Homework 2

Total Points - 100

Due Date: September 22, 2021

1. You are given sample documents with corresponding classes. Documents are annotated as A, or B. These documents are splitted as training and testing set shown in the below table. Assign the most probable class to the test sentence given below using naïve bayes classification approach. Please mention each step clearly (i.e., prior probability, conditional probability, etc.). [**Points 25**]

Doc #	words	Class	
d1	Chinese Beijing Chinese	B	Training
d2	Chinese Chinese Shanghai	B	Training
d3	Tokyo Japan Chinese	A	Training
d4	Chinese Macao	B	Training
d5	**Chinese Chinese Chinese Tokyo Japan**	?	**Testing**

2. What is cross-validation? Would you please give an example and explain how cross-validation work? When to use cross-validation during an experiment? [**Points 15**]

3. Explain how gradient descent algorithm works? Please explain the effect of learning rate on the learning algorithm while updating parameters using the following equations.

$$\mathrm{w}^{\mathrm{t+1}} = \mathrm{w}^{\mathrm{t}} - \frac{d}{\mathrm{dw}} f(x, w)$$

Please show differences among different types of gradient descent – mini-batch, batch, and stochastic gradient descent. [**Points 15**]

4. You are given the following text documents. Please answer the following questions. [**Total Points 45**]

Text:

It is going to rain today.

Today I am not going outside.

NLP is an interesting topic.

NLP includes ML, DL topics too.

I am going to complete NLP homework, today.

i) Would you please calculate the TF-IDF vector for each of the tokens? Please show each step (i.e., tokenization, vocabulary, TF, IDF, etc.). [**points 15**]

ii) Please calculate term-term co-occurrence matrix with a given context window ± 3. [**points 15**]

iii) Please find the most similar words (a pair of words) from their vector representations for both TF-IDF and co-occurrence matrix based vector representation cases. To compute most similarity score, you may use cosine similarity score. Show details similarity calculation for both cases – (a) TF-IDF and (b) co-occurrences-based representations. [**points 15**]