Project 1

Zheyi Yi

Part 1.

This following graph shows that the change of perplexity with the increase of the size of training data by three different predictive methods: the maximum likelihood estimate, MAP estimate, predictive distribution for unigram model.

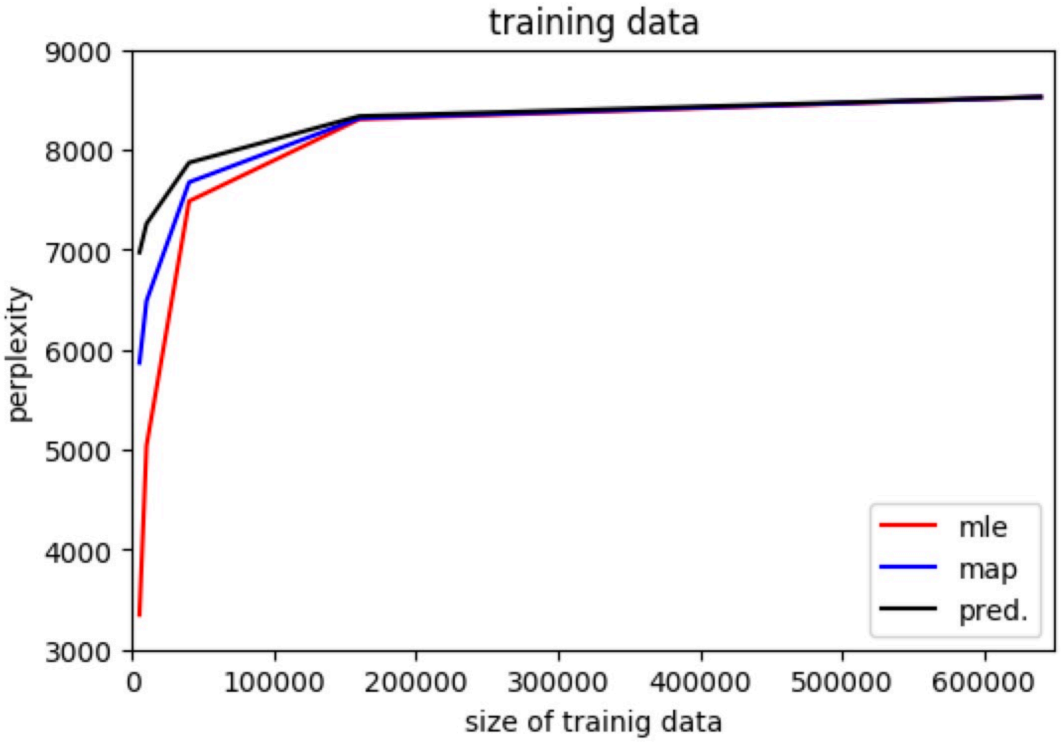

Figure 1

The Figure 2 shows that the change of perplexity with the increase of the size of training data by three different predictive methods: the maximum likelihood estimate, MAP estimate, predictive distribution

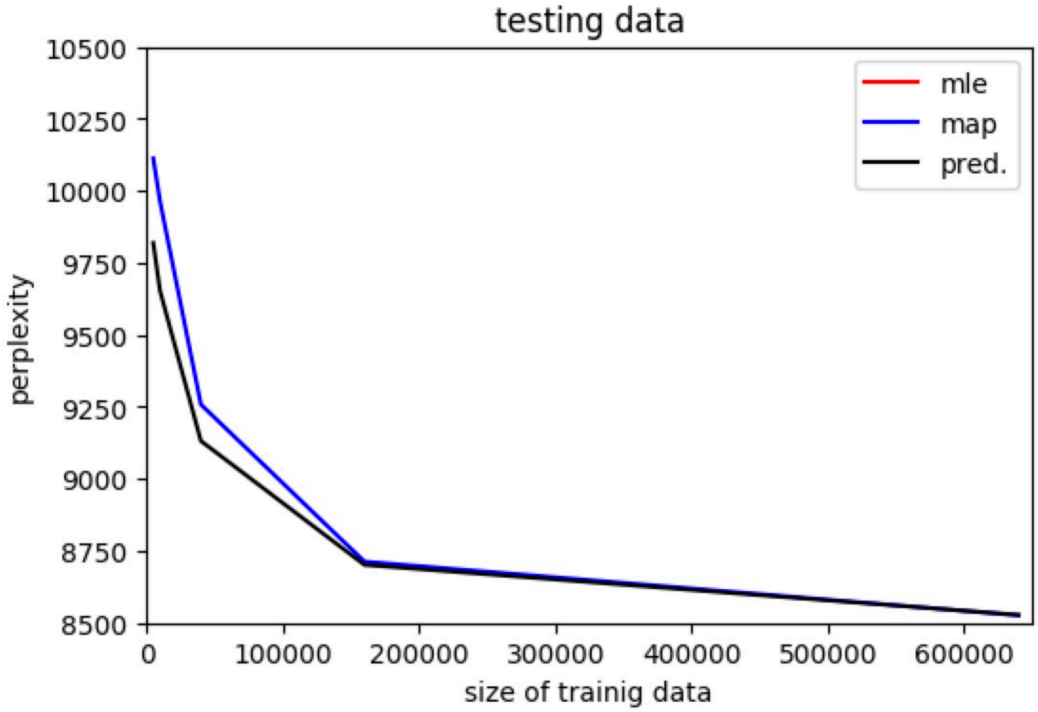

Figure 2

What happened to the test set perplexities of the different method with respect to each other as the training set size increase?

There are undefined values (infinity) for maximum likelihood estimate until the whole training set.

For map and predictive distribution, perplexity decrease as the training set increased. All models converge to roughly the same perplexity on the whole size training data. In conclusion, accuracy will increase as the training set size increase

What is the obvious shortcoming of the maximum likelihood estimate for a unigram model? How do the other two approaches address this issue?

The big shortcoming of maximum likelihood estimate is that from figure 2, we can see that perplexity is undefined until we train the

whole dataset. If a test word is not in the training set, so propability of this word will be 0, and the perplexity won't be defined because In(0) = negative infinity. So the shortcoming of maximum likelihood estimate is it cannot deal with unknown data which doesn't appear in training data.

A vector of counts alpha is introduced in other two methods. With the parameter, even the word in testing data doesn't appear in the training data, we still keep a small amount of probability on it, which avoid In(0), so we can calculate the perplexity.

For the full training set, how sensitive do you think the test set perplexity will be to small changes in alpha? Why?

I think it won't be sensitive for small changes in alpha. So compared the numbers of words, we just plus a small changing alpha on it, it won't have big change because compared the value of numbers of words, the small change of alpha only occupy a very small part.

Part 2.

The figure 3 shows that the change of perplexity of predictive distribution as a function of alpha on the N/128 training set.

