CSC 446 Notes: Lecture 3

January 31, 2013

1 Smoothing

The Naive Bayes classifier that we are using in the homework is

$$P(y|\mathbf{x}) \propto P(y) \prod_{j=1}^{J} P(x_j|y)$$

By making the "Naive" independence assumption last time, we were able to factor and get rid of a lot of potential zeros in the product. Since all of our probabilities are based on counts, any unseen combination of a single feature x and the class label y results in

$$\begin{aligned} P(x|y) &= \frac{c(x,y)}{c(y)} \\ &= 0. \end{aligned}$$

These zeros can ruin the entire classifier. For example, say there's one bill where all the Republicans we know about voted "no". Now, say we are trying to classify an unknