Assignment 1

CSE 517: Natural Language Processing – University of Washington

Winter 2020 – Due: January 20, 2020, 11:59 pm

From the Textbook

Complete these exercises from the Eisenstein book. They are ordered roughly the same way the content is ordered in lectures.

- Language modeling: 6.4, 6.5, and 6.6 (pp. 135–6)
- Linear classification: 2.5 and 2.6 (pp. 43–44)
- Implementation: 4.6 (p. 89); the last part, performing statistical significance tests, is extra credit
- Nonlinear classification: 3.4, 3.5, and 3.6 (p. 65)

Maximum *a Posteriori* Estimation

In the noisy channel paradigm, we consider two random variables, X and Y, and imagine that the observable X is a noisy or corrupted version of Y.

$$\boxed{\text{source}} \longrightarrow Y \longrightarrow \boxed{\text{channel}} \longrightarrow X$$

For example, if we observe a Chinese sentence (a value for X) but believe it was "originally" an English sentence (a value for Y), we can build source and channel models to capture, respectively, the distribution over English sentences and the translation of English into Chinese. Then, we use Bayes' rule to "decode":

$$\hat{y} = \operatorname*{argmax}_y p(y \mid x) = \operatorname*{argmax}_y \frac{p(y) \cdot p(x \mid y)}{p(x)} = \operatorname*{argmax}_y p(y) \cdot p(x \mid y)$$

Question 1: Why can we drop the $p(x)$ factor in the denominator?

Interestingly, this same "noisy channel" idea can also be applied when we are estimating the parameters of a probabilistic model, though it goes by a different name: maximum *a posteriori* (MAP) estimation. The story looks like this:

$$\boxed{\text{source}} \longrightarrow \Theta \longrightarrow \boxed{\text{channel}} \longrightarrow X$$

Instead of Y, we use the random variable Θ for the unknown parameters of our model. X corresponds to the training data.

Applying Bayes' rule gives us:

$$\hat{\theta} = \operatorname*{argmax}_{\theta} p(\theta) \cdot p(x \mid \theta)$$

Here, we're using the notation $p(x \mid \theta)$ instead of $p_\theta(x)$, since we're thinking of θ as a random variable (its name would be Θ) with its own probability distribution, but both are usually fine and they mean the same thing.

It's helpful to think about this for a specific model, so let's take the simplest one we know: the unigram model. The parameters for the unigram model are a vector $\boldsymbol{\theta} \in \triangle^V$.[1] The evidence is a sequence of words in a text, denoted $\boldsymbol{x}$, though—as discussed in class—we really only need to know the count of each word, not the order in which they occur.

Let's consider the "channel" model first; it defines the probability of $\boldsymbol{x}$ given the parameters. Recall this model from class:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \theta_{x_i} = \prod_{v \in \mathcal{V}} \theta_v^{\text{count}_{\boldsymbol{x}}(v)} \tag{1}$$

where $\text{count}_{\boldsymbol{x}}(v)$ is the count of word v in the data $\boldsymbol{x}$.

We discussed maximum likelihood estimation (MLE), in which we select $\hat{\boldsymbol{\theta}}$ to maximize the above quantity, and proved that relative frequency estimation accomplishes MLE. You should think of MAP as an *alternative* to MLE.

The "source" model is a bit harder to think about at first. We need a probability distribution over $\triangle^V$. An attractive candidate for $p(\Theta)$ is the **Dirichlet** distribution.[2]

Here is the form of the Dirichlet distribution:

$$p_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{\Gamma\left(\sum_{v \in \mathcal{V}} \alpha_v\right)}{\prod_{v \in \mathcal{V}} \Gamma(\alpha_v)} \prod_{v \in \mathcal{V}} \theta_v^{(\alpha_v - 1)} \tag{2}$$

This is somewhat daunting, so we'll break it down:

- The parameters $\boldsymbol{\alpha} \in \mathbb{R}^V$ are a vector of (strictly) positive values. If we normalize them (divide each by their sum), we get the mean value of the Dirichlet; if we sampled many different values of Θ, their average would be close to this normalized version of $\boldsymbol{\alpha}$. As these values get larger, the tendency of draws from the Dirichlet to be close to the mean increases. When they are smaller, draws from the Dirichlet will be more diffuse.
- The Γ function is an extension of the factorial function to the nonnegative reals. For natural numbers x, $\Gamma(x) = (x-1)!$. The more general form is not important for this assignment; if you are curious, read more at `http://mathworld.wolfram.com/GammaFunction.html`.
- Fortunately for us, the Γ factors are constant with respect to $\boldsymbol{\theta}$, so they don't affect our estimation procedure.

