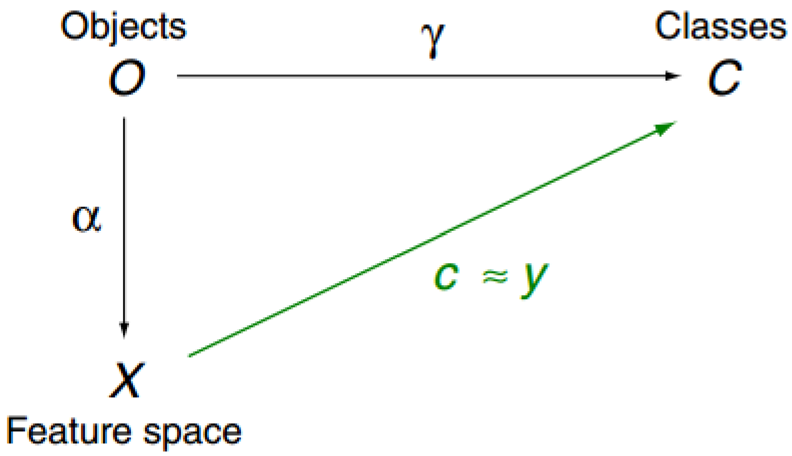
Classification

Classification is the statistical analysis of data, where there is predetermined knowledge about the data. This predetermination is used to guide the statistical analysis based on predefined classes. In contrast to clustering which is an unsupervised learning method, classification is instead supervised. As mentioned supervised means that there is already some knowledge about the data, that is, all the objects of the data has been assigned a class.

With classification the learning method is different from clustering. Classification is done by learning the pattern that specifies each class of the data. The learning is therefore done by finding similarities between the objects in each class, which means that once new data is added, they can be classified based on the current classes. This also means that new classes cannot be added at a later point without having to relearn the entire dataset again.

Hypothesis Space

The hypothesis space is all the functions that approximate the correct result.



Let's start off from O. O indicates objects in our real world, for instance human, rain or not, gender and etc. X is some features, also called feature space of an object, for example, hair color, voice and etc for human. And between O and X, α is the process of abstraction from O to X. And the rest capital C is the class of the Object.

As we can see, there are two approaches from O to C: O ⇢⇢ C directly and O ⇢⇢ X ⇢⇢ C. Since it's almost impossible for us to take the first path duo to the unbounded features of an object, we should do that with the help of X(pattern learning). Assuming that we have a finite features, hair color and etc, in order to tell if a human is female or male, we hope to have an idealized function C to classify the samples(of object) by their features, without any error. But that's not feasible. We thus need a function y to approximate that ideal function.

Then all the possibilities of y form the hypothesis space.

Bias

Bias is when an algorithm tends to lean towards some hypothesis, due to restrictions of its hypothesis space.

- The bias is error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The variance is error from sensitivity to small fluctuations in the training set. High variance can cause overfitting: modeling the random noise in the training data, rather than the intended outputs.

However, a high bias is often the result of a lack of data, because less data is used to describe the different classes.

Models with a high bias are generally more simple and don’t overfit, however, they may not capture the real picture as detailed as a lower bias would.

In contrast as low bias generally fits the training data very well, but may not fit as well on new unseen data.

Example

An example is the *k*-nearest neighbor algorithm. The lower *k* is, the higher the bias, since it classifies the object based on fewer close objects. A high *k* is resulting in a low bias, since it bases the classification on many more objects.

Generalization

Generalization is a measure of how accurately an algorithm is able to predict outcome values for previously unseen data. Because learning algorithms are evaluated on finite samples, the evaluation may be sensitive to sampling error. Generalization can be minimized by avoiding overfitting in the learning algorithm.

The concepts of generalization and overfitting are closely related. Overfitting occurs when the learned function becomes to sensitive to noise in the sample. As a result the function will perform well on the training data, but not so well on new data. Thus, the more overfitting occurs, the larger the generalization error.

Evaluation

Binary evaluation

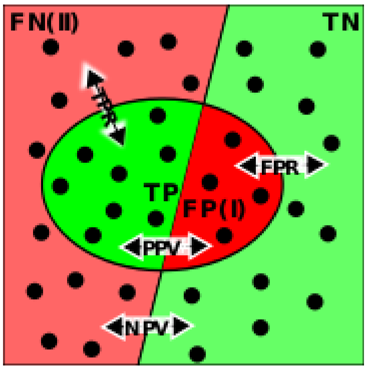
For an example with sickness:

True positive: The percentage of sick people who are identified as sick.

True negative: The percentage of healthy people who are identified as healthy.

False positive: The percentage of sick people who are falsely identified as being healthy.

False negative: The percentage of healthy people who are falsely identified as being sick.


FN(II)
TN
TPR
FPR
TP
FP(I)
PPV
NPV