Lab 3

Design of the Lab

The goal of this lab is to understand following concepts:

1. Feature extraction.
2. Feature Normalization
3. Dimensionality reduction techniques
4. Metrics to measure the accuracy of a classifier

1 Experiment 1

1.1 Feature Selection

Let us take the IRIS dataset for this task. The dataset has 3 classes and 4 features. For this experiment we will use only two classes out of three.

- Let X be the two class IRIS Dataset. Let A be a 2×4 matrix as given below:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

  - Compute $X' = AX$
  - Plot this X' on 2D graph with class1 as red and class2 as blue.

- Repeat the above with the following alternates for A.

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} or A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} or A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Which of the above plots show best visualization for 2 class IRIS Dataset?
- Matrix A was selecting features to visualize. Which A will you choose for IRIS flower classification problem?

1.2 Covariance and Correlation

From the above experiment it is clear that all features are not equally important to perform tasks like classification. So, sometimes we may need to reduce the feature dimension to work with better set of features. In this exercise we will look at the simple maths that gives us a framework to identify the features in lower dimensional space.

- Take different X' obtained in previous question. Find U=X'[0] and V=X'[1]. Compute the value of covariance and correlation co-efficient using the formula:

$$Cov(U,V) = \frac{\sum_i (U_i - \bar{U})(V_i - \bar{V})}{n-1}$$

where 'n' is the number of samples.

$$r(U,V) = \frac{Cov(U,V)}{\sigma(U)\sigma(V)}$$

where, r(U,V) is correlation coefficient between feature vector U and V.

> Covariance is a measure of how much two random variables vary together. Its similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.

- Which X' will you choose such that there is maximum variance between two feature vector?

- For n-dimensional feature vector lets try to find out best features in lower dimensional space.

  1. Compute n-dimensional mean vector and store it in 'm'
  2. Compute the covariance matrix which takes into consideration all feature combinations using direct function np.cov
  3. Compute Eigenvalues and Eigenvectors of the covariance matrix. How many Eigenvalues and Eigenvectors are you getting?
  4. Visualize the eigenvector
  5. Verify Eigenvalue-Eigenvector calculation
  6. Can you sort the Eigenvector by Eigenvalues? What is the order of Eigenvalues that you are getting?
  7. Can you choose 2 eigenvectors corresponding to top 2 largest eigenvalues and store it in variable 'T'?
  8. Can you transform the matrix using the formula $X' = TX$
  9. Can you plot the new feature space X' ? What can you see?

2 Experiment 2 – Feature Normalization

In this lab we will use WINE dataset and explore the concept Scaling and Standardizing the data.

- Load the WINE dataset from here.The Wine dataset consists of 3 different classes where each row correspond to a particular wine sample. The class labels (1, 2, 3) are listed in the first column, and the columns 2-14 correspond to 13 different attributes (features):

- Can you extract the features Alcohol (percent/volumne) and Malic acid (g/l). Are these features in different scale?

- Plot a graph between Alcohol (percent/volumne) and Malic acid (g/l). Can you see some sparsity and non-symmetry in the dataset?

2.1 Min-Max scaling

To visualize the dataset clearly we do feature scaling. The simplest method is Min-Max scaling to map the features in the range of [0, 1] or [1, 1]. The formula for min-max scaling is below:

$$x_norm = \frac{x - x_min}{x_max - x_min}$$

Plot a graph between Alcohol (percent/volumne) and Malic acid (g/l). Can you see some difference in min-max plotting vs plotting raw data?

2.2 Z-Score Normalization

- Given x is the original data, μ is the mean of a particular feature and σ is the standard deviation scale the features in the range of [0, 1]. The formula for feature Normalization is:

$$x_{norm} = \frac{x - \mu}{\sigma}$$

Plot a graph between Alcohol (percent/volumne) and Malic acid (g/l). Can you see some difference in min-max plotting vs plotting raw data?

