Homework Two: Natural Language Processing

Quin'darius Lyles-Woods

September 22, 2021

1 Naïve Bayes Classification

Document	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

$$Count_B = 8 \tag{1}$$

$$Count_A = 3 \tag{2}$$

$$Probablitiy(Chinese|B) = \frac{5}{8} \tag{3}$$

$$Probablitiy(Chinese|A) = \frac{1}{3} \tag{4}$$

$$Probablitiy(Tokyo|B) = \frac{L(0) = 1}{8} \tag{5}$$

$$Probablitiy(Japen|B) = \frac{\mathcal{L}(0) = 1}{8} \tag{6}$$

2 Cross Validation

2.1 What is Cross Validation

Cross Validation is a type of statistical analysis on how well a model can predict future outcomes with the given dataset. Through various methods that can be used in cross validation the average performance of the methods will give insight to the models prediction performance.

2.2 How does Cross Validation Work

There are a couple different ways that Cross Validation can be done but the basis of all of these methods are:

- Splitting up the data into three different sets.
- Set on part to the **Past Set**, this will be the largest set.
- Set another part to the **Test Set**, this is the second to largest.
- Set the last smallest part to be the "**Future Set**".

2.3 When to use Cross Validation

Cross Validation is best used when the task is very interpolative in nature and it because a lot harder to extrapolate into the future with Cross Validation. Examples of good ones would be image voice and text classification. An example of a bad use case of cross validation would be with ballistic missile calculation. It would be really hard to teach the machine the laws of physics and we would want to actually hard code those values into such a model.

3 Gradient Descent and its Variations

4 Tokenization Matrices and Vectors

Text:
It is going to rain today.
Today I am not going outside.
NLP is an intersting topic.
NLP includes ML, DL topics too.
I am going to complete NLP homework, today.

4.1 Tokenization

Text:
It is going to rain today.

Today I am not going outside.
NLP is an intersting topic.
NLP includes ML, DL topics too.
I am going to complete NLP homework, today.

Tokens: 31

4.2 Vocabulary

Text:
It is going to rain today.
Today I am not going outside.
NLP is an intersting topic.
NLP includes ML, DL topics too.
I am going to complete NLP homework, today.

is	2
going	3
to	3
today	3
I	2
am	2
NLP	3
topic	2
It	1
rain	1
not	1
outside	1
an	1
intresting	1
includes	1
ML	1
DL	1
complete	1
homework	1

4.3 Term Frequency

Document	is	going	to	today	I	am	NLP	topic	It	rain	not	outside	an	interesting	includes
1	1	0	1	1	0	0	0	0	1	1	0	0	0	0	0
2															
3															
4															
5															

Document	ML	DL	complete	homework
0	0	0	0	0
1				
2				
3				
4				
5				

4.4 Inverse Document Frequency