IITM-CS4011 : Principles of Machine Learning — Release date: 29 Oct, 2017, 11:55 am

Programming Assignment #3 — Deadline: 10 Nov, 2017, 11:55 pm

- The goal of this assignment is to play around with the plethora of clustering techniques you have learnt, and implement a Naive Bayes Classifier.
- This is an individual assignment. Collaborations are strictly prohibited (you know which kind). If you had to refer to any sources, make sure you cite them.
- You need to use `Weka` for this assignment.
- Turn in a detailed report containing the results of the experiments, the inferences you have drawn, the explanations you come by with. As usual, it should be typeset in LaTeX.
- Check Moodle regularly for updates regarding the assignment.
- Please read all the questions carefully. Don't miss any sub-parts (there are plenty).

The datasets are available here. Use your smail IDs for access.

1 Clustering

You have been provided with the following 8 2-dimensional datasets for clustering: Aggregation, Compound, Path-based, Spiral, D31, R15, Jain, Flames. First two columns are the features and the third column is the class label. In all your experiments, make sure that you are not giving the third column also as input to the clustering algorithm. You need to turn in the visualizations of your results for each question.

1. Convert all 8 datasets into ARFF format.
2. Visualize all 8 datasets. You need to turn in all your plots. Analyze each dataset by visualization and explain how these clustering algorithms will perform in these data (with reasons) : K-means, DBSCAN, hierarchical clustering with single link and complete link.
3. Run K-means with R15 dataset. Set $k = 8$. Report the cluster purity. Vary the value of k from 1 to 20 and study the effect of k on cluster purity. Plot a graph which explains your study.
4. Run DBSCAN with Jain dataset. Again report cluster purity. Study the effect of *minpoints* and *epsilon* on cluster purity.
5. Run DBSCAN and hierarchical clustering on Path-based, Spiral and Flames. Compare their performance on each dataset. For hierarchical clustering, you need to experiment with all types of linkages available in `Weka` to find the one that best suits the data.
6. Run K-means with D31 dataset. Can you recover all 31 clusters with $k = 32$? If not, can you recover all clusters by increasing the value of k? What happens when you apply DBSCAN? Apply hierarchical clustering with Ward's linkage. How does it perform?

2 Naive Bayesian Classifier

Design and implement a Bayesian Spam Filter that classifies email messages as either spam (unwanted) or ham (useful), that is, $y_i \in \{\texttt{spam, ham}\}$ using a Naive Bayes Classifier (explained later) for the following four scenarios:

1. Maximum Likelihood Estimation assuming likelihood $L \sim \text{Multinomial}(n_1, n_2, \cdots, n_k, N)$, where k is the size of the vocabulary, n_w is the number of times word w appears in the document d and $N = \sum_i n_i$
2. Maximum Likelihood Estimation assuming likelihood $L \sim$ Bernoulli (i, p), where p is the parameter of the Bernoulli Distribution and $i \in \{0, 1\}$. In our case, we have k Bernoulli Distributions.
3. Bayesian Parameter Estimation assuming that prior $p \sim \text{Dir}(\alpha_1, \cdots, \alpha_k)$, where $(\alpha_1, \alpha_2, \cdots, \alpha_k)$ are the parameters of the Dirichlet distribution.
4. Bayesian Parameter Estimation, assuming that prior $p \sim \text{Beta}(\alpha, \beta)$, where α and β are the parameters of the Beta distribution.

Theory

Naive Bayesian Classifier takes a text as input. The text is just a collection of words. It predicts the category of the input text. Naive Bayes algorithm for text classification involves two stages - training and classification. In the training stage, various probabilities are estimated based on the counts of training example features. In the classification stage, the estimated probabilities are used to evaluate the likelihood of each class for the input text. The text is then assigned a label with the highest likelihood score. Let $C = \{C_1, \cdots, C_m\}$ be the set of labels and $V = \{w_1, w_2, \cdots, w_n\}$ be all the words in the vocabulary. You would need to estimate the following propbabilities in the training stage:

- Class priors: $p(C_i)$ for $i = 1, 2, \cdots, m$. Note that $\sum_{i=1}^{m} p(C_i) = 1$.
- Within class word probabilities: $p(w_j|C_i)$ for $i = 1, 2, \cdots, m$ and $j = 1, 2, ..., n$. Note that $\sum_{j=1}^{n} p(w_j|C_i) = 1$ for $i = 1, 2, 3, \cdots, m$

Note - For some words, the conditional probability can end up becoming 0. To prevent this, one can perform "Add-one smoothing".

Description of the Dataset:

The data set contains a collection of spam and legitimate emails. Each token (word, number, punctuations, etc.) is replaced by a unique number throughout the dataset. In order to get the emails to an usable by Naive Bayes, some preprocessing is required. Each document should be represented by a term frequency vector of size k, where the i_{th} element is the frequency of the i_{th} term in the vocabulary (set of all distinct words in any text document from the training set). Do not consider the token "Subject:" as part of the vocabulary. Files whose names have the form spmsg*.txt are spam messages. Files whose names have the form *legit*.txt are legitimate examples.

Tasks

- You are required to implement Naive Bayesian classifier for 4 different scenarios as described previously. In the last two cases you are expected to try out different parameters for the prior distribution. Report the results after performing 5-fold cross validation (80-20 split). You have 10 folders in the dataset. Use 1-8 for training, 9-10 for testing in one fold. In the next fold, use 3-10 for training, 1-2 for testing and so on.
- Refer to Chapter 13 of Manning, Raghavan and Schutze for further reference on implementation of Naive Bayes for text classification.
- Comment on the impact of choice of parameters on the performance. Also comment on the performance of the classifier under different scenarios. Plot a PR-curve for each scenario and for different parameter setting in the last 2 scenarios.
- Your code should take $\texttt{trainMatrix}_{p\times k}$, $\texttt{trainLabel}_{p\times 1}$, $\texttt{testMatrix}_{r\times k}$ and $\texttt{testLabel}_{r\times 1}$ in the first two scenarios. p and r are number of documents in training and test set respectively and k is the size of the vocabulary. In the last 2 scenarios it should also take the parameters for the prior distribution as input. The code should output precision, recall, f1-measure for spam class and plot PR-Curve for the best model obtained in terms of performance.

Submission Instructions

Submit a single tarball/zip file containing the following files in the specified directory structure. Use the following naming convention: 'rollno_PA3.tar.gz'.

rollno_PA3

Dataset

spiral.arff

...

Report

rollno-report.pdf

Code

all your code files