Practice Exam Problems: Log-Linear Modeling

Natural Language Processing (JHU 601.465/665)

Prof. Jason Eisner

1. In homework 2, we met beavers who communicated with cries from the vocabulary $V = \{\texttt{bwa}, \texttt{bwee}, \texttt{kiki}\}$. Let's build a simple language model of beavers.

   First let's assume that each sentence consists of a single cry. Then the language model only has to specify a distribution over V. In a training corpus, you observe empirical probabilities

   $$\tilde{p}(\texttt{bwa}) = 0.3 \qquad \tilde{p}(\texttt{bwee}) = 0.2 \qquad \tilde{p}(\texttt{kiki}) = 0.5$$

   We'll use a log-linear model with three features, one per cry:

   $$p_\theta(\texttt{bwa}) = \tfrac{1}{Z}\exp\theta_{\texttt{bwa}} \qquad p_\theta(\texttt{bwee}) = \tfrac{1}{Z}\exp\theta_{\texttt{bwee}} \qquad p_\theta(\texttt{kiki}) = \tfrac{1}{Z}\exp\theta_{\texttt{kiki}}$$

   (a) If you train the parameter vector θ to maximize likelihood, then
   $p_\theta(\texttt{bwee}) =$ __________

   (b) Give a parameter vector that would maximize likelihood. (Hint: Think about the gradient.)

   $\theta_{\texttt{bwa}} =$ __________ $\theta_{\texttt{bwee}} =$ __________ $\theta_{\texttt{kiki}} =$ __________

   (c) The maximizer is not unique. Give a *different* parameter vector that would *also* maximize likelihood.

   $\theta_{\texttt{bwa}} =$ __________ $\theta_{\texttt{bwee}} =$ __________ $\theta_{\texttt{kiki}} =$ __________

   (d) Which parameter vector would an L_1 regularizer prefer?
   MY ANSWER FROM (b) MY ANSWER FROM (c) *(circle one)*

Suppose you simplify your model to the following:

$$p_\theta(\texttt{bwa}) = \tfrac{1}{Z}\exp\theta_{\texttt{bw}} \qquad p_\theta(\texttt{bwee}) = \tfrac{1}{Z}\exp\theta_{\texttt{bw}} \qquad p_\theta(\texttt{kiki}) = \tfrac{1}{Z}\exp\theta_{\texttt{kiki}}$$

This model has only has 2 parameters, since now **bwa** and **bwee** share a parameter $\theta_{\texttt{bw}}$.

(e) Under the new model, the answer to question 1a is $p_\theta(\texttt{bwee}) =$ __________

Now we discover that beaver sentences can be of any length, so we need a distribution over sentences $x \in V^*$:

$$p(x) = \tfrac{1}{Z}\exp\left(\theta_{\texttt{bwa}}\, c_x(\texttt{bwa}) + \theta_{\texttt{bwee}}\, c_x(\texttt{bwee}) + \theta_{\texttt{kiki}}\, c_x(\texttt{kiki})\right)$$

where $c_x(w)$ is the count of word w in sentence x. In our new training corpus of sentences, we again observe

$$\tilde{p}(\texttt{bwa}) = 0.3 \qquad \tilde{p}(\texttt{bwee}) = 0.2 \qquad \tilde{p}(\texttt{kiki}) = 0.5$$

Also, the average training sentence has length 4.0.

(f) Note that our model doesn't have any features for sentence length or for EOS. If you train the parameter vector θ to maximize likelihood, will the trained model correctly predict that sentences have length 4.0? Answer YES or NO, and explain your answer.

2. Dr. Abecedarian is interested in predicting how long it takes a student to learn a language. The possible answers are 0 years, 1 year, 2 years, 3 years, . . . 10 years. (No one ever admits to taking more than 10 years, so larger numbers don't show up in his training data!)

   For example, many people can have a basic conversation in Esperanto after 1 year, or Italian after 2 years. But it usually takes a student 4 or more years to achieve conversational ability in Polish, although some students may be faster because of luck or skill.

