UNIT-1

Chapter-1

The ingredients of machine learning

1. **Tasks: the problems that can be solved with machine learning**
2. **Models: the output of machine learning**
3. **Features: the workhorses of machine learning**

Tasks: The problems that can be solved with machine learning

Spam e-mail recognition was described in the Prologue. It constitutes a binary classification task, which is easily the most common task in machine learning which figures heavily throughout the book. One obvious variation is to consider classification problems with more than two classes. For instance, we may want to distinguish different
kinds of ham e-mails, e.g., work-related e-mails and private messages. We could approach this as a combination of two binary classification tasks: the first task is to distinguish between spam and ham, and the second task is, among ham e-mails, to distinguish between work-related and private ones.

a) **Multiclass** or **multinomial classification**

b) Regression analysis

c) Cluster analysis

d) Association rule learning

In machine learning, **multiclass** or **multinomial classification** is the problem of classifying instances into one of three or more classes. (Classifying instances into one of two classes is called binary classification.)

While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms; these can, however, be turned into multinomial classifiers by a variety of strategies.

Multiclass classification should not be confused with multi-label classification, where multiple labels are to be predicted for each instance.

Regression analysis

regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which a researcher finds the line (or a more complex linear function) that most closely fits the data according to a specific mathematical criterion.

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

Models: The output of machine learning

Models form the central concept in machine learning as they are what is being learned from the data, in order to solve a given task. There is a considerable – not to say bewildering – range of machine learning models to choose from.

1. *Geometric models*
2. *Probabilistic models*
3. *Logical models*
4. *Grouping and grading*

→A ***geometric model*** is constructed directly in instance space, using geometric concepts such as lines, planes and distances. One main advantage of geometric classifiers is that they are easy to visualise, as long as we keep to two or three dimensions.

→**probabilistic classifier** is a classifier that is able to predict, given an observation of an input, a probability distribution over a set of classes, rather than only outputting the most likely class that the observation should belong to. Probabilistic classifiers provide classification that can be useful in its own right or when combining classifiers into ensembles.

→Logical models

Logic models are hypothesized descriptions of the chain of causes and effects leading to an outcome of interest (e.g. prevalence of cardiovascular diseases, annual traffic collision, etc). While they can be in a narrative form, logic model usually take form in a graphical depiction of the "if-then" (causal) relationships between the various elements leading to the outcome. However, the logic model is more than the graphical depiction: it is also the theories, scientific evidences, assumptions and beliefs that support it and the various processes behind it.

→Grouping and grading

Grouping models do this by breaking up the instance space into groups or *segments*, the number of which is determined at training time. One could say that grouping models have a fixed and finite 'resolution' and cannot distinguish between individual instances beyond this resolution.

Features: the workhorses of machine learning

→***Univariate model***

→***Binary splits***

→***Univariate model :*** In mathematics, **univariate** refers to an expression, equation, function or polynomial of only one variable. Objects of any of these types involving more than one variable may be called multivariate. In some cases the distinction between the univariate and multivariate cases is fundamental; for example, the fundamental theorem of algebra and Euclid's algorithm for polynomials are fundamental properties of univariate polynomials that cannot be generalized to multivariate polynomials.

→**Binary splitting** is a technique for speeding up numerical evaluation of many types of series with rational terms. In particular, it can be used to evaluate hyper geometric series at rational points.

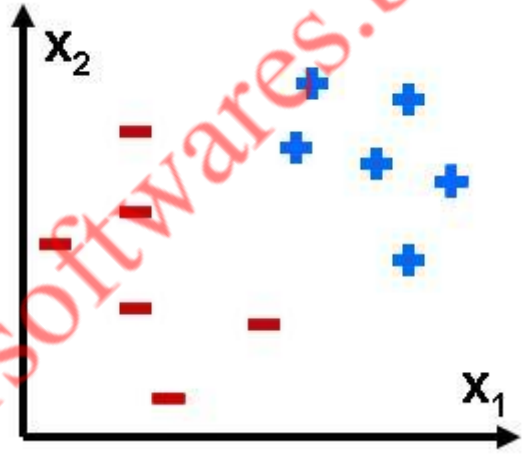
Chapter-2

Binary classification and related tasks

Classification systematic arrangement in groups or categories according to established criteria. Classification is the most common task in machine learning. A *classifier* is a mapping $\hat{c}: X \rightarrow C$, where $C = \{C1, C2, \ldots, Ck\}$ is a finite and usually small set of *class labels*. We will sometimes also use *Ci* to indicate the set of examples of that class. We use the 'hat' to indicate that $\hat{c}(x)$ is an estimate of the true but unknown function $c(x)$.