[1] The probability simplex in V dimensions, denoted $\triangle^V$, is defined as the set of all V-length vectors that are (i) nonnegative and (ii) sum to one.

[2] We mentioned this family briefly when we talked about latent Dirichlet allocation, but did not go into details.

The fascinating thing about Dirichlet distributions and categorical distributions like the unigram model's $\boldsymbol{\theta}$ is that the form of the posterior distribution $p(\Theta \mid \boldsymbol{x})$ is also a Dirichlet distribution. When this happens, Bayesian statisticians get very excited and use the term "conjugate prior" to denote the special relationship between a family of priors (here, Dirichlet distributions) and likelihood functions (here, categorical distributions). In fact, this property is why the Dirichlet was chosen as the prior in latent Dirichlet allocation in the first place.

Consider the logarithm of this quantity as a function of $\boldsymbol{\theta}$:

$$\log p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \log p_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) + \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) - \log p(\boldsymbol{x}) \tag{3}$$

$$= \log p_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) + \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) + \text{constant} \tag{4}$$

$$= \log \frac{\Gamma\left(\sum_{v \in \mathcal{V}} \alpha_v\right)}{\prod_{v \in \mathcal{V}} \Gamma(\alpha_v)} \prod_{v \in \mathcal{V}} \theta_v^{(\alpha_v - 1)} + \log \prod_{v \in \mathcal{V}} \theta_v^{\text{count}_{\boldsymbol{x}}(v)} + \text{constant} \tag{5}$$

$$= \log \prod_{v \in \mathcal{V}} \theta_v^{(\alpha_v - 1)} + \log \prod_{v \in \mathcal{V}} \theta_v^{\text{count}_{\boldsymbol{x}}(v)} + \text{constant} \tag{6}$$

where we've just plugged in the expressions from (1) and (2). Starting on line (4), we fold all terms that do not depend on $\boldsymbol{\theta}$ into "constant."[3]

Question 2: Conjugacy means that the above posterior can be expressed as a Dirichlet distribution with some parameters; call them $\boldsymbol{\alpha'}$. Write α'_v as a function of $\boldsymbol{\alpha}$ and the counts $\text{count}_{\boldsymbol{x}}(*)$.

MAP estimation requires that we find the most probable value of $\boldsymbol{\theta}$ under this posterior—in other words, its mode. The mode has a closed form whenever all $\alpha_v > 1$ for all $v \in \mathcal{V}$. The closed form is:

$$\left[\operatorname*{argmax}_{\boldsymbol{\theta}} p_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \cdot p(\boldsymbol{x} \mid \boldsymbol{\theta})\right]_v = \frac{\alpha'_v - 1}{\sum_{v' \in \mathcal{V}} \alpha'_{v'} - V} \tag{7}$$

for each $v \in \mathcal{V}$.

Question 3: For an appropriate choice of $\boldsymbol{\alpha}$, the Dirichlet can be "uninformative," meaning that it assigns equal probability to all possible $\boldsymbol{\theta}$. In this case, MAP is the same as MLE! Derive the values of $\boldsymbol{\alpha}$ that will make this happen.

Question 4: For a (different) appropriate choice of $\boldsymbol{\alpha}$, we make MAP estimation equivalent to Laplace smoothing.[4] Derive the values of $\boldsymbol{\alpha}$ that will make this happen.

Question 5: A *symmetric* Dirichlet distribution is one where all α_v take the same value (say, a). Show that, for any a that is sufficiently large, we can write the MAP estimate as an interpolation between the MLE (weighted by some λ) and a uniform distribution over $\mathcal{V}$ (weighted by $1 - \lambda$). Derive λ as a function of a.

[3] What we mean here is only that they are constant *with respect to* $\boldsymbol{\theta}$ and therefore do not affect our objective of finding $\boldsymbol{\theta}$ to maximize the (log-)likelihood.

[4] In lecture, we may have mentioned Laplace smoothing casually but probably not in detail or with the name "Laplace smoothing." It's what the instructor would have described as "the simplest smoothing approach you can think of." Before normalizing counts into probabilities (as you do in MLE), take the count of every $v \in \mathcal{V}$ (including the v that have a count of zero) and add one to it. So now, everything you didn't see, you're pretending you saw once. Everything you *did* see, you're pretending you saw it one more time than you did. If, instead of adding one, you add a (positive) value λ, then this method is called "add-λ" smoothing.