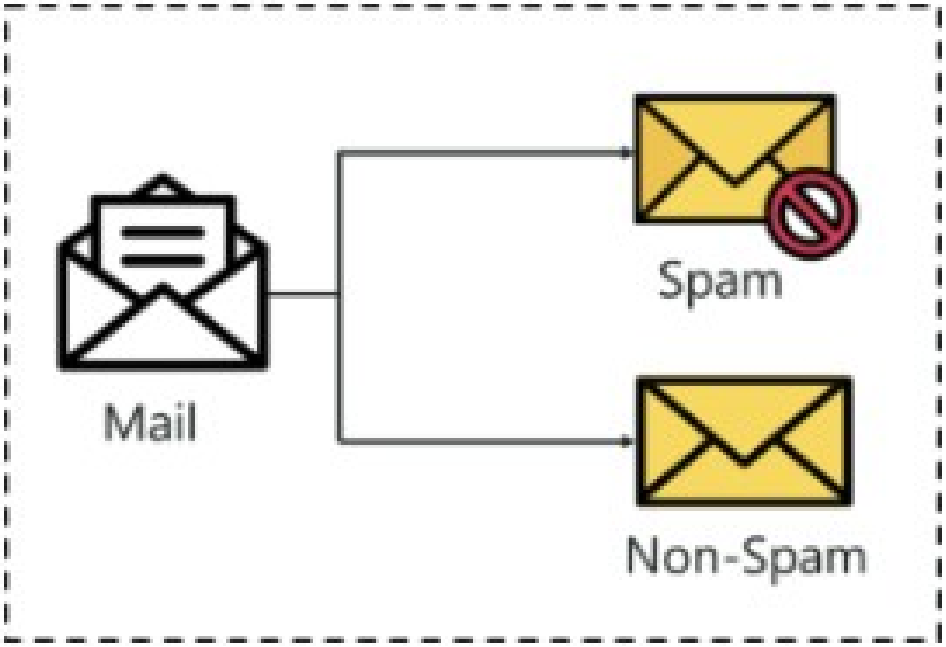
Classification Characterization

Tushar B. Kute,
http://tusharkute.com

What is Classification?

- Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data.
- The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.
- The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables.
- The main goal is to identify which class/category the new data will fall into.

Example:



Example:

- Heart disease detection can be identified as a classification problem, this is a binary classification since there can be only two classes i.e has heart disease or does not have heart disease.
- The classifier, in this case, needs training data to understand how the given input variables are related to the class. And once the classifier is trained accurately, it can be used to detect whether heart disease is there or not for a particular patient.
- Since classification is a type of supervised learning, even the targets are also provided with the input data.

Types of Classification

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

Binary Classification

- Binary classification refers to those classification tasks that have two class labels.
- Examples include:
  - Email spam detection (spam or not)
  - Churn prediction (churn or not).
  - Conversion prediction (buy or not).
- Typically, binary classification tasks involve one class that is the normal state and another class that is the abnormal state.

Binary Classification – Example

- For example “not spam” is the normal state and “spam” is the abnormal state. Another example is “cancer not detected” is the normal state of a task that involves a medical test and “cancer detected” is the abnormal state.
- The class for the normal state is assigned the class label 0 and the class with the abnormal state is assigned the class label 1.
- It is common to model a binary classification task with a model that predicts a Bernoulli probability distribution for each example.

Evaluation of Binary Classifier

- There are many metrics that can be used to measure the performance of a classifier or predictor; different fields have different preferences for specific metrics due to different goals.
- In medicine sensitivity and specificity are often used, while in information retrieval precision and recall are preferred.
- An important distinction is between metrics that are independent of how often each category occurs in the population (the prevalence), and metrics that depend on the prevalence – both types are useful, but they have very different properties.

Evaluation of Binary Classifier

- Given a classification of a specific data set, there are four basic combinations of actual data category and assigned category: true positives TP (correct positive assignments), true negatives TN (correct negative assignments), false positives FP (incorrect positive assignments), and false negatives FN (incorrect negative assignments).

	Condition positive	**Condition negative**
Test outcome positive	True positive	False positive
Test outcome negative	False negative	True negative

Confusion Matrix

- In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix).
- Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class (or vice versa).
- The name stems from the fact that it makes it easy to see whether the system is confusing two classes (i.e. commonly mislabeling one as another).

