

L02: Classification

Classification predicts the class or category of a new observation based on previous examples. Several key elements are needed to develop a classification model:

1. **Features that can be quantified:** Identify the relevant features or variables that can help distinguish between different classes. These features, such as numerical or categorical data, should be measurable and quantifiable to facilitate analysis.
2. **Labels are known:** In supervised machine learning, labelled data is required to train the model. These labels should accurately reflect the true class or category of the observations, allowing the model to learn from these examples and make predictions on new data.
3. **Method to measure similarity:** measure the similarity or distance between observations, such as *Euclidean distance, cosine similarity, or Jaccard similarity*.

Exercise 1: Develop a classification model that can accurately predict the type of flowers a customer will likely purchase based on the similarity to their most recent purchase.

Answer

- i. Data Collection: Collect the data, such as the type of flowers purchased, the date, and other relevant information, such as the occasion for the purchase.
- ii. Feature Extraction: Extract the relevant features, such as the type of flowers purchased, the frequency of purchase, and the seasonality of purchase.
- iii. Labelling: Label the data based on the type of flowers purchased. For example, roses, lilies, tulips, etc.
- iv. Similarity Measure: Use similarity measure to predict the type of flowers a customer is likely to purchase. This could be based on the type of flowers purchased, the frequency of purchase, and the seasonality of purchase.
- v. Model Training: Train a classification model using the labeled data and similarity measures. Various classification algorithms, such as logistic regression, decision trees, or neural networks, can be used for this purpose.
- vi. Model Evaluation: The model's accuracy can be evaluated using a test set of data that was not used during training. This can be done by comparing the predicted type of flowers with the actual type of purchased flowers.
- vii. Prediction: Finally, the developed model can predict the type of flowers a customer will likely purchase based on the similarity to their most recent purchase.

Types of classification tasks:

- Binary Classification
- Multi-Class Classification
- Multi-Label Classification
- Imbalanced Classification

Binary Classification:

- Classify items into one of two categories.
- Examples include spam detection (is an email spam or not), fraud detection (is a transaction fraudulent or not), and sentiment analysis (is a tweet positive or negative).
- The output is a binary decision, such as yes or no, true or false, or 0 or 1.
- Other examples include disease diagnosis (does a patient have a disease or not) and credit risk assessment (will a borrower default on a loan or not).

Multi-Class Classification:

- To classify items into one of three or more categories.
- Examples include image classification (classifying images into various objects, scenes, or activities), topic classification (categorizing text into various topics), and language identification (identifying the language of a given text).
- The output is a categorical decision, such as class labels or text descriptions.
- Other examples include handwritten digit recognition (classifying handwritten digits from 0 to 9), facial expression recognition (recognizing facial expressions such as happy, sad, or angry), and disease classification (classifying types of cancer or types of infections).

Multi-Label Classification:

- To assign multiple labels to a single instance.
- Examples include text categorization (assigning multiple tags or categories to a given text), image tagging (assigning multiple labels to an image), and music genre classification (assigning multiple genres to a song).
- The output is a set of labels, which can be binary or multi-class.
- Other examples include product categorization (assigning multiple categories to a product), and video classification (assigning multiple labels to a video based on its content).

Imbalanced Classification:

- To classify items when one class is much less frequent than the others.
- Examples include fraud detection (detecting rare fraudulent transactions), disease diagnosis (detecting rare diseases), and rare event prediction (predicting rare natural disasters).
- The output is usually a binary decision, but the performance is measured with specialized metrics that account for the class imbalance.
- Other examples include credit card fraud detection (detecting rare cases of credit card fraud), and defect detection in manufacturing (detecting rare defects in products during manufacturing).

Binary Classification example:

- To classify a credit card transaction as either fraudulent or not fraudulent. The features used in this classification task: the transaction amount, location, and time.
- The labels are known because the credit card companies have historical data of fraudulent and non-fraudulent transactions.
- To measure similarity, machine learning algorithms can be trained on historical data to identify patterns of fraudulent transactions.

Multi-Class Classification example:

- To classify products such as electronics, clothing, or home appliances. This task is important for e-commerce websites as it helps organise the products and display them to the users in a meaningful way.
- The features used: product descriptions, images, and specifications.
- The labels are known because the website owners have pre-defined product categories.
- To measure similarity, machine learning algorithms can be trained to learn patterns in the product descriptions, images, and specifications to assign the products to the appropriate categories.

Multi-Label Classification example:

- To assign tags or labels to the content of the blog post or news article that describes its content or topic.
- The features used: the text of the blog post or news article.
- The labels are not mutually exclusive because a blog post or news article can cover multiple topics.
- To measure similarity, machine learning algorithms can be trained to identify patterns in the text of the blog post or news article and assign the appropriate tags to it.

Imbalanced Classification example:

- To detect a rare disease.
- The features used: medical symptoms, family history, and medical history.
- The labels are known because medical experts have labeled the diseases as rare or common based on their frequency.
- To measure similarity, machine learning algorithms can be trained on historical medical data to identify patterns of rare disease symptoms and detect the disease in the patient's medical history.

Exercise 2

A bank is looking to improve its credit risk assessment process using machine learning. For example, they have a dataset of past loan applications and their outcomes. The bank wants to predict whether or not a new loan application is likely to be approved or denied based on the applicant's information, such as income, credit score, age, and other factors.

- i) What is the purpose of the classification task in the bank's scenario?
- ii) What are the features used in the classification task for the bank's scenario?
- iii) What are the labels in the classification task for the bank's scenario?
- iv) What method is used to measure similarity in the classification task for the bank's scenario?
- v) What type of classification task is being used in the bank's scenario?
- vi) Give one example of a binary classification task in the context of the bank's scenario.
- vii) Give one example of a multi-class classification task in the context of the bank's scenario.
- viii) Give one example of a multi-label classification task in the context of the bank's scenario.
- ix) What is the main challenge of imbalanced classification in the context of the bank's scenario?

Answers

- i) To predict whether a new loan application is likely to be approved or denied based on the applicant's information.
- ii) The features used: income, credit score, age, and other factors of the loan applicant.
- iii) The labels used: approved or denied.
- iv) The method used to measure similarity (distance metric): Euclidean distance or cosine similarity, depending on the specific machine learning algorithm used.
- v) Binary classification.
- vi) To predict whether a loan applicant will default on their loan or not.
- vii) To predict whether a loan applicant will be approved, denied, or offered a different type of loan such as a higher interest rate or lower loan amount.
- viii) To predict whether a loan applicant is likely to default on their loan and whether they are likely to make their payments on time.
- ix) The main challenge of imbalanced classification in the context of the bank's scenario is that there may be many more approved loans than denied loans in the dataset.