CS 486/686 Introduction to Artificial Intelligence—Fall 2016

Assignment 3 Topic Classification Using Machine Learning Classifiers

GLOSSARY OF TERMS

(from Witten et al. Data mining: Practical machine learning tools and techniques, 4th ed, 2016)

Document Classification: An important domain for machine learning is document classification, in which each data instance represents a document, and the instance's *class* is the document's *topic*. In our case, documents are news articles, and the classes are "ship", "wheat", "corn", "grain", etc. Documents are characterized by the words that appear in them—a document can be viewed as a *"bag of words"*, i.e., a set that contains all the words in the document, with multiple occurrences of a word appearing multiple times.

Vector Space Representation: By "vector space representation", or "vector space model", we mean that a document is converted into a vector of elements where each element represents the "weight" of a particular term (typically a word) in the document. A term may also be a "keyword", e.g., "investment" for a financial document, or even a phrase. If a certain term does not appear in a given document then a value of **zero** will be assigned for that term's entry in the vector. The *number of terms* in the vector is its *dimensionality*. If words are used as terms then the dimensionality of the vector will be the total number of words in the vocabulary of the document set. Typically therefore in text categorization problems the vectors involved will have *very high dimensionality* (thousands of terms) but can be very *sparse* (i.e., contain many zero values).

Filtering—for feature (attribute) selection:

Most machine learning algorithms are designed to learn which are the most appropriate *attributes* to use for making their decisions. For example, decision tree methods choose the most promising attribute to split on at each point and should—in theory—never select irrelevant or unhelpful attributes. Having more features should—in theory—result in more discriminating power, never less. [However] adding irrelevant or distracting attributes to a dataset often confuses machine learning systems. …
Naïve Bayes robustly ignores irrelevant attributes. It assumes by design that all attributes are independent of one another…But through this very same assumption, Naïve Bayes pays a heavy price in other ways because its operation is damaged by adding redundant attributes. …
…Because of the effect of irrelevant attributes on most machine learning schemes, it is common to precede learning with an attribute selection stage that strives to eliminate all but the most relevant attributes. [With the *filter* method] the attribute set is filtered to produce the most promising subset before learning commences.

StringToWordVector filter:
Converts a document to a vector space representation—uses the attributes below (and others) on the data instances (set of documents) to build a "dictionary" of the terms (words) and their occurrences that the classifier will work from.

StringToWordVector filter attributes:

tokenizer: Conversion into words—*"tokenization"*—is not as simple as it sounds. Tokens may be formed from contiguous alphabetic sequences with non-alphabetic characters discarded. An alphanumeric sequence may be regarded as a single token. Whitespace, punctuation, etc. make tokenization problematic.

stemmer: A stemming algorithm may be used to remove a word's prefixes and suffixes to reduce it to its root form, e.g., "runs"→ "run", "running"→run. Related words should be mapped to the same root.

lowerCaseTokens, useStopList: All words may be converted to lower case before being added to the dictionary. Words on a fixed, pre-determined list of function words or *"stopwords"*—such as *the*, *and*, and *but*—could be ignored.

minTermFreq: Low frequency words may be discarded. Sometimes it is found beneficial to keep the most *k* frequent words after the stopwords have been removed, or the top *k* words for each class.

TFTransform, IDFTransform: A measure of word (i.e., term) frequency that is widely used in information retrieval is **TF-IDF**: "term frequency times inverse document frequency". Here, the term frequency is modulated by a factor that depends on how commonly the word is used in other documents. The idea is that a document is basically characterized by the words that appear often in it, *except* that words used in every or almost every document are useless as features for classifying the topic of the document.