# Lecture\_2b

Hello and welcome to the second lecture in this week. We're going to be looking at classification, which is one of the forms of supervised learning. In this video, we are going to have a recap of supervised learning. We're going to discuss what classification is in the context of machine learning.

We are going to discuss the types of classification and machine learning.

Supervised learning techniques are a form of classification or regression. Classification techniques mainly predict discrete responses and regression techniques mainly predict continuous responses.

Classification techniques are mainly employed if the dataset can be tagged or categorised or separated into specific groups or classes. In machine learning, classification can be simply defined as the predictive modelling problem where a unique class label, our group is predicted for a given set of input data. An example is this: if we have a dataset of mail, we could want to classify a given mail or observation from this dataset as spam or non-spam. Classification requires a training dataset with many sets of inputs and outputs to learn, train and build classification models. Typically, a classification model will use the training datasets to determine how best to map examples of input data (i.e., each observation) to specific class labels or groups. And it's very important that training dataset sufficiently represents the problem by having as many examples of each class lable as possible. A rule of thumb to remember is this: using larger training datasets for supervised learning often yields models that generalise suitably and accurately well for new data.

Some common terms used in machine learning-based classification include, but are not limited to: observation, which refers to a data point, row, or sample in a dataset. Training data set, which refers to a set of observations used to generate machine learning models. Validation dataset, which refers to a set of observations used during model training to derive feedback on how well the current parameters, that is model parameters, generalise beyond the training set. We also have terms like test dataset, which refers to a set of observations used at the end of model training and validation to evaluate the predictive power and the predictive accuracy of the model. We also have label, which is the answer portion of an observation in supervisor landing. It could also be called a class, a group or category. We have a feature. Now, if we have a dataset to be used to classify flowers into different species. the features might include the petal length and the petal width, while the label would be the flowers' species. Classification problems could take on several forms in machine learning. The most popular forms are: unary classification, which is a special case and this has been included for emphasis, Binary classification, multiclass classification and multi-label classification. We also have imbalanced classification. Unary classification is also called one-class classification or class modelling and it works mainly by trying to identify objects of a specific class of group amongst all objects, by learning from a training dataset containing mainly or only the objects of that class and it's mainly employed for outlier detection. So, as illustrated in the diagram here, which is to classify apples from pairs, you can see that an apple that has been bitten into, which is obviously an outlier, is currently classified, or categorised as a pair, which is not correct.

Unary classification can be employed to address such situation. Compared to conventional classification problems in machine learning, when we have training dataset for all classes or groups, unary classification is much more complex. We have a number of unary classification learning logarithms or methods and they are mainly employed at the risk of repetition for outliner detection, anomaly detection and novelty detection. Some real world examples of unary classification problems include: the classification of the network traffic in a secure software-defined network as normal. In this scenario, there are a few, if any, examples of the traffic under an attack, that is the network traffic now when the network is under an attack or during an intrusion. However, the statistics of normal traffic, that is normal network traffic are well known and established. Another example could be monitoring the operational status of nuclear power plants as normal. There are few, if any, examples of catastrophic system states for the nuclear power plants. However, the statistics of normal operations of the nuclear power plants are well-known and established, and these two examples are classification problems that require unary classification algorithms. Binary classification, sometimes called binomial classification.

In statistical classification, any label is a member of a finite set of classes and if the size of the set of classes is two, then the classification problem is a binary or binomial classification problem, provided there is one label per training sample or observational data point as the case may be. Binary classification can be defined simply as the task of classifying the elements of a set into two groups based on the classification rule. Take for instance, if we have a dataset of patients, we could model a binary classifier that classifies the patients into a sick bucket or a healthy bucket. Many binary classification learning algorithms or methods are also available. Multiclass classification is also sometimes called multinomial classification. And as stated before, in statistical classification, any label is a member of a finite set of classes. If the size of the set of classes is three or more, then the classification problem is a multiclass or multinomial classification problem, provided there is one label per training sample or observation.

Multiclass classification, as you can tell, is very similar to binary classification. However, there are more than two classes or categories or labels involved. Similar to unary classification and binary classification, in terms of the algorithms that can be used to implement multiclass classification, many classification learning algorithms naturally allow for more than two classes, while others are by nature binary classification algorithms. They are also efficient strategies in place which can allow binary classification learning algorithms to be augmented into multiclass algorithms. Examples of multiclass classification problems could be qualitative assessment-based credit scoring, where you have to assign poor, fair, good or excellent credit scores to each observation. Multilabel classification is somewhat similar to multiclass classification, based on the number of labels involved. However, in multilabel classification each training example or observation doesn't just have one label, but could have a number of labels. Take for instance, if we had to describe an image, we could assign several labels to it. We could have people, truck, vegetation, lion, all four label at the same time for a single picture like we have. Analogously, algorithms that naturally can be made multiclass can be applied to multilabel classification problems as well.

Imbalance classification strongly focuses on classification tasks where the number of observations in each class is unequally distributed. In other words, the distribution of observations across the known classes of labels is biased or skewed, and distribution could also vary from a slight bias to a severe imbalance. Take for instance one observation in the minority class for hundreds, thousands or even millions of observations in the majority class. or classes, as the case may be. Conventionally imbalanced classification tasks can be modelled as binary classification tasks. Majority of observations in the training dataset can be said to belong to one class, that is the normal class and a minority of observations to the other class abnormal class. Intuitively, we can tell that some real world examples that would employ imbalanced classification that has been modelled into binary classification would be fraud detection, outlier detection and medical diagnostic tests. In this video, we've had a recap of supervised learning, we've discussed what is classification in the context of machine learning and we've also discussed the types of classification in machine learning.