   Dr. Abecedarian has collected N triples of the form (student, language, number of years) = (s, ℓ, y).

(a) Define a single feature function f_1 that allows a log-linear model to capture that fact that Esperanto is fast to learn.

$f_1(\underline{\hspace{4cm}}) = \underline{\hspace{8cm}}$

(b) The students of Esperanto in Dr. A.'s training data took an average of 1.6 years to learn Esperanto. Suppose f_1 is the *only* feature in the model. If you train the model by maximum conditional likelihood, then what is the optimal value of θ_1?

(You would need a calculator to get the exact value. Don't worry about that; just show how you would calculate θ_1, and indicate whether you expect it to be positive or negative.)

(c) What does your trained model predict is the most common amount of time needed to learn Esperanto?

(d) No one in training data was able to learn a language in 0 years. Define a single feature function f_2 that allows a log-linear model to capture this fact.

$f_2(\underline{\hspace{4cm}}) = \underline{\hspace{8cm}}$

(e) Suppose that f_1 and f_2 are the *only* features in your model. If you train the model by maximum conditional likelihood, then what is the optimal value of θ_2?

(f) What additional feature function or functions would help the model capture the fact that some students are "very good at languages"? For example, if training data shows that Helen has learned Polish unusually quickly, in only 2 years rather than the usual 4, the model might guess that she will also be able to learn Italian unusually quickly.

(g) Given all the features above, Dr. Abecedarian wants to predict whether *you* will be good at learning Italian. You did not appear in training data (and you have no other features that make you similar to people in training data). Does the trained model think *you* are good at languages? Discuss.

3. Homework 2 had a problem about the color of automobiles. Here's another.

An annual survey from DuPont Automotive[1] reported the 2010 distribution of automobile colors (in percentages), for each geographic region:

	China	Europe	India	N. America	S. America
Black	31	24	8	18	23
Blue	2	9	6	9	3
Brown	1	5	11	5	3
Gray	18	19	9	15	13
Red	4	7	8	11	9
Silver	33	17	24	17	33
White	9	14	29	21	13
OOV	2	5	5	4	3

Each column sums to 100%. I've chosen to limit our color vocabulary to the above set of 7 words. All other colors are therefore considered to be OOV. Notice that some OOV colors (e.g., green, purple, gold) had positive counts in DuPont's data.

You would like to build a log-linear model that you can take with you on your travels. When you arrive in a new region r, your model should tell you what colors c you can expect to see. It specifies conditional probabilities of the form

$$p(\text{Color} = c \mid \text{Region} = r)$$

for all (c, r) pairs. If your model is good, then it matches the table, e.g.,

$$p(\text{Color} = \text{black} \mid \text{Region} = \text{china}) \approx 0.31$$

(a) Let's start with a super-simple model. It has only a *single* feature f_1, defined by

$$f_1(c, r) = \begin{cases} 1 \text{ if } c = \text{black and } r = \text{china} \\ 0 \text{ otherwise} \end{cases}$$

Suppose you set weight $\theta_1 = 1$.

- In China, what set of features fires on black cars? ____________________
- In China, what set of features fires on white cars? ____________________

[1] http://www2.dupont.com/Media_Center/en_US/color_popularity/

- In India, what set of features fires on black cars? ____________________
- In India, what set of features fires on white cars? ____________________

Fill in the numerators and denominators of the following probabilities (you don't need to compute their numeric values):

$$p(c = \text{black} \mid r = \text{china}) = \frac{u(c = \text{black}, r = \text{china})}{Z(\text{china})} = \underline{\hspace{7cm}}$$

$$p(c = \text{black} \mid r = \text{india}) = \frac{u(c = \text{black}, r = \text{india})}{Z(\text{india})} = \underline{\hspace{7cm}}$$