3 Case Study 1: Twitter Sentiment Analysis

We have Twitter Dataset. We have to convert given tweets into features which can be used for sentiment classification. Every tweet can be classified as having either a positive or negative sentiment.

- **Few Positive tweets:**
  - @Msdebramaye I heard about that contest! Congrats girl!!
  - UNC!!! NCAA Champs!! Franklin St.: I WAS THERE!! WILD AND CRAZY!!!!!! Nothing like it...EVER http://tinyurl.com/49955t3
- **Few Negative Tweets:**
  - no more taking Irish car bombs with strange Australian women who can drink like rockstars...my head hurts.
  - Just had some bloodwork done. My arm hurts

We have 100,000 tweets for training and 300,000 tweets for testing. The Ground truth is 1 for positive tweet and 0 for negative tweet. Let's try to make a sentiment Analyzer using this dataset.

1. Modify the tweets such that the irrelevant words and characters are removed. To this end apply the following preprocessing.

   Case Convert the tweets to lower case.

   URLs We don't intend to follow the (short) urls and determine the content of the site, so we can eliminate all of these URLs via regular expression matching or replace with generic word URL.

   Username We can eliminate "@username" via regex matching or replace it with generic word AT_USER

   #hashtag hash tags can give us some useful information, so replace them with the exact same word without the hash. E.g. #nike replaced with 'nike'.

   Whitespace Replace multiple whitespaces with a single whitespace.

   Stop words a, is, the, with etc. The full list of stop words can be found at Stop Word List. These words don't indicate any sentiment and can be removed.

   Repeated letters If you look at the tweets, sometimes people repeat letters to stress the emotion. E.g. hunggrryyy, huuuuuuungry for 'hungry'. We can look for 2 or more repetitive letters in words and replace them by 2 of the same.

   Punctuation Remove punctuation such as comma, single/double quote, question marks at the start and end of each word. E.g. beautiful!!!!!! replaced with beautiful

   non-alpha Words Remove all those words which don't start with an alphabet. E.g. 15th, 5.34am

After preprocessing you should have obtain as shown below:

Tweet	Preprocessed words
cici hey cici sweetheart! just wanted to let u know i luv u! oh! and will the mixtape drop soon? fantasy ride may 5th!!!!	'hey', 'cici', 'luv', 'mixtape', 'drop', 'soon', 'fantasy', 'ride'
just had some bloodwork done. my arm hurts	'bloodwork', 'arm', 'hurts'

2. Count the number of words corresponding to each tweet after preprocessing
3. Count the number of positive and negative words in each tweet using positive_words.txt file and negative_words.txt which contain positive words and negative words respectively?
4. Take the features as the number of positive words and number of negative words for each tweet and use them in a KNN Classifier. What is the accuracy obtained?

4 Case Study 2: Product Rating

We have Amazon Product review Dataset Please download Toys and games 5-core dataset. We have to convert given reviews into features which can be used for rating the product. Each review can have an integer rating value from 1 to 5 . Example:

- **Few Reviews with rating 5:**
  - I like the item pricing. My granddaughter wanted to mark on it but I wanted it just for the letters.
  - Bought one a few years ago for my daughter and she loves it, still using it today. For the holidays we bought one for our niece and she loved it too
- **Few Reviews with rating 1:**
  - no more taking Irish car bombs with strange Australian women who can drink like rockstars...my head hurts.

Let's try to make a sentiment Analyzer using this dataset which has 167,597 reviews. The Ground truth is in range of 1-5 from 1 being bad review to 5 being good review.

1. Modify the reviews such that the irrelevant words and characters are removed. To this end apply the preprocessing we did in the previous case study.
2. Use positive_words.txt file and negative_words.txt which contain positive words and negative words respectively. Count the number of positive and negative words in each review.
3. Take the features as the number of positive words and number of negative words for each review and use them in a 5 class KNN Classifier.