DATA ANALYTICS

Unit 4:Text Classification and Clustering

Jyothi R.
Department of Computer Science and
Engineering

Introduction to Text Classification

- The problem of text classification is closely related to that of classification of records with set-valued features.
- However, this model assumes that only information about the presence or absence of words is used in a document.
- In reality, the frequency of words also plays a helpful role in the classification process, and the typical domain-size of text data (the entire lexicon size) is much greater than a typical set-valued classification problem.

Introduction to Text Classification

- The problem of text classification finds applications in a wide variety
- of domains in text mining.
- Some examples of domains in which text classification is commonly used are as follows:
- **News filtering and Organization:** Most of the news services today are electronic in nature in which a large volume of news articles are created very single day by the organizations.
- In such cases, it is difficult to organize the news articles manually.
- Therefore, automated methods can be very useful for news categorization in a variety of web portals.
- This application is also referred to as text filtering.

Introduction to Text Classification

- **Document Organization and Retrieval:** The above application is generally useful for many applications beyond news filtering and organization.
- A variety of supervised methods may be used for document organization in many domains.
- These include large digital libraries of documents, web collections, scientific literature or even social feeds.
- Hierarchically organized document collections can be particularly useful for browsing and retrieval.

Introduction to Text Classification

- **Opinion Mining:** Customer reviews or opinions are often short text documents which can be mined to determine useful information from the review.
- Details on how classification can be used in order to perform opinion mining.

- **Email Classification and Spam Filtering:** It is often desirable
- to classify email in order to determine either the subject or to determine junk email in an automated way.
- This is also referred to as spam filtering or email filtering.

DATA ANALYTICS

Neighborhood-Based Collaborative Filtering

- A wide variety of techniques have been designed for text classification.
- Note that these classes of techniques also generally exist for other data domains such as quantitative or categorical data.
- Since text may be modeled as quantitative data with frequencies on the word attributes, it is possible to use most of the methods for quantitative data directly on text.
- However, text is a particular kind of data in which the word attributes are sparse, and high dimensional, with low frequencies on most of the words.
- Therefore, it is critical to design classification methods which effectively account for these characteristics of text.

Methods Used for Text Classification

- Some key methods, which are commonly used for text classification are as follows:
- **Decision Trees:** Decision trees are designed with the use of a hierarchical division of the underlying data space with the use of different text features.
- The hierarchical division of the data space is designed in order to create class partitions which are more skewed in terms of their class distribution.
- For a given text instance, we determine the partition that it is most likely to belong to, and use it for the purposes of classification.

Methods Used for Text Classification

- Some key methods, which are commonly used for text classification are as follows:
- **Decision Trees:** Decision trees are designed with the use of a hierarchical division of the underlying data space with the use of different text features.
- The hierarchical division of the data space is designed in order to create class partitions which are more skewed in terms of their class distribution.
- For a given text instance, we determine the partition that it is most likely to belong to, and use it for the purposes of classification.

Methods Used for Text Classification

- Some key methods, which are commonly used for text classification are as follows:
- **Decision Trees:** Decision trees are designed with the use of a hierarchical division of the underlying data space with the use of different text features.
- The hierarchical division of the data space is designed in order to create class partitions which are more skewed in terms of their class distribution.
- For a given text instance, we determine the partition that it is most likely to belong to, and use it for the purposes of classification.

Methods Used for Text Classification

- **Pattern (Rule)-based Classifiers**: In rule-based classifiers we determine the word patterns which are most likely to be related to the different classes.
- We construct a set of rules, in which the left-hand side corresponds to a word pattern, and the right-hand side
- corresponds to a class label. These rules are used for the purposes of classification.
- **SVM Classifiers:** SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes.
- The key in such classifiers is to determine the optimal boundaries between the different classes and use them for the purposes of classification.

Methods Used for Text Classification

- **Neural Network Classifiers:** Neural networks are used in a wide variety of domains for the purposes of classification.
- In the context of text data, the main difference for neural network classifiers is to adapt these classifiers with the use of word features.
- We note that neural network classifiers are related to SVM classifiers; indeed, they both are in the category of discriminative classifiers, which are in contrast with the generative classifiers.

DATA ANALYTICS

Methods Used for Text Classification

- Bayesian (Generative) Classifiers: In Bayesian classifiers (also called generative classifiers), we attempt to build a probabilistic classifier based on modeling the underlying word features in different classes.
- The idea is then to classify text based on the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents.
- Other Classifiers: Almost all classifiers can be adapted to the case of text data.
- Some of the other classifiers include nearest neighbor classifiers, and genetic algorithm-based classifiers.
- We will discuss some of these different classifiers in some detail and their use for the case of text data.