Examples for a classifier take the form $(x, c(x))$, where $x \in X$ is an instance and $c(x)$ is the true class of the instance. Learning a classifier involves constructing the function $\hat{c}$ such that it matches c as closely as possible (and not just on the training set, but ideally on the entire instance space X).

In the simplest case we have only two classes which are usually referred to as *positive* and *negative*, ⊕ and ⊖, or +1 and −1. Two-class classification is often called *binary classification* (or *concept learning*, if the positive class can be meaningfully called a concept). Spam e-mail filtering is a good example of binary classification, in which spam is conventionally taken as the positive class, and Hamas the negative class (clearly, positive here doesn't mean 'good'!). Other examples of binary classification include medical diagnosis (the positive class here is having a particular disease) and credit card fraud detection.



Visualising classification performance

→*Coverage plot*: data is displayed graphically in a coverage plot. The more sequence reads you have in a region, the higher the plot is. More RNA sequence reads means more gene expression.

→*Degrees of freedom:* each of a number of independently variable factors affecting the range of states in which a system may exist, in particular any of the directions in which independent motion can occur.

Scoring and ranking

Variable **Ranking** is the process of ordering the features by the value of some **scoring** function, which usually measures feature-relevance. Resulting set: The **score** S(fi) is computed from the training data, measuring some criteria of feature fi. By convention a high **score** is indicative for a valuable (relevant) feature.

List of scoring modules

Machine Learning Studio (classic) provides many different scoring modules. You select one depending on the type of model you are using, or the type of scoring task you are performing:

- Apply Transformation: Applies a well-specified data transformation to a dataset.

  Use this module to apply a saved process to a set of data.

- Assign Data to Clusters: Assigns data to clusters by using an existing trained clustering model.

  Use this module if you want to cluster new data based on an existing K-Means clustering model.

  This module replaces the Assign to Clusters (deprecated) module, which has been deprecated but is still available for use in existing experiments.

- Score Matchbox Recommender: Scores predictions for a dataset by using the Matchbox recommender.

  Use this module if you want to generate recommendations, find related items or users, or predict ratings.

- Score Model: Scores predictions for a trained classification or regression model.

  Use this module for all other regression and classification models, as well as some anomaly detection models.

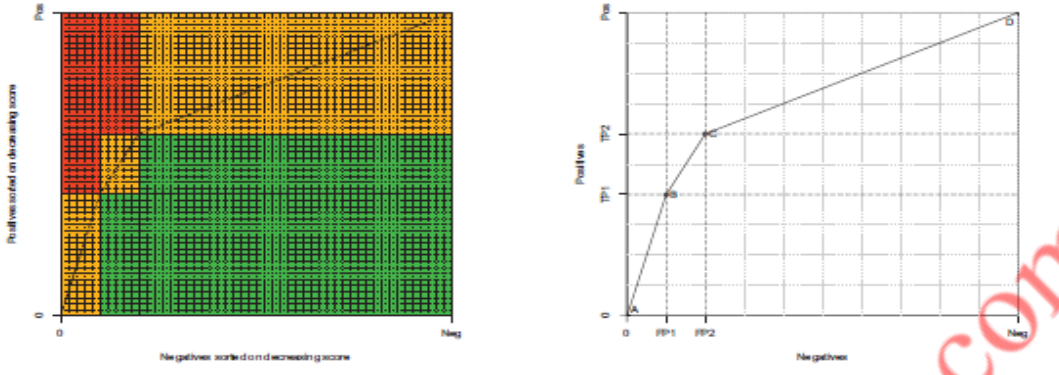


Figure 2.7. (left) Each cell in the grid denotes a unique pair of one positive and one negative example: the green cells indicate pairs that are correctly ranked by the classifier, the red cells represent ranking errors, and the orange cells are half-errors due to ties. **(right)** The coverage curve of a tree-based scoring classifier has one line segment for each leaf of the tree, and one (*FP*, *TP*) pair for each possible threshold on the score.

Class probability estimation

A **probabilistic** classifier assigns the **probabilities** to each **class** , where the **probability** of a particular **class** corresponds to the **probability** of the image belonging to that **class**. This is called **probability estimation**.

Turning rankers into class probability estimators

→ **Concavity** relates to the rate of change of a function's derivative. A function f is concave up (or upwards) where the derivative f' is increasing. This is equivalent to the derivative of f' , which is f′f, start superscript, prime, prime, end superscript, being positive.

-----XXX-----