Homework Two: Natural Language Processing

Quin'darius Lyles-Woods

September 23, 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

$$Count_B = 8$$

$$Count_A = 3$$

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

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

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

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

$$Probablitiy(Tokyo|A) = \frac{1}{3}$$

$$Probablitiy(Japen|A) = \frac{1}{3}$$

$$Probablitiy(B)Probablity(d5|B) = \frac{5*1*1}{8*8*8} = \frac{5}{8^3} = \frac{5}{512} = 0.0097$$

$$Probablitiy(A)Probablity(d5|A) = \frac{1*1*1}{3*3*3} = \frac{1}{3^3} = \frac{1}{27} = \underbrace{0.037}_{Chosen\ Set=A}$$

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

The gradient descent algorithm works to calculate the miniumum loss function of a given neural network in our context. You find the loss function, compare it to what you value should've been graph it out. If the tangent line is in the negative you shift the function by the **learning rate**.

In the context of the equation:

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

The w^{t+1} is the weight for the next iterative step in the gradient descent algorithm.

The w^t is the current weight of the loss function given to the gradient descent algorithm.

The $\frac{d}{dw} f(x, w)$ is the slope of the loss function relative to the local minimum.

3.0.1 Mini-Batch

You are taking **very small** steps along the parabola to get the local minimum.

3.0.2 Scholastic

You are taking **every** step along the parabola to get to the local minimum.

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

4.3 Term Frequency

Document	is	going	to	today	I	am	NLP	topic	It
1	1	1	1	1	0	0	0	0	1
2	0	1	0	1	1	1	0	0	0
3	1	0	0	0	0	0	1	1	0
4	0	0	1	0	0	0	1	1	0
5	0	1	1	1	1	1	1	0	0

Document	rain	not	outside	an	interesting	includes
1	1	0	0	0	0	0
2	0	1	1	0	0	0
3	0	0	0	1	1	0
4	0	0	0	0	0	1
5	0	0	0	0	0	0

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

4.4 Inverse Document Frequency

is : $\log \frac{2}{5}$	NLP : $\log \frac{3}{5}$	an : $\log \frac{1}{5}$
going : $\log \frac{3}{5}$	topic : $\log \frac{2}{5}$	intresting : $\log \frac{1}{5}$
to : $\log \frac{3}{5}$	It : $\log \frac{1}{5}$	includes : $\log \frac{1}{5}$
today : $\log \frac{3}{5}$	rain : $\log \frac{1}{5}$	ML : $\log \frac{1}{5}$
I : $\log \frac{2}{5}$	not : $\log \frac{1}{5}$	DL : $\log \frac{1}{5}$
am : $\log \frac{2}{5}$	outside : $\log \frac{1}{5}$	complete : $\log \frac{1}{5}$
		homework : $\log \frac{1}{5}$

4.5 Co-Occurrence Matrix ± 3

I have chosen to do it with bounds on the sentences but still use the entire vocabulary for the matrix. Probably could do just the vocabulary in the sentence in the future.

	It	is	going	to	rain	today
is	1	0	1	1	1	0
going	1	1	0	1	1	1
to	1	1	1	0	1	1
today	0	0	1	1	1	0
I	0	0	0	0	0	0
am	0	0	0	0	0	0
NLP	0	0	0	0	0	0
topic	0	0	0	0	0	0
It	0	0	0	0	0	0
rain	0	1	1	1	0	0
not	0	0	0	0	0	0
outside	0	0	0	0	0	0
an	0	0	0	0	0	0
interesting	0	0	0	0	0	0
includes	0	0	0	0	0	0
ML	0	0	0	0	0	0
DL	0	0	0	0	0	0
complete	0	0	0	0	0	0
homework	0	0	0	0	0	0

	Today	I	am	not	going	outside
is	0	0	0	0	0	0
going	0	1	1	1	0	1
to	0	0	0	0	0	0
today	0	0	0	0	0	0
I	1	0	1	1	1	0
am	1	1	0	1	1	1
NLP	0	0	0	0	0	0
topic	0	0	0	0	0	0
It	0	0	0	0	0	0
rain	0	0	0	0	0	0
not	1	1	1	0	1	1
outside	0	0	1	1	1	0
an	0	0	0	0	0	0
interesting	0	0	0	0	0	0
includes	0	0	0	0	0	0
ML	0	0	0	0	0	0
DL	0	0	0	0	0	0
complete	0	0	0	0	0	0
homework	0	0	0	0	0	0

	NLP	is	an	interesting	topic
is	1	0	1	1	1
going	0	0	0	0	0
to	0	0	0	0	0
today	0	0	0	0	0
I	0	0	0	0	0
am	0	0	0	0	0
NLP	0	0	0	0	0
topic	0	1	1	1	0
It	0	0	0	0	0
rain	0	0	0	0	0
not	0	0	0	0	0
outside	0	0	0	0	0
an	1	1	0	1	1
interesting	1	1	1	0	1
includes	0	0	0	0	0
ML	0	0	0	0	0
DL	0	0	0	0	0
complete	0	0	0	0	0
homework	0	0	0	0	0

	NLP	includes	ML	DL	topics	to
is	0	0	0	0	0	0
going	0	0	0	0	0	0
to	0	0	0	0	0	0
today	0	0	0	0	0	0
I	0	0	0	0	0	0
am	0	0	0	0	0	0
NLP	0	0	0	0	0	0
topic	0	0	0	0	0	0
It	0	0	0	0	0	0
rain	0	0	0	0	0	0
not	0	0	0	0	0	0
outside	0	0	0	0	0	0
an	0	0	0	0	0	0
interesting	0	0	0	0	0	0
includes	1	0	1	1	1	0
ML	1	1	0	1	1	1
DL	1	1	1	0	1	1
complete	0	0	0	0	0	0
homework	0	0	0	0	0	0

	I	am	going	to	complete	NLP	homework	today
is	0	0	0	0	0	0	0	0
going	1	1	0	1	1	1	0	0
to	1	1	1	0	1	1	1	0
today	0	0	0	0	1	1	1	0
I	0	0	0	0	0	0	0	0
am	1	0	1	1	1	0	0	0
NLP	0	0	1	1	1	0	1	1
topic	0	0	0	0	0	0	0	0
It	0	0	0	0	0	0	0	0
rain	0	0	0	0	0	0	0	0
not	0	0	0	0	0	0	0	0
outside	0	0	0	0	0	0	0	0
an	0	0	0	0	0	0	0	0
interesting	0	0	0	0	0	0	0	0
includes	0	0	0	0	0	0	0	0
ML	0	0	0	0	0	0	0	0
DL	0	0	0	0	0	0	0	0
complete	0	1	1	1	0	1	1	1
homework	0	0	0	1	1	1	0	1

4.6 Term Frequency-Inverse Document Frequency