DATA ANALYTICS

Feature Selection for Text Classification

- Before any classification task, one of the most fundamental tasks that needs to be accomplished is that of document representation and feature selection.
- While feature selection is also desirable in other classification tasks, it is especially important in text classification due to the high dimensionality of text features and the existence of irrelevant (noisy) features.
- In general, text can be represented in two separate ways.
- The first is as a bag of words, in which a document is represented as a set of words, together with their associated frequency in the document.

Feature Selection for Text Classification

- Such a representation is essentially independent of the sequence of words in the collection.
- The second method is to represent text directly as strings, in which each document is a sequence of words.
- Most text classification methods use the bag-of-words representation because of its simplicity for classification purposes.
- In this section, we will discuss some of the methods which are used for feature selection in text classification.

Feature Selection for Text Classification

- The most common feature selection which is used in both supervised and unsupervised applications is that of stop-word removal and stemming.
- In stop-word removal, we determine the common words in the documents which are not specific or discriminatory to the different classes.
- In stemming, different forms of the same word are consolidated into a single word.
- For example, singular, plural and different tenses are consolidated into a single word.

Feature Selection for Text Classification

- We note that these methods are not specific to the case of the classification problem, and
- Are often used in a variety of unsupervised applications such as clustering and indexing.
- In the case of the classification problem, it makes sense to supervise the feature selection process with the use of the class labels.
- This kind of selection process ensures that those features which are highly skewed towards the presence of a particular class label are picked for the learning process.

DATA ANALYTICS

Feature selection methods

1. Gini Index
2. Information Gain
3. Mutual Information
4. $\chi 2$-Statistic

DATA ANALYTICS

Interaction of Feature Selection with Classification

- Since the classification and feature selection processes are dependent upon one another, it is interesting to test how the feature selection process interacts with the underlying classification algorithms.
- In this context, two questions are relevant:
- Can the feature-specific insights obtained from the intermediate results of some of the classification algorithms be used for creating feature selection methods that can be used more generally by other classification algorithms?
- Do the different feature selection methods work better or worse with different kinds of classifiers?

Interaction of Feature Selection with Classification

- In regard to the first question, it was shown in that feature selection which was derived from linear classifiers, provided very effective results.
- In regard to the second question, it was shown in that the sophistication of the feature selection process itself was more important than the specific pairing between the feature selection process and the classifier.

DATA ANALYTICS

Interaction of Feature Selection with Classification

- Linear Classifiers are those for which the output of the linear predictor is defined to be

$$p = \overline{A} \cdot \overline{X} + b,$$

- where $\overline{X}$ = (x1 . . . xn) is the normalized document word frequency vector,

$\overline{A}$ = (a1 . . . an) is a vector of linear coefficients with the same dimensionality as the feature space, and b is a scalar.

Interaction of Feature Selection with Classification

- Both the basic neural network and basic SVM classifiers
- belong to this category.
- The idea here is that if the coefficient a is close to zero, then the corresponding feature does not have a significant effect on the classification process.
- On the other hand, since large absolute values of aj may significantly influence the classification process, such features should be selected for classification.

DATA ANALYTICS

Interaction of Feature Selection with Classification

- In the context of the SVM method, which attempts to determine linear planes of separation between the different
- classes, the vector A is essentially the normal vector to the corresponding plane of separation between the different classes.
- This intuitively explains the choice of selecting features with large values of |*aj* |.
- This class of feature selection methods was quite robust, and performed well even for classifiers such as the Naive Bayes method, which were unrelated to the linear classifiers from which these features were derived.

DATA ANALYTICS

Decision Tree Classifiers

- A decision tree is essentially a hierarchical decomposition of the
- (training) data space, in which a predicate or a condition on the attribute value is used in order to divide the data space hierarchically.
- In the context of text data, such predicates are typically conditions on the presence or absence of one or more words in the document.
- The division of the data space is performed recursively in the decision tree, until the leaf nodes contain a certain minimum number of records, or some conditions on class purity.

Decision Tree Classifiers

- The majority class label (or cost-weighted majority label) in the leaf node is used for the purposes of classification.
- For a given test instance, we apply the sequence of predicates at the nodes, in order to traverse a path of the tree in top-down fashion and determine the relevant leaf node.
- In order to further reduce the overfitting, some of the nodes may be pruned by holding out a part of the data, which are not used to construct the tree. The portion of the data which is held out is used in order to determine whether or not the constructed leaf node should be pruned or not.