Confusion Matrix

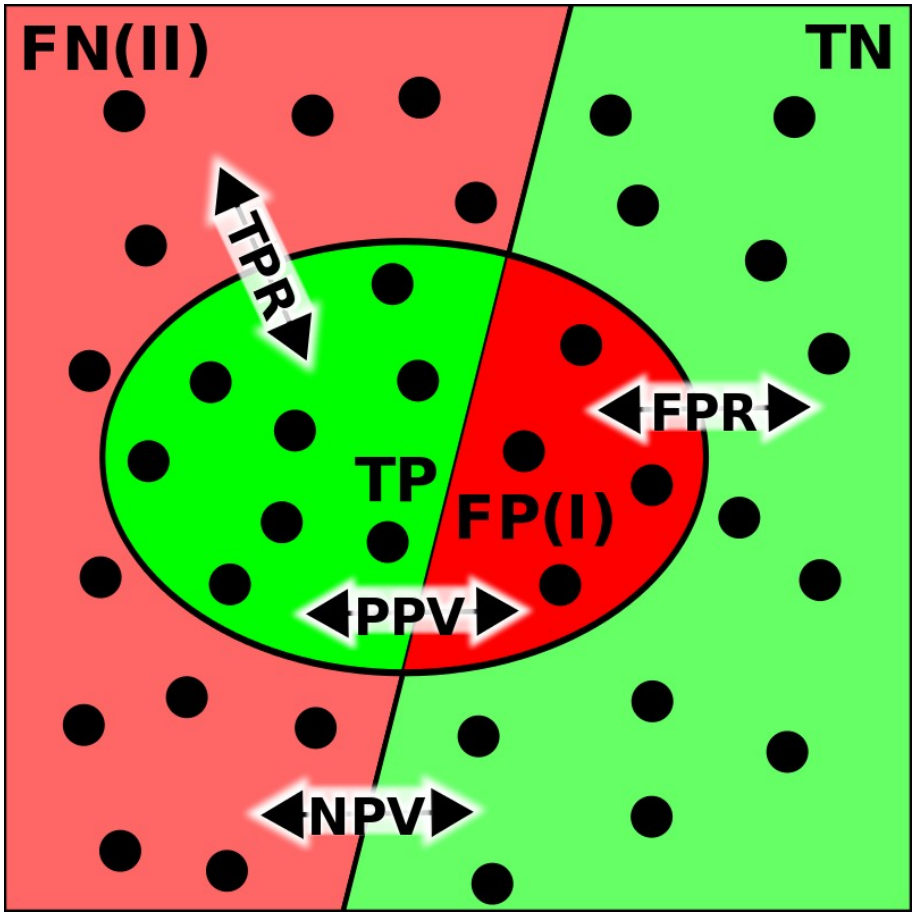
- Given a sample of 13 pictures, 8 of cats and 5 of dogs, where cats belong to class 1 and dogs belong to class 0,
  - actual = [1,1,1,1,1,1,1,1,0,0,0,0,0],
- assume that a classifier that distinguishes between cats and dogs is trained, and we take the 13 pictures and run them through the classifier, and the classifier makes 8 accurate predictions and misses 5: 3 cats wrongly predicted as dogs (first 3 predictions) and 2 dogs wrongly predicted as cats (last 2 predictions).
  - prediction = [0,0,0,1,1,1,1,1,0,0,0,1,1]

Confusion Matrix

		Actual class	
		Cat	**Dog**
Predicted class	**Cat**	**5**	2
	Dog	3	**3**

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	**5 true positives**	2 false positives
	Non-cat	3 false negatives	**3 true negatives**

Evaluation of Binary Classifier



Eight basic ratios

- There are eight basic ratios that one can compute from this table, which come in four complementary pairs (each pair summing to 1).
- These are obtained by dividing each of the four numbers by the sum of its row or column, yielding eight numbers, which can be referred to generically in the form "true positive row ratio" or "false negative column ratio".
- There are thus two pairs of column ratios and two pairs of row ratios, and one can summarize these with four numbers by choosing one ratio from each pair – the other four numbers are the complements.

Column Ratios

- true positive rate (TPR) = (TP/(TP+FN)), aka sensitivity or recall. These are the proportion of the population with the condition for which the test is correct.
  - with complement the false negative rate (FNR) = (FN/(TP+FN))
- true negative rate (TNR) = (TN/(TN+FP), aka specificity (SPC),
  - with complement false positive rate (FPR) = (FP/(TN+FP)), also called independent of prevalence

Row Ratios

- positive predictive value (PPV, aka precision) (TP/(TP+FP)). These are the proportion of the population with a given test result for which the test is correct.
  - with complement the false discovery rate (FDR) (FP/(TP+FP))
- negative predictive value (NPV) (TN/(TN+FN))
  - with complement the false omission rate (FOR) (FN/(TN+FN)), also called dependence on prevalence.

More Ratios

- There are a number of other metrics, most simply the accuracy or Fraction Correct (FC), which measures the fraction of all instances that are correctly categorized; the complement is the Fraction Incorrect (FiC).
- The F-score combines precision and recall into one number via a choice of weighing, most simply equal weighing, as the balanced F-score (F1 score).
- Some metrics come from regression coefficients: the markedness and the informedness, and their geometric mean, the Matthews correlation coefficient.
- Other metrics include Youden's J statistic, the uncertainty coefficient, the phi coefficient, and Cohen's kappa.

F1 Score / Harmonic Mean

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}.$$

Multi Class Classification

- Multi-class classification refers to those classification tasks that have more than two class labels.
- Examples include:
  - Face classification.
  - Plant species classification.
  - Optical character recognition.
- Unlike binary classification, multi-class classification does not have the notion of normal and abnormal outcomes. Instead, examples are classified as belonging to one among a range of known classes.

Multi Class Classification

- The number of class labels may be very large on some problems. For example, a model may predict a photo as belonging to one among thousands or tens of thousands of faces in a face recognition system.
- Problems that involve predicting a sequence of words, such as text translation models, may also be considered a special type of multi-class classification.
- Each word in the sequence of words to be predicted involves a multi-class classification where the size of the vocabulary defines the number of possible classes that may be predicted and could be tens or hundreds of thousands of words in size.

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com