(b) Now let's make the model a little bit more complicated. Keep f_1 with $\theta_1 = 1$, but add two more features:

$$f_2(c, r) = \begin{cases} 1 \text{ if } c = \text{black} \\ 0 \text{ otherwise} \end{cases} \qquad f_3(c, r) = \begin{cases} 1 \text{ if } r = \text{china} \\ 0 \text{ otherwise} \end{cases}$$

with $\theta_2 = 2$ and $\theta_3 = 3$. Answer the same questions as above:

- In China, what set of features fires on black cars? ____________________
- In China, what set of features fires on white cars? ____________________
- In India, what set of features fires on black cars? ____________________
- In India, what set of features fires on white cars? ____________________

Fill in expressions for the following (you don't need to compute their numeric values):

$$p(c = \text{black} \mid r = \text{china}) = \frac{u(c = \text{black}, r = \text{china})}{Z(\text{china})} = \underline{\hspace{7cm}}$$

$$p(c = \text{black} \mid r = \text{india}) = \frac{u(c = \text{black}, r = \text{india})}{Z(\text{india})} = \underline{\hspace{7cm}}$$

(c) Now imagine that you can include as many features as you want, and you can set their weights $\vec{\theta}$ however you want. You might like to find a "perfect" setting of $\vec{\theta}$ such that your model $p(c \mid r)$ perfectly predicts the probabilities in the table.

If such a “perfect” setting of $\vec{\theta}$ exists, one way to find it would be to use an optimizer. A rather direct approach would be to minimize the total squared error of your predictions (known as least-squares regression):

$$
\begin{array}{ccccccc}
 & (0.31 - p(\text{black} \mid \text{china}))^2 & + & (0.24 - \cdot p(\text{black} \mid \text{europe}))^2 & + & (0.08 - \cdot p(\text{black} \mid \text{india}))^2 & + \cdots \\
+ & (0.02 - p(\text{blue} \mid \text{china}))^2 & + & (0.09 - \cdot p(\text{blue} \mid \text{europe}))^2 & + & (0.06 - \cdot p(\text{blue} \mid \text{india}))^2 & + \cdots \\
+ & (0.01 - p(\text{brown} \mid \text{china}))^2 & + & (0.05 - \cdot p(\text{brown} \mid \text{europe}))^2 & + & (0.11 - \cdot p(\text{brown} \mid \text{india}))^2 & + \cdots \\
 & & \vdots & & & &
\end{array}
$$

where the coefficients are drawn from the table.[2] This function is minimized when all the squared terms are 0—in which case the predictions are perfect.

However, you could equally well find the perfect setting by maximizing

$$
\begin{array}{ccccccc}
 & 31 \cdot \log p(\text{black} \mid \text{china}) & + & 24 \cdot \log p(\text{black} \mid \text{europe}) & + & 8 \cdot \log p(\text{black} \mid \text{india}) & + \cdots \\
+ & 2 \cdot \log p(\text{blue} \mid \text{china}) & + & 9 \cdot \log p(\text{blue} \mid \text{europe}) & + & 6 \cdot \log p(\text{blue} \mid \text{india}) & + \cdots \\
+ & 1 \cdot \log p(\text{brown} \mid \text{china}) & + & 5 \cdot \log p(\text{brown} \mid \text{europe}) & + & 11 \cdot \log p(\text{brown} \mid \text{india}) & + \cdots \\
 & & \vdots & & & &
\end{array}
$$

This second version can be regarded as maximizing the conditional log-probability of a particular training “corpus,” consisting of cars observed in various countries. **How many car tokens are in that “corpus”? __________**

Should the first objective include the OOV row of the table?

YES NO DOESN’T MATTER *(circle one)*

Should the second objective include the OOV row of the table?

YES NO DOESN’T MATTER *(circle one)*

(d) In the previous question, we didn’t specify a particular set of features. Let’s do that now.

