Lesson 3

Text Classification Using KNN and Naïve Bayes Classifier Supervised Machine Learning

Supervised Learning Method

- Supervised learning Algorithm exploits the training data (where zero or more categories) to model a classifier
- Classifies new text documents and labels each document
- Classification results from labeled documents and additional knowledge from experts

Classification

- Using Categories to which Document Labels
- Considered as a positive example for all categories with which it is labeled
- Negative example to all others

Training Algorithm Task

- Find a weight vector which best classifies new text documents

Different approaches for Supervised Learning

(i) K-Nearest Neighbour Method

(ii) Support Vector Machine

(iii) Naïve Bayes Method

(iv) Decision Tree

(v) Decision Rule

K-Nearest Neighbour Method

- Assumes that close-by objects are more probable in the same category
- Finds k objects in the large number of text documents, which have most similar query responses

K-Nearest Neighbour Method

- Predictions are based on a method to predict new (not observed earlier) text data
- Predictions by (i) majority vote method (for classification tasks) and (ii) averaging (for regression) method over a set of K-nearest examples

Naïve-Bayes Classifier

- Parallel algorithm
- Widely preferred text analytics
- Medium to Large Datasets 1 M to 100 M training examples which take too long time on SGD (Sequential, online incremental execution) or SVM (Sequential execution)
- Uses Posteriori Probability

Meaning of posteriori

- Posteriori means relating to or involving inductive reasoning from particular facts or effects to a general principle
- Posteriori example─ Combined conditional probability relates to individual condition probability using inductive reasoning

Meaning of posterior probability in statistics

- The probability assigned to some parameter or to an event on the basis of its observed frequency in a sample, and calculated from a prior probability by Bayes theorem

Bayes Classification Assumption

- Naïve independence assumptions (conditional independence)
- The classifier computes condition probabilities for the conditional independence

Naïve Bayes Classifier

- Naïve means unsophisticated, ..., a simple classifier
- Probabilistic and statistical classifier
- Based on Bayes theorem (from Bayesian statistics) with assumption of strong (Naïve) independence and maximum posteriori (MAP) hypothesis

Naïve Bayes Classifier

- A supervised learning technique, which uses non-parametric approach
- Uses assumption that features have strong independences
- “maximum a posteriori (MAP)” used to obtain the most likely class (Posteriori means at the back of something, for example, hypothesis)

Document classification in Text Analytics

- Use the bag-of-words model
- The pre-processing of a document first provides a document with a bag of words
- The occurrence (frequency) of each word as a feature used for training a classifier [Refer Section 9.2.2 Example 9.3]

Bayes Classification

- Probability that a bag-of-words x belong to k^{th} class equals the product of individual probabilities of those words.

$$P(\mathbf{x}|c_k) = \prod_{i=1}^{n} P(x_i|c_k)$$, where x_i is a discrete random variable (word), $i = 1, 2, \ldots, n$, when n is number of words in the bag.

Meaning of Symbols

- Π is sign for the product of n terms. P(xi|ck) means probability of condition that state the value = x_i and of c = c_k

Naïve Bayes Analysis

- Example 9.3 for "maximum a posteriori (MAP)" used to obtain the most likely class and take a decision

Naïve Bayes Classifier

- Requires a small amount of training data to estimate the parameters
- Not sensitive to irrelevant features as well

Naïve Bayes Classifier Applications

- Document categorization
- Language detection
- Authorship identification, age/gender identification
- Sentiment detection
- Email spam detection
- Personal email sorting

Summary

We learnt:

- Training data used to learn by a classifier, which classifies new text documents and labels each document
- KNN method– close-by objects are more probable in the same category, Finds k objects in the large number of text documents, which have most similar query responses

Summary

We learnt:

- Naïve Bayes Classifier
- Use the concept that probability that a bag-of-words x belong to k^{th} class equals the product of individual probabilities of those words

End of Lesson 3 on Text Classification Using KNN and Naïve Bayes Classifier Supervised Machine Learning