DATA ANALYTICS

Rule-based Classifiers

- Decision trees are also generally related to rule-based classifiers.
- In rule-based classifiers, the data space is modeled with a set of rules, in which the left hand side is a condition on the underlying feature set, and the right hand side is the class label.
- The rule set is essentially the model which is generated from the training data.
- For a given test instance, we determine the set of rules for which the test instance satisfies the condition on the left hand side of the rule.
- We determine the predicted class label as a function of the class labels of the rules which are satisfied by the test instance.

DATA ANALYTICS

Probabilistic and Naive Bayes Classifiers

- Probabilistic classifiers are designed to use an implicit mixture model for generation of the underlying documents.
- This mixture model typically assumes that each class is a component of the mixture. Each mixture component is essentially a generative model, which provides the probability of sampling a particular term for that component or class.
- This is why this kind of classifiers are often also called generative classifier.

DATA ANALYTICS

Probabilistic and Naive Bayes Classifiers

- The naive Bayes classifier is perhaps the simplest and also the most commonly used generative classifiers.
- It models the distribution of the documents in each class using a probabilistic model with independence assumptions about the distributions of different terms.
- Two classes of models are commonly used for naive Bayes classification.
- Both models essentially compute the posterior probability of a class, based on the distribution of the words in the document.
- These models ignore the actual position of the words in the document, and work with the “bag of words” assumption.

Probabilistic and Naive Bayes Classifiers

- The major difference between these two models is the assumption in terms of taking (or not taking) word frequencies into account, and the corresponding approach for sampling the probability space:
- **Multivariate Bernoulli Model**: In this model, we use the presence or absence of words in a text document as features to represent a document.
- Thus, the frequencies of the words are not used for the modeling a document, and the word features in the text are assumed to be binary, with the two values indicating presence or absence of a word in text.
- Since the features to be modeled are binary, the model for documents in each class is a multivariate Bernoulli model.

Probabilistic and Naive Bayes Classifiers

- **Multinomial Model:** In this model, we capture the frequencies of terms in a document by representing a document with a bag of words.
- The documents in each class can then be modeled as samples drawn from a multinomial word distribution.
- As a result, the conditional probability of a document given a class is simply a product of the probability of each observed word in the corresponding class.

Probabilistic and Naive Bayes Classifiers

- No matter how we model the documents in each class (be it a multivariate Bernoulli model or a multinomial model), the component class models (i.e., generative models for documents in each class) can be used in conjunction with the Bayes rule to compute the posterior probability of the class for a given document, and the class with the highest posterior probability can then be assigned to the document.

Mixture Modeling for Text Classification

- We note that the afore-mentioned Bayes methods simply assume that each component of the mixture corresponds to the documents belonging to a class.
- A more general interpretation is one in which the components of the mixture are created by a clustering process, and the class membership probabilities are modeled in terms of this mixture.
- Mixture modeling is typically used for unsupervised (probabilistic) clustering or topic modeling, though the use of clustering can also help in enhancing the effectiveness of probabilistic classifiers.

DATA ANALYTICS

Mixture Modeling for Text Classification

- These methods are particularly useful in cases where the amount of training data is limited.
- In particular, clustering can help in the following ways:
- The Bayes method implicitly estimates the word probabilities
- $P(t_i \in T | C^T = i)$ of a large number of terms in terms of their fractional presence in the corresponding component.
- This is clearly noisy.
- By treating the clusters as separate entities from the classes, we now only need to relate (a much smaller number of) cluster membership probabilities to class probabilities.
- This reduces the number of parameters and greatly improves classification accuracy.

DATA ANALYTICS

Mixture Modeling for Text Classification

- The use of clustering can help in incorporating unlabeled documents into the training data for classification.
- The premise is that unlabeled data is much more copiously available than labeled data, and when labeled data is sparse, it should be used in order to assist the classification process.
- While such unlabeled documents do not contain class-specific information, they do contain a lot of information about the clustering behavior of the underlying data.

Mixture Modeling for Text Classification

- This can be very useful for more robust modeling, when the amount of training data is low. This general approach is also referred to as co-training.
- The common characteristic of both the methods is that they both use a form of supervised clustering for the classification process.
- While the goal is quite similar (limited training data), the approach used for this purpose is quite different.

References

Text Book:

"Business Analytics, The Science of Data-Driven Making", U. Dinesh Kumar, Wiley 2017

"Recommender Systems, The text book, Charu C. Aggarwal, Springer 2016 Section 1.and Section 2.

DATA ANALYTICS

Image Courtesy

https://link.springer.com/chapter/10.1007/978-1-4614-3223-4_6

THANK YOU

Jyothi R.
Assistant Professor,
Department of Computer Science
jyothir@pes.edu