- Let’s rename our old feature f_1 to $f_{\text{black,china}}$, and replace it with a full set of $7 \times 5 = 35$ features of the form $f_{c,r}$.
- Let’s rename our old feature f_2 to f_{black}, and replace it with a full set of 7 features of the form f_c.
- Let’s rename our old feature f_3 to f_{china}, and replace it with a full set of 5 features of the form f_r.

[2] Only the part of the equation corresponding to the upper left corner of the table is shown here. The rest of the equation (both rows and columns) is given by “$\cdots$”.

So now we have a model with $35 + 7 + 5 = 47$ features. There are now *many* "perfect" settings of $\vec{\theta}$. Intuitively, that's because the optimizer's ability to control 47 numeric parameters gives it lots of ways to match the < 47 numeric probabilities from the table.

In this class, you've learned a way to add a "regularizer" to the objective function so that it prefers some of these perfect settings over others.

- What is a formula for the regularizer here? In class we summed over k, but here the features are indexed by c and r, so you should write sums over c and r.
- Let's assume from now on that you maximize the second objective above, plus a regularizer. Will the optimizer still choose one of the "perfect" settings of $\vec{\theta}$? YES NO EITHER IS POSSIBLE
- In what sense does the regularizer help? That is, how might the model chosen with regularization be a "better" or more useful model of car colors, compared to one chosen without regularization?
- What values do you expect the optimizer to choose for:
  θ_{brown} ? POSITIVE ZERO NEGATIVE
  θ_{india} ? POSITIVE ZERO NEGATIVE
  $\theta_{\text{brown,india}}$? POSITIVE ZERO NEGATIVE
- You have to go to Australia on short notice. You're nervous that you don't know anything about Australian cars. But happily, you realize that your trained model can still predict car colors in Australia, even though they're not in the DuPont data!
  What does the model think about $p(\text{brown} \mid \text{Australia})$? *(circle one)*
  0 BETWEEN 0 AND 1/8 1/8 BETWEEN 1/8 AND 1
  (*Hint:* Even though you have no Australia-specific features in the model, there are other features that will fire here.)
- You could add Australia-specific features to your model, but they would get

weight 0. Why?
(*Hint*: None of the cars in your training corpus were from Australia. Think about how this affects the training objective, or the gradient of the objective.)

(e) "Uncolored" cars seem to be popular everywhere. You can see that some regions strongly prefer black cars to white, or vice-versa. But in every region, black+white together are always 36–40% of the total.

- Design a new binary feature that will help model this consistent preference for "uncolored" cars:

$$f_{\text{new}}(c, r) = \begin{cases} 1 \text{ if } \underline{\hspace{8cm}} \\ 0 \text{ otherwise} \end{cases}$$

- If you add this feature to the model above, it should get a positive weight (because "uncolored" cars are popular). What will probably happen as a result to the weights for

$\theta_{\text{black,india}}$? INCREASE SAME DECREASE
$\theta_{\text{white,india}}$? INCREASE SAME DECREASE

(f) Europe currently has far more cars than China or India. Prof. Bizbuz suggests that you could reflect this fact by multiplying the Europe column by 10, as follows:

$$\begin{array}{lrcrcrl}
 & 31 \cdot \log p(\text{black} \mid \text{china}) & + & \mathbf{240} \cdot \log p(\text{black} \mid \text{europe}) & + & 8 \cdot \log p(\text{black} \mid \text{india}) & + \cdots \\
+ & 2 \cdot \log p(\text{blue} \mid \text{china}) & + & \mathbf{90} \cdot \log p(\text{blue} \mid \text{europe}) & + & 6 \cdot \log p(\text{blue} \mid \text{india}) & + \cdots \\
+ & 1 \cdot \log p(\text{brown} \mid \text{china}) & + & \mathbf{50} \cdot \log p(\text{brown} \mid \text{europe}) & + & 11 \cdot \log p(\text{brown} \mid \text{india}) & + \cdots \\
 & & \vdots & & & & \\
+ & \text{regularizer} & & & & &
\end{array}$$

How would this change affect your distribution $p(c \mid r)$?