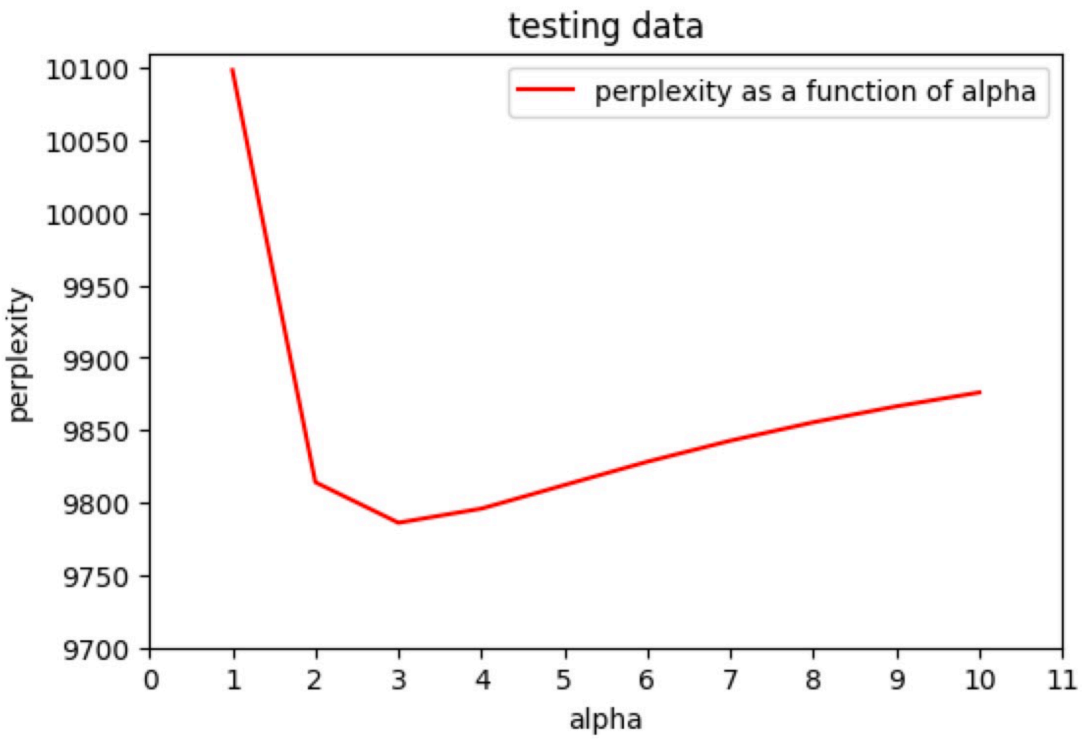

Figure 3

The figure 4 shows that the change of log evidence of predictive distribution as a function of alpha on the test set.

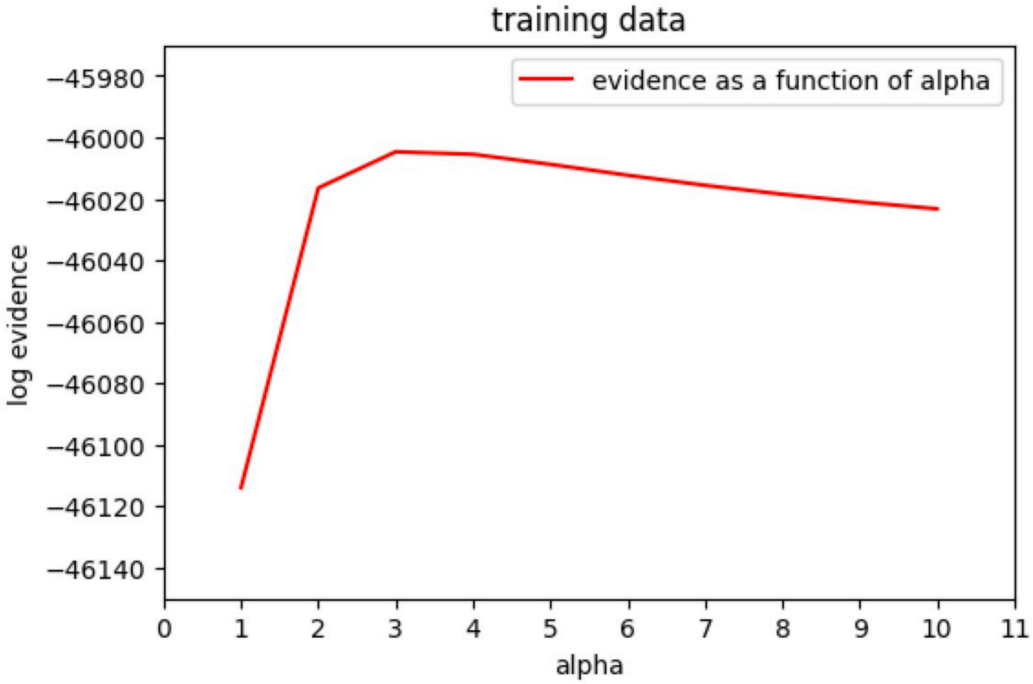

Figure 4

Is maximizing the evidence function a good method for model selection on this dataset?

Yes, it is good method. From figure 4, we see the top value is on the alpha = 3, and at the same time, the lowest value of perplexity is on the alpha =3 as well if we see figure 3. This is because evidence function is reverse proportional to model.

Part 3

We apply our unigram predictive distribution model to identify authors. We choose pg121 file and predictive distribution model with alpha = 2 to evaluate the perplexity on other two file pg141 and pg1400. We can classify them by selecting the text which has the lowest perplexity performance from the trained model. We got

4784.50 for pg141 and 6397.17 for pg1400, so we choose the same pg141 has the same author with pg121. It is successful because if texts written by the same author, the frequency of words should be the similar, and then got low perplexity. After remove the words which appear less than 50 times, the perplexity become 8761.38 for pg141, and 10231.86 for pg1400. It is still successful.
Removing small occurrence data has a little affect on the result of perplexity, we see perplexity increase, but it won't affect the final outcome because we just remove a small occurrence words.