us a sparse feature matrix, where most of the values are zeros.
2. Label encoding is another technique that converts categorical variables into numerical features. In this method, every category is assigned to a unique integer value, and these integer values are used as numerical features. This method assumes an order or hierarchy between categories, which may not be appropriate at all times.
3. Binary encoding is a technique that combines the benefits of one-hot encoding and label encoding. In this method, each category is assigned a unique integer value, and these integer values are shown in binary format.

These are the most important preprocessing techniques to transform categorical variables into numerical features but there are still more of them like:

1. frequency encoding
2. target encoding
3. ordinal encoding

Choosing an encoding technique can have significant implications for model accuracy, computational complexity, and interpretability. For example, one-hot encoding can increase computational complexity, but it also can improve model accuracy. Label encoding results in a more compact feature matrix but does not capture non-ordinal relationships between categories, which will lead to reduced model accuracy.

When selecting an appropriate encoding technique, we must consider some factors like :

1. nature of the categorical variable.
2. the number of categories, the size of the dataset.
3. specific machine learning algorithm being used.

## 1-4-1- Question:

Is the decision tree a parametric model or non-parametric? Why?

## 1-4-2- Answer:

A decision tree is a non-parametric model because it doesn't make any assumptions about the underlying data distribution or the functional form of the relationship between the input features and the target variable.

## 1-5-1- Question:

How do you choose the appropriate splitting criteria for a decision tree?

## 1-5-2- Answer:

Choosing the appropriate splitting criteria for a decision tree is selecting a metric that can effectively partition the data based on the values of the input features. Some of the used metrics for this job are:

1. Gini Index: It measures the impurity of a node by calculating the probability of misclassifying a random sample. The lower the Gini index, the better the split.
2. Information Gain: It measures the reduction in uncertainty in a node after splitting based on a particular feature. The higher the information gain, the better the split.
3. Chi-Square: It measures the difference between the expected and observed frequencies of each class in a node. The higher the chi-square value, the better the split.
4. Reduction in Variance: This measures the reduction in variance of the target variable after splitting a node. The higher the reduction in variance, the better the split.

Overall choosing between this metrics for splitting a decision tree is heavily dependent on the nature of the data and the problem itself.

## 1-6-1- Question:

Consider the following table, apply decision tree classification with ID3 algorithm and plot it. Show all the steps as well as the formulas used. Which of the following features is the most important one?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Id** | **F1** | **F2** | **F3** | **F4** | **F5** | **Target** |
| **1** | **-** | **-** | **-** | **+** | **+** | **1** |
| **2** | **-** | **+** | **-** | **+** | **+** | **1** |
| **3** | **+** | **-** | **-** | **+** | **-** | **0** |
| **4** | **+** | **+** | **+** | **+** | **-** | **1** |

## 1-6-2- Answer:

For answering this question first of all we have to calculate the entropy of the dataset.

Where **n** is the total number of classes in the target column and pᵢ is the probability of class i. The calculated Entropy is:

Next, we calculate the information gain for each feature:

where F is a feature, Sa(v) is the set of training examples of S such for which attribute a is equal to v.

Now we calculate the Information Gained from the features:

As we can see the information gained from the feature F1,F2,F5 are equal. So we choose one of them and pursue our task.

We choose F1 as our root node in the decision tree. As we can see if the value of F1 is (-) target will be 1, but if the f1 is (+) it depends on the next feature. For selecting the best node to decide what feature would we consider for placing after the f1 node we have to calculate the gained data from these features again.

## 1-7-1- Question:

## 1-7-2- Answer:

## 1-8-1- Question:

## 1-8-2- Answer:

## 1-9-1- Question:

## 1-9-2- Answer:

## 1-10-1- Question:

## 1-10-2- Answer:

## 1-11-1- Question:

## 1-11-2- Answer:

## 1-12-1- Question:

## 1-12-2- Answer: