## Week 1 – Logistic Regression

1. When performing logistic regression on sentiment analysis, you represented each tweet as a vector of ones and zeros. However your model did not work well. Your training cost was reasonable, but your testing cost was just not acceptable. What could be a possible reason?
   1. The vector representations are sparse and therefore it is much harder for your model to learn anything that could generalise well to the test set.
2. What are examples of text preprocessing?
   1. Stemming.
   2. Lowercasing.
   3. Removing stopwords, punctuation, handles and URLs.
3. The sigmoid function is defined as . Which of the following is true?
   1. Large positive values of will make close to 1, and large negative values of will make close to -1.
4. The cost function for logistic regression is defined as

. Which is true about the cost function?

* 1. When , as approaches 0, the cost function approaches .
  2. When , as approaches 0, the cost function also approaches 0.

1. For what value of in the sigmoid function does .
2. When performing logistic regression for sentiment analysis, you have to:
   1. Perform data preprocessing.
   2. Create a dictionary that maps the word and the class that word is found in to the number of times that word is found in the class.
   3. For each tweet, create a positive feature with the sum of positive counts of each word in that tweet, and a negative feature with the sum of negative counts of each word in that tweet.
3. When training a logistic regression mode, what order are the operations performed in?
   1. Initialise parameters
   2. Classify/predict
   3. Get gradient
   4. Update
   5. Get loss
   6. Repeat.
4. Assuming we got the classification correct, where for some specific example . This means that . What has to hold?
   1. Our prediction, , for this specific training example is greater than
5. What is the purpose of gradient descent?
   1. It allows us to learn the parameters in logistic regression as to minimise the loss function .
   2. *grad\_theta* allows us to update the parameters by computing

.

1. What is a good metric that allows you to decide when to stop training/trying to get a good model?
   1. When your accuracy is good enough on the test set.
   2. When you plot the cost versus the number of iterations and see that your loss is converging (i.e. no longer changing a lot).

## Week 2 – Naïve Bayes

1. Assume that there are 2 happy people and 2 unhappy people in a room. Persons A and B are happy and persons C and D are unhappy. If you were to randomly pick a person from the room, what is the probability that the person is happy.
2. Assume that there are 2 happy people and 2 unhappy people in a room. Persons A and B are happy and persons C and D are unhappy. If a friend showed you the part of the room where the two happy people are, what is the probability that you choose person B?
3. Express the probability of a tweet being positive given that it contains the word happy in terms of the probability of a tweet containing the word happy given that it is positive.
4. Bayes rule is defined as:
5. Suppose that in your dataset, 25% of the positive tweets contain the word ‘happy’. You also know that a total of 13% of the tweets in your dataset contain the word 'happy', and that 40% of the total number of tweets are positive. You observe the tweet: ''happy to learn NLP'. What is the probability that this tweet is positive?
6. The log likelihood for a certain word ​ is defined as
   1. Positive numbers imply that the word is positive.
   2. Negative numbers imply that the word is negative.
7. The log likelihood, which is the log of the ratio between two probabilities, is bounded between:
   1. and
8. When implementing naive Bayes, in which order should the steps be implemented?
   1. Get or annotate a dataset with positive and negative tweets.
   2. Pre-process the tweets.
   3. Compute the frequency of each word within each class.
   4. Calculate and .
   5. Calculate
   6. Compute the logprior:
9. What is required to test a naïve Bayes model?
   1. , , ,
10. What is not an application of naïve Bayes?
    1. Numerical predictions.

## Week 3 – Vector Space Models

1. Given a corpus A, encoded as and corpus B encoded as , what is the Euclidean distance between the two corpora?
2. Given the previous problem, a user came up with a corpus C defined as and you want to recommend a similar document. Would you recommend corpus A or corpus B?

so document A.

1. What is true about Euclidean distance?
   1. When comparing similarity between two corpuses, it does not work well when the documents are of different sizes.
   2. It is the norm of the difference between two vectors.
2. What is the range of a cosine similarity score, in the case of information retrieval where the vectors are positive?
3. Given a corpus A, encoded as and corpus B encoded as , what is the cosine similarity score of the two corpora?
4. Using the following vectors and cosine similarity, identify what country Ankara is the capital of:

so Ankara is the capital of Turkey

1. PCA is used to:
   1. Reduce the dimension of your data.
   2. Visualise word vectors.
2. What is correct about PCA?
   1. Eigenvectors are uncorrelated features of your data.
   2. The eigenvalues tell you the amount of information retained by each feature.
   3. Computing the covariance matrix is critical.
3. In which order do you perform the PCA operations?
   1. Mean normalise your data
   2. Calculate the covariance matrix
   3. Perform singular value decomposition
   4. Calculate the dot product to project the data
   5. Calculate the percentage of retained variance.
4. Vector space models allow us to:
   1. Represent words and documents as vectors.
   2. Build useful applications including information extraction, machine translation and chatbots.
   3. Create representations that capture similar meaning.

## Week 4 – Hashing and Machine Translation

1. Assuming that the objective is to minimise the transformation of as similar to as possible, what would you optimise to get R?
   1. Minimise the distance between and .
2. When solving for , which of the following is true?
   1. Initialise R.
   2. Create a for-loop.
   3. Inside the for-loop compute the gradient then update the loss.]
3. The Frobenius norm of is:
4. Assuming , , , what is the gradient of ?
5. You are visiting a city in the US. If you search for friends that are living in the US, would you be able to determine your two closest friends around the world?
   1. No.
6. What is the purpose of using a function to hash vectors into values?
   1. To speed up the time it takes when comparing similar vectors.
   2. To not have to spend the time comparing vectors with other vectors that are completely different.
7. Given the following vectors, determine the true statement:

, , ,

* 1. and have the same sign.

1. If we define to be the number of planes and to be 1 or 0 depending on the sign of the dot product with plane , what is the equation used to calculate the hash for several planes?
2. How can you speed up the look for similar documents?
   1. Approximate nearest neighbours
   2. Locality sensitive hashing
3. Hash tables are useful because they:
   1. Allow us to divide vector space into regions.
   2. Speed up look-up.
   3. Can always be reproduced. (You will always hash the same vector to the same bucket with the same hash function.)