Erin Sarlak

Collaboration Statement: I received help from Patrick and Kyle during the office hours.

Total hours spent: 10

Links: [CP1 instructions] [Course collaboration policy]

Contents

1a: Figure for Experiment 1: Test-set Log Likelihood vs Train Set Size

1a: Solution

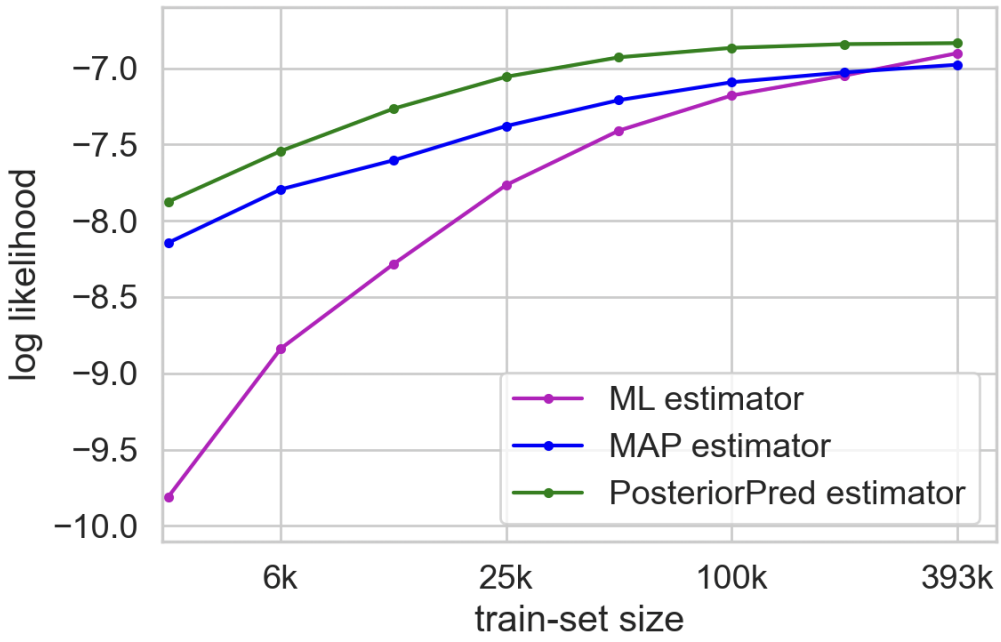

Figure 1: Performance of three estimators against varying train set size.

1b: Problem Statement

Will the ML estimator eventually be clearly superior to the others, given enough training data? Why or why not? (Hint: what happens as $N \to \infty$).

1b: Solution

Based on the pattern in the graph, it seems like given enough training data, when $N \to \infty$, ML estimator will become clearly superior to other estimator models, but this is not the case. Given enough training data, the prior terms in the formula for Map estimator and Posterior Predictive estimator will become insignificant, and both of those formula will converge to the ML estimator, $\frac{n_v}{N}$. Thus, ML estimator would not be superior to other and instead all estimator will yield about the same results.

1c: Problem Statement

What test-set log likelihood performance would you get if you simply predicted the next word using a uniform probability distribution over the vocabulary? Please report a concrete numerical value, comparable to the scores in Fig 1a above. Does the ML estimator always beat this "dumb" baseline?

1c: Solution

If we predict the next word using a uniform probability distribution over the vocabulary, then the log likelihood of each word will be assigned as the log of $\frac{1}{V}$. Since the vocabulary size is 15205, the log likelihood is $\ln \frac{1}{15205} = -9.6$. This value remains the same with over varying train set size.

The ML estimator does not always beat the 'dumb' baseline when the training set size is too small, around 2k based on the results. The reason is when the data set is too small, there are a lot of unseen words, which are assigned very small probability values (based on the epsilon), and the seen data therefore receive relatively high probability values. However, as the train set size increases, because there are less and less unseen words, each word is assigned a reasonable probability, and the MLEstimator surpasses the baseline.

2a: Figure for Experiment 2: Selecting a Value of Hyperparameter α

2a: Solution

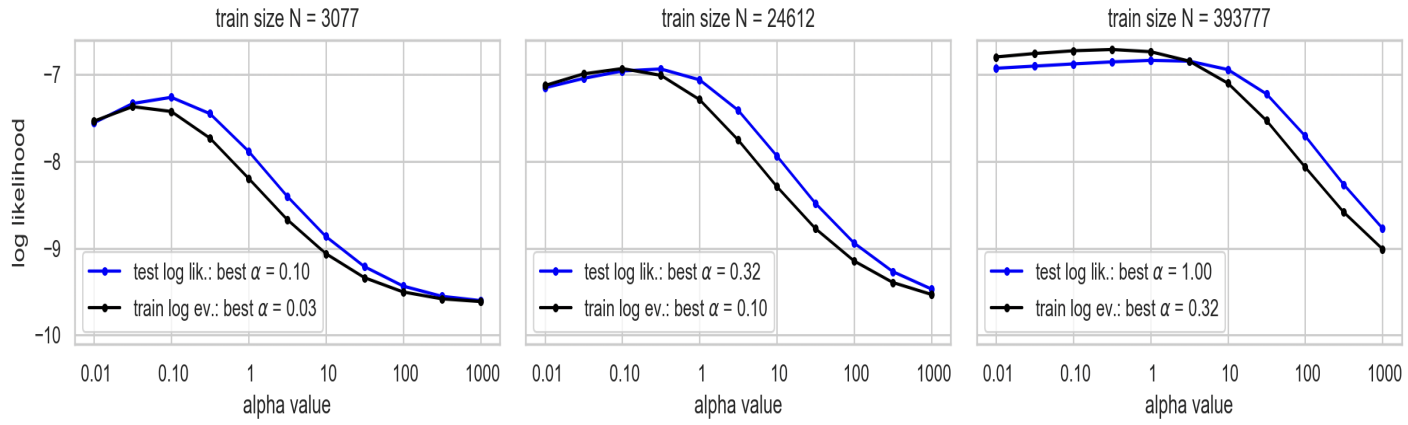

Figure 2: Performance variation of predictive posterior estimator against different alpha values (for three different train sets).

2b: Takeaway from Figure 2a

Across all 3 dataset sizes, compare the best α value as selected by the evidence to the α selected by looking directly at test-set likelihood. Do these peaks occur at "similar" locations, if we define similar to x as within (0.2x, 5x)? What does this suggest about the viability of using the evidence to select α that might predict well on future state-of-the-union speeches?

2b: Solution

Across all 3 dataset sizes, the highest test-set log likelihood occurs when the alpha value is between 0.10 and 1. For N = 3077, the estimator's performance peaks when alpha = 0.10. For N = 24612, the highest likelihood is when alpha is around 0.60. Lastly, for N = 393777, the ideal alpha is 1, but still performs almost equal well for

values to 0.10.
High log likelihood on a test set for an estimator is a sign of good performance. Since all three of the estimators perform the best for alpha values in between 0.10 and 1. I would select a value in this range for predicting the future state of the union speeches. As a note, I would prefer values closer to 1 in that range for an estimator trained on higher training set sizes.