Tutorial - Machine Learning and Natural Language Processing (17B1NCI731)

Tutorial 1 (24-July to 29-July)

Topic: Introduction to machine learning

Each of the cases below, assume that you have appropriate dataset available for the algorithm

1. Consider that you are supposed to develop a text summarization system, which type of machine learning algorithm (Supervised/Unsupervised) will be suitable for this application and why?
2. A weather prediction system that predicts the weather as Summer/Winter/Rainy has to be designed. Which learning algorithm (Classification/Regression) will be suitable for this system? Justify
3. To develop a spam filtering system, which learning algorithm will be appropriate?
4. For a stock market prediction system which predicts the price of particular stock the next day, which learning algorithm will be suitable and is this a classification problem to be considered?

Tutorial 2 & 3 (31-July to 12-Aug)

Topic: Linear Algebra, Tokenization

1) Which (if any) of the following pairs that are stemmed to the same form by the Porter stemmer definitely shouldn't be conflated? Explain your answers.

   a. abandon/abandonment
   b. absorbency/absorbent
   c. marketing/markets
   d. university/universe
   e. volume/volumes

2) The following is part of the beginning step of the stemming algorithm developed by Porter (1980). Porter's algorithm consists of a set of rules, in the form of S1 -> S2, which means that if a word ends with the suffix S1, then S1 is replaced by S2. If S2 is empty, then S1 is deleted:

   IES -> I
   SS -> SS
   S ->

   In a grouped set of rules written beneath each other (as above), only one is applied, and this will be the one with the longest matching S1 for the given word.

   a. What is the purpose of including an identity rule such as SS -> SS?
   b. Given the above set of rules, what do the following words map to?

      PONIES, TIES, CARESS, *CIRCUS, CANARIES, BOSS*

c. What rules should you add to correctly stem the following?
CARESSES, PONY
d. The stemming for PONIES and PONY given above might seem strange. Does it have a deleterious effect on retrieval?

Tutorial 4 (21-Aug to 27-Aug)

Topic: POS tagging – Rule Based Tagger, Stochastic based Tagger

1) Consider the following annotated corpus for the POS tagger,
The/DT race/NN was/VBD awesome/JJ ./.
The/DT boy/NN raced/VBD to/TO the/DT school/NN ./
Modern/NNP people/NNS are/VBP busy/JJ in/IN the/DT rat/NN race/NN ./.
Assuming, the default probability to be 0.01, design a rule based tagger using the above corpus and test the tagger for the input sentence – "The race to the school was awesome".

2) Considering the above annotated corpus mentioned in Q1, design a stochastic based tagger and test the input sentence "The race to the school was awesome".
3) Compare the output of the results of the above two questions and discuss how to overcome the difficulties if any.
4) Install NLTK in your laptop and perform tokenization and tagging (rule based and stochastic based) using the brown corpus.

Tutorial 5 (28-Aug to 03-Sep)

Topic: POS tagging – HMM based POS Tagger

1) A researcher proposes to POS tag the sentences from right to left instead of from left to right. However, he/she uses a pre trained HMM which was trained using the usual left to right order. Will the researcher get a different set of tags than if what were to use the left to right order? First examine the question when bigram assumption is made. Then generalize to k-order markov model.
2) Considering the corpus given in the previous tutorial (Tutorial-4 Q1), design a hidden markov model based POS tagger and test the sentence "The race to the school was awesome".
3) Will HMM with n-gram assumption works better compared to a bigram one ?
4) Using NLTK, develop a python script for a Hidden Markov Model based tagger and test it with other POS Taggers.