NLP Questions on Language Modelling and Classification (weeks 2 and 3)

Joe Stacey

January 2024

1 Naive Bayes

1.1

State the independence assumption used in Naive Bayes.

1.2

For a Bag of Words model, provide an example of two features (words) where this is not an accurate assumption.

1.3

Consider the training corpus in Table 1. Create a Binary Naive Bayes model based only on the features 'good', 'great', 'bad' and 'awful'. Use this model to predict the class of the following review: "tom cruise did it again , so great it's a masterpiece , don't listen to the bad reviews , just enjoy this great film"

Training corpus (after some pre-processing)	Class
almost as good as the first top gun	+
fun throughout , definitely recommend this great film	+
awful , tom cruise had no depth once again . why is his acting so bad	-
better than i expected , not bad at all	+
lived up to the name top gun , what a great film	+
wish i had just rewatched the original , this one was nowhere near as good	-
tom cruise should do the stunts and leave the acting to someone else	-

Table 1: Film reviews, categorised as being either positive (+) or negative (-)

2 Logistic Regression

2.1

What is one limitation of a Naive Bayes model that logistic regression helps to overcome.

2.2

What is the difference between a generative and discriminative algorithm? Is Logistic Regression generative or discriminative? Provide your justification.

2.3

Write out the logistic function of an input x.

2.4

How many parameters do we have for a 3-class logistic regression model with 10 different Bag of Words input features?

3 Accuracy and F1-scores

3.1

State the definition of the F1-score for a single class, writing this in terms of precision and recall.

3.2

State the definition for micro-averaged F1, and show how this is equivalent to accuracy. Avoid making any assumptions about the number of classes.

3.3

How is macro-averaged F1 different from micro-averaged F1? What are the advantages of using the macro-averaged F1 score?

4 N-grams and perplexity

4.1

Write an expression for approximating $P(w_n|w_1^{n-1})$ using an N-gram model

	hamish	is	the	sweetest	cat
hamish	0.01	0.4	0.2	0	0
is	0.02	0	0.4	0.08	0.01
the	0.01	0	0	0.45	0.3
sweetest	0.01	0	0.01	0	0.5
cat	0	0.15	0.03	0	0

Table 2: The column on the left hand side are the words in the history, while the top row are the words being predicted. The values in the table are the model probabilities for the words in the top row, given the words in the left hand side column.

4.2

Decompose the joint probability of words $P(w_1, ..., w_n)$ into a product of conditional probabilities after applying a trigram assumption.

4.3

Given the probability table in Table 2, calculate the probability of the sentence "hamish is the sweetest cat", using a bi-gram model. You can use the information that $P(hamish| < s >) = 0.1$, and $P(< /s > |cat) = 0.1$.

4.4

Would you expect implementing Add-1 smoothing to increase or decrease this probability? Why?

4.5

What would the perplexity be for this example? Please include $P(< /s > |cat)$ in this calculation.

4.6

In general, provide an expression for the cross-entropy loss (using log base 2) in terms of perplexity.

4.7

Is perplexity an intrinsic or extrinsic measure of performance?

4.8

Why might it be preferable to evaluate the extrinsic performance of a language model when this is possible?

5 Feed-forward neural language models

5.1

Consider a feed-forward language model using 4 words of context, 200-dimensional word embeddings and a vocabulary of 20,000 words. Assume no bias terms in the model.

How many parameters are there in the model? Include parameters in the embedding layer.

5.2

Name two advantages that a feed-forward language model has over an N-gram model.

5.3

With a large training corpus, would you expect a feed-forward language model to outperform an n-gram model?

6 GRUs

6.1

Consider a GRU with:

- 200 dimensional word embeddings
- 100 dimensional hidden states
- 3 output logits (for a multi-class classification task)
- No bias terms in the network
- A maximum of 50 words per observation in the training data
- The GRU is bidirectional (and is a single layer GRU)

Excluding the parameters in the embedding layer, how many model parameters are there in total in the GRU? Please include any parameters in the output layer.

7 LSTMs

7.1

Provide the full equations used to define an LSTM (without including bias terms). Do not include an output layer.

7.2

State the dimensions for each matrix and vector mentioned in the equations above, where we have E dimensional word embeddings and H dimensional hidden states.

7.3

Explain why LSTMs can help to mitigate the vanishing gradient problem seen in vanilla-RNNs, mentioning the role of the forget gate.

7.4

One option for a sentence-level classification task is to apply a classifier to the final hidden state of an LSTM. Suggest two different strategies we could use instead of using the final LSTM hidden state.

8 True or False questions

8.1

Binary classification models can either have a single logit output (using a sigmoid), or they can have two logit outputs (using a softmax). Both methods are acceptable.

8.2

A non-linear activation function usually precedes softmax in a model.

8.3

Prior to transformer and RNN models, neural networks in NLP usually had more than 5 layers.

8.4

De-biasing methods usually have no impact on a model's performance on in-domain test sets.

8.5

Excluding the embedding layer, a single-layered bi-directional LSTM has exactly twice the number of parameters of a single-layered unidirectional LSTM.