SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601– Data Science and Analytics

III Year / VI Semester

Unit 3 – PREDICTIVE MODELING AND MACHINE LEARNING

Topic 7: Supervised Learning

Supervised Learning

Supervised Learning

Supervised ML algorithms is a type of ML technique that can be applied according to what was previously learned to get new data using labeled data and to predict future events or labels.

In supervised learning, we give it a set of training data, that the model learns from. It can then infer relationships between the features and the categories that we want, and apply that to unseen new values - and predict information about them.

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output..

Supervised Learning

Types of Supervised Learning Algorithms

- Linear Regression
- Logistic Regression
- Naive Bayes Classifiers
- Decision Trees
- Random Forest
- Support Vector Machine

Supervised Learning - Naive Bayes Classifiers

Naive Bayes Classifiers

The Bayes theorem is a mathematical formula for calculating conditional probability in probability and statistics.

In other words, it's used to figure out how likely an event is based on its proximity to another. Simply put, it is a way of calculating conditional probability.

The probability of A given B is equal to the probability of A times the probability of B given A over the probability of B.

The key insight is that the probability of something that depends on B depends very much on the base probability of B and A.

Supervised Learning - Naive Bayes Classifiers

Example of Spam Filtering
Let's just figure out the probability of an e-mail being spam given that it contains the word "free".

The probability of an email being spam given that you have the word "free" in that e-mail works out to the overall probability of it being a spam message times the probability of containing the word "free" given that it's spam over the probability overall of being free:

$$P(Spam \mid Free) = \frac{P(Spam)P(Free \mid Spam)}{P(Free)}$$

Supervised Learning – Decision Tree

Decision Trees
A decision tree employs a structure of test points (called nodes) and branches, which represent the decision being made. A node without further branches is called a leaf node. The leaf nodes return class labels.

Decision trees have two varieties: classification trees and regression trees.

Classification trees usually apply to output variables that are categorical—often binary—in nature, such as yes or no, purchase or not purchase, and so on.

Regression trees, on the other hand, can apply to output variables that are numeric or continuous, such as the predicted price of a consumer good or the likelihood a subscription will be purchased.

Supervised Learning – Decision Tree

Decision Tree Algorithms
Multiple algorithms exist to implement decision trees, and the methods of tree construction vary with different algorithms.

Some popular algorithms include

- ID3,
- C4.5, and
- CART

ID3

- ID3 (or Iterative Dichotomiser 3) is one of the first decision tree algorithms, and it was developed by John Ross Quinlan.
- Let A be a set of categorical input variables, P be the output variable (or the predicted class), and T be the training set.
- ID3 follows the rule — A branch with an entropy of zero is a leaf node and A branch with entropy more than zero needs further splitting.
- It uses Entropy and Information gain to select most informative attribute.

Supervised Learning – Decision Tree

C4.5

- The C4.5 algorithm introduces a number of improvements over the original ID3 algorithm.
- The C4.5 algorithm can handle missing data. If the training records contain unknown attribute values, the C4.5 evaluates the gain for an attribute by considering only the records where the attribute is defined.
- Both categorical and continuous attributes are supported by C4.5.
- For the corresponding records of each partition, the gain is calculated, and the partition that maximizes the gain is chosen for the next split
- The ID3 algorithm may construct a deep and complex tree, which would cause overfitting.

Supervised Learning – Decision Tree

CART

- CART (or Classification And Regression Trees) is often used as a generic acronym for the decision tree, although it is a specific implementation.
- Similar to C4.5, CART can handle continuous attributes.
- Whereas C4.5 uses entropy based criteria to rank tests, CART uses the Gini diversity index defined.
- Whereas C4.5 employs stopping rules, CART constructs a sequence of subtrees, uses cross-validation to estimate the misclassification cost of each subtree, and chooses the one with the lowest cost.

THANK YOU