Patrick Trinkle

CMSC676 - Spring 2009

Term Paper

Unsupervised Document Classification Outline

General Definition: Purpose of science is to automatically place a document in a specific category based on its contents. Also used for information filtering.

Why should we care? Because they help reduce time required for query processing, can be used to determine if a document is spam or ham; and also to automatically organize data sets too large for a human.

How is it presently done? Naive Bayes, SVMs, Neural Networks, Centroid-based techniques, k-nearest neighbor, etc

What is the next logical step? Other applications of the technology and ideas? No idea. Possibly a more complicated centroid based bayes theorem that is less naive.

### SVM - Support Vector Machine

"Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring datapoints of both classes, since in general the larger the margin the lower the generalization error of the classifier."-Wikipedia.org

### Rocchio's algorithm

Nothing found.

### Nearest Neighbor

"Nearest neighbor search (NNS), also known as proximity search, similarity search or closest point search, is an optimization problem for finding closest points in metric spaces. The problem is: given a set S of points in a metric space M and a query point q ∈ M, find the closest point in S to q. In many cases, M is taken to be d-dimensional Euclidean space and distance is measured by Euclidean distance or Manhattan distance."-Wikipedia.org

### Naive Bayes

"A naive Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even though these features depend on the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple."-Wikipedia.org

### Latent Semantic Analysis

"Latent semantic analysis (LSA) is a technique in natural language processing, in particular in vectorial semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms."-Wikipedia.org

### Neural Network

"Artificial neural networks are made up of interconnecting artificial neurons (programming constructs that mimic the properties of biological neurons). Artificial neural networks may either be used to gain an understanding of biological neural networks, or for solving artificial intelligence problems without necessarily creating a model of a real biological system. The real, biological nervous system is highly complex and includes some features that may seem superfluous based on an understanding of artificial networks."-Wikipedia.org

### Centroids

"In geometry, the centroid, geometric center, or barycenter of a plane figure X is the intersection of all straight lines that divide X into two parts of equal moment about the line. Informally, it is the "average" of all points of X. The definition extends to any object X in n-dimensional space: its centroid is the intersection of all hyperplanes that divide X into two parts of equal moment."-Wikipedia.org

Centroid based techniques are very fast because they are simple to calculations.