Data Mining 2
Hands-on-Exercises

Dr. Shailesh Kumar

Google, Inc.

Instructions

- Use ANY Software you are comfortable with
  - R, MATLAB, FreeMat, SAS, etc.
- We will focus on TWO datasets:
  - Mushroom dataset
  - MNIST dataset
- We will focus on TWO areas
  - EXPLORATION – Data visualization and understanding
  - CLASSIFICATION – With various classification methods
- What you need to submit:
  - Create a document where you can just put pictures and tables and brief "observations" of what you learnt from each of those.
  - Each Slide explains step by step what to try, what to observe, and what to report in the final report.
  - Mention the slide title in each section of your report.

Mushroom – Exploration

- **Some of the values are missing.**
  - Substitute the most common value of that feature for missing.
- **Summarize the following for Mushroom data**:
  - Number of FEATURES:
  - Number of DATA POINTS:
  - Number of CLASSES:
  - PRIOR probability of each class:
  - Draw the histogram of for each feature:
- **Information Gain of each FEATURE**
  - Compute the information gain of each feature
  - List all the features in descending sort order of this value.

Mushroom – Naive Bayes classifier

- Partition the Mushroom dataset into 40% testing and 60% training data.
- Make sure you get proportional points from each class while sampling (or just do random sampling).
- Build a Naïve Bayes classifier using the training data
- Evaluate the Naïve Bayes classifier on test data.
- Now build the NB classifier with top k features based on the information gain (k = 5, k = 10, k = all)
- Report the test accuracy on each of these classifiers.

Mushroom – Decision Tree classifier

- Using the same training test split as before.
- Learn a Decision Tree Classifier using any of the tools.
- Build a HIGH complexity and a LOW complexity decision tree.
  - Complexity could be in terms of maximum depth or number of leaf nodes, etc. (depending on the parameters allowed in your tool).
  - Report accuracy on the high and low complexity decision trees.
  - Also define how you created the high and low complexity DT's.

Mushroom – Nearest Neighbor

- Define similarity between two data points as the fraction of features that match.
- Using the training test split built above, report the test accuracy with k = 1, 3, 5, and 7 nearest neighbor classifiers.

MNIST – Explore

- Fisher Projection of the entire data:
  - For the entire data, compute the Fisher projection.
  - Sample 50 points of class 3, 5, and 8.
  - Plot those in the top two Fisher projections.
  - Color code the points with their class label.
  - Repeat above exercise for classes 1, 7, and 9.
- PCA projection of the entire data:
  - Compute the top two PCA projections for the data.
  - Using the same 50 points of class 3, 5, 8 as above.
  - Plot these points in top two PCA projections.
  - Color code the points by their class labels.
  - Repeat the above exercise for classes 1, 7, and 9.

MNIST – Logistic Regression

- Partition each class data into 40% training and 60% testing
- Build the following classifiers for each PAIR of 10 classes
  - Logistic Regression model with all the 784 features.
  - Logistic Regression model with Top 9 PCA features.
  - Logistic Regression model with Top 9 Fisher features.
- Take the average of the 45 classifiers in each case.
- Report average accuracy w.r.t. different way of projection.

MNIST – K-Nearest Neighbor

- Sample 50 examples in each class as TRAINING data.
- Sample 50 examples in each class as TEST data.
- Build the k-Nearest neighbor classifier for k = 1, 3, 5, 7.
- Do this in Original Space, PCA(9) space, and Fisher(9) space.
- Compare the performance in the table with k values on one side and transformation on the other (no transformation, PCA(9) and Fisher(9)).
- NOTE: PCA(9) means top 9 PCA components.

MNIST – Bayesian Classifier

- Partition each class into 40% testing and 60% training.
- Build a Bayesian classifier for each pair of classes
  - Assume single full covariance matrix Gaussian for each class.
  - Do this over 784 dimensional full data.
  - Do this over PCA(9) dimensions.
  - Do this over Fisher(9) dimensions.
- Take average of 45 classifiers for each method.
- Report accuracies for the three feature projection methods.