Homework 2 Problem Set Solutions

Ankur Parikh

October 23, 2018

More Language Modeling (10 points):

Let n be the order of the language model i.e.

$$P(w_i|w_1, ..., w_{i-1}) \approx P(w_i|w_{i-n+1}, ..., w_{i-1}) = P(w_i|w_{i-n+1}^{i-1})$$

For this question, the term *inference time* refers to one call to the language model i.e. computing $P(w_i|w_{i-n+1}^{i-1})$ under the language model for one choice of w_{i-n+1}^{i-1}, w_i.

- Consider the particular neural language model (Mnih and Hinton 2007) below (where $\mathbf{e}_w \in \mathbb{R}^d$ is a trainable vector for the word w, $C_j \in \mathbb{R}^{d \times d}$ is a trainable transformation matrix, and $b_w \in \mathbb{R}$ is a scalar bias.

$$\mathbf{p}_i = \sum_{j=1}^{n-1} \mathbf{e}_{w_{i-j}} C_j$$

$$\nu_i(w) = \mathbf{p}_i^T \mathbf{e}_w + b_w$$

$$P(w_i|w_{i-n+1}^{i-1}) = \frac{\exp(\nu_i(w_i))}{\sum_{v \in \mathcal{V}} \exp(\nu_i(v))}$$

  What is the memory and inference complexity of the above neural language model in big-O notation?

Inference: The dominant cost for computing the numerator is in computing $\mathbf{p}_i$ which takes $O(nd^2)$. Computing the denominator then takes $O(Vd)$ so the total inference complexity is $O(nd^2+Vd)$ (where V is the size of the vocabulary).

Note that unlike in n-gram models the denominator cannot be precomputed.

Memory: Storing $\mathbf{e}_w$ for all w takes $O(Vd)$ and storing C_j for all j takes $O(nd^2)$. All other terms are negligible. Thus the total memory complexity is $O(Vd + nd^2)$.

Repeated Features in Naive Bayes / Logistic Regression (15 points): Consider a classification problem with 3 binary features x_1, x_2, x_3 where $x_2 = x_3$ (i.e. they are identical features) and a binary class label y. Ideally, the addition of repeated (identical) features should not affect the classification decision. We will study the effect of these repeated features on Naive Bayes and Logistic Regression.

Assume all models below are trained with Maximum Likelihood to convergence.

- Mathematically characterize the effect (or non-effect) of repeated features on Naive Bayes e.g. how A Naive Bayes model trained only on (x_1, x_2) compares to one trained on (x_1, x_2, x_3).

  We note that

  $$P(y, x_1, x_2, x_3) = P(y)P(x_1|y)P(x_2|y)P(x_3|y) \tag{1}$$

  This means that the log likelihood decomposes: $\log P(y, x_1, x_2, x_3) = \log P(y) + \log P(x_1|y) + \log P(x_2|y) + \log P(x_3|y)$ As a result, when optimizing the likelihood, these probabilities are estimated separately of one another i.e.

  $$\widehat{P}(x_i = a|y = b) = \frac{\text{count}(x_i = a, y = b)}{\text{count}(y = b)} \tag{2}$$

  Thus, repeated features will get magnified by Naive Bayes since the model will effectively become:

  $$P(y, x_1, x_2, x_3) = P(y)P(x_1|y)P(x_2|y)^2 \tag{3}$$

  This effectively overweights the $x_2 = x_3$ feature. This is one of the primary weaknesses of Naive Bayes.

- Mathematically characterize the effect (or non-effect) of repeated features on Logistic Regression (without any regularization) e.g. how a Logistic Regression model trained only on (x_1, x_2) compares to one trained on (x_1, x_2, x_3).

  Without regularization logistic regression will learn to ignore repeated features. To see this consider the conditional likelihood with x_3 present ($\mathbf{w}$ is the weight vector)

$$\begin{aligned} P(y = a|x_1, x_2, x_3, \mathbf{w}) &= \log \frac{\exp(w_{a,1} \cdot x_1 + w_{a,2} \cdot x_2 + w_{a,3} \cdot x_3)}{\sum_y (\exp(w_{y,1} \cdot x_1 + w_{y,2} \cdot x_2 + w_{y,3} \cdot x_3))} \\ &= \log \frac{\exp(w_{a,1} \cdot x_1 + (w_{a,2} + w_{a,3}) \cdot x_2)}{\sum_y (\exp(w_{y,1} \cdot x_1 + (w_{y,2} + w_{y,3}) \cdot x_2))} \\ &= \log \frac{\exp(w_{a,1} \cdot x_1 + w_{a,4} \cdot x_2)}{\sum_y (\exp(w_{y,1} \cdot x_1 + w_{y,4} \cdot x_2))} \end{aligned}$$

where in the second line we have used the fact that $x_2 = x_3$ and in the last line we have just done a change of variables. Thus, without regularization logistic regression will ignore the repeated features.

- Does addition of ℓ_2 regularization to logistic regression affect its sensitivity to repeated features? (If so, how?)

  Regularization will make logistic regression more sensitive to repeated features. Let us consider the derivation above again:

$$P(y = a|x_1, x_2, x_3, \mathbf{w}) = \log \frac{\exp(w_{a,1} \cdot x_1 + w_{a,2} \cdot x_2 + w_{a,3} \cdot x_3)}{\sum_y (\exp(w_{y,1} \cdot x_1 + w_{y,2} \cdot x_2 + w_{y,3} \cdot x_3))} + \lambda \sum_y \sum_{i=1}^{3} \|w_{y,i}\|^2$$

$$= \log \frac{\exp(w_{a,1} \cdot x_1 + (w_{a,2} + w_{a,3}) \cdot x_2)}{\sum_y (\exp(w_{y,1} \cdot x_1 + (w_{y,2} + w_{y,3}) \cdot x_2))} + \lambda \sum_y \sum_{i=1}^{3} \|w_{y,i}\|^2$$

  Now we note that we cannot simply do a change of variables because of the regularization term. Due to the nature of the regularizer (i.e. each weight is squared), x_2 will effectively be up-weighted since it is associated with multiple weights.