Winter Workshop on Data Science and Machine Learning
26th - 30th December, 2017

Task Sheet - Day 4

Supervised Learning: Sentiment Prediction

Definition:

Sentiment of a tweet can be positive, negative or neutral. Sentiment is obtained using an external API (textblob) and whatever that external API returns is assumed to be correct sentiment for a tweet.

DataSet:

Take the dataset prepared in the Day-2 tasks which comprise of CSV file in which each row comprise of following features belonging to a tweet. However, to do tasks below, extend the data set to include 'all' tweets posted by a username X (SrBachchan).

a. Length of tweet
b. Number of hashtags in tweet
c. Number of @ mentions in tweet
d. Likes received by the tweet
e. Retweets received by the tweet
f. Sentiment expressed in the tweet (refer TextBlob API)
g. Hour when tweet was posted, eg. If a tweet is posted at 7:35 pm, then hours = 19

Algorithms:

To perform experiments given below, assume following algorithms for classification.

a. K-Nearest Neighbors (KNN), take different values of K in experiments.
b. Random Forest
c. Decision Tree
d. Naive Bayes

Experiments:

Research Question: Can tweet sentiment be predicted using features as hour of day, tweet length, number of hashtags, number of @mentions, number of likes and number of retweets ?

Class or Dependent Variable: Sentiment (possible values positive, negative and neutral)
Independent Variables: tweet length, number of hashtags, number of @mentions, number of likes and number of retweets, hour of day.

Run all algorithms (a,b,c,d), perform 10-fold cross validation with 80-20 split between training and test data. For algorithm 'a', run for K = 1, 2, 3, 4, 5.

Find following evaluation parameters:-

a. Accuracy
b. Precision
c. Recall
d. F-score

(i) Compare the above evaluation parameters for all algorithms as bar-plot in single graph.
X-axis: Evaluation Parameter x for all algorithms (one bar for one algorithm)
Y-axis: Value of Evaluation Parameter (on scale of 0 to 100)
Note: For KNN, use that K value in above graph for which accuracy is highest.

(ii) Compare the evaluation parameters for different values of K in a line-plot, all K values in same graph.

Supervised Learning: Popularity Prediction

Definition:
Popularity of a tweet is a numeric value which is sum of number of likes and retweets received on a tweet.

Note: Above is a simple definition, in real world popularity may have other components as well. For instance, a tweet which spreads more farther in Twitter network is more popular and so on ... However, for now, we restrict ourselves to above simple definition. In any project that you do, such definitions have to be made beforehand with absolute clarity.

DataSet:
Use the same dataset as prepared for above task.

Algorithms:
As you can observe that based on our definition of popularity, this is a problem of numeric prediction. Assume linear dependent and perform linear regression using R.

Experiments:

Research Question: Is tweet popularity *linearly* related to other features namely hour of day, tweet length, number of hashtags, number of @mentions and tweet sentiment ?

=> If tweet popularity can be predicted with high accuracy using linear regression on above parameters, then we can say that it is linearly related.

Class or Dependent Variable: Popularity (likes + retweets)
Independent Variables: tweet length, number of hashtags, number of @mentions, hour of day.

Run linear regression, perform 10-fold cross validation with 80-20 split between training and test data.

Find following evaluation parameters:-

a. Accuracy
b. Precision
c. Recall
d. F-score

Compare the above evaluation parameters as bar-plot in single graph.
X-axis: Evaluation Parameter x for all algorithms (one bar for one algorithm)
Y-axis: Value of Evaluation Parameter (on scale of 0 to 100)

Feature Importance: Perform correlation between tweet popularity and all other features X, one by one and use only those features which are comparatively highly correlated in above experiment. Do you get better accuracies ?

Unsupervised Learning: Tweet Categorization

DataSet:
Use the same dataset as prepared for above task.

Algorithms:
K-means algorithm.

Experiments:

Research Question: Can tweets be *categorized* based on features namely tweet length, number of hashtags, number of @mentions, number of likes, number of retweets, tweet sentiment and hour of day ?

=> If tweets can be grouped into K clusters with reasonable Within Cluster Distance for some value of K, then we say that tweet categorization is possible with above features.

Run K-means algorithm for different values of K.

Plot within cluster distance for various values of K and find elbow point.