Homework # 4

August 12, 2019

1 Random Walk

A 1-dimension random walk is defined as a successive movements, where at each step an object can either move forward ($+1$, of probability = 0.5) or backward (-1 of probability = 0.5).

(1) Simulate a random walk of 1000 step. Set the object at position 0 initially. At each step

- Draw a random number r from the uniform distribution [0, 1)
- Update uhe position x
  $x \leftarrow x + 1$ if $r > 0.5$
  $x \leftarrow x - 1$ if $r \leq 0.5$

(2) Simulate 2000 random walk trials and make a histogram of the position at 1000^{th} step, i.e. x^i_{1000}, $i = 1, 2, 3, ...2000$.

(3) Calculate the mean and variance of x_{1000} for the 2000 trials, does it make sense?

(4) Now, simulate 2000 random walk trials and make a histogram of the position at 3000^{th} step, i.e. x^i_{3000}, $i = 1, 2, 3, ...2000$.

(5) Calculate the mean and variance of x_{3000} for the 2000 trials, does it make sense?

(6) Explain your result using central limit theorem.

2 NLP on Medical Transcripts

Understand medical notes is a challenging NLP problem. Lots of good application can be made if a machine can read doctors' notes and interpret the underlying medical conditions and severity. In this excercise, you are presented a simple data of 5000 medical cases "**medicaltranscriptions.csv**". Each case has the trascript and the associated medical specialty. Please

(1) For the "transcription" of each individual, use "word_tokenize" function from nltk and convert the corpus into a list of words.

(2) Convert all the tokens in question-1 to a continuous vector, using the pretrained word to vector dictionary "PubMed-and-PMC-w2v.bin". You may download the data from `http://evexdb.org/pmresources/vec-space-models`

(3) Calculate the cosine-similarity of the following word pair: "zyrtec-allergra"; "coronary-heart"; "heart-liver". Do the simlilarity measures make sense to you?

(4) Convert each transcription to a vector representation by taking the average of the vectors created in question-2

(5) Build a classification model and use the transcipts to predict the medical specialty by considering of the followings:

- Find a way to convert each transcript into structured vector
- Use log-loss as your error measure

- Build a multi-class classification model using random forest
- Build a multi-class classification model using glm

(6) Compare the model performance use corss-validation methods. Which model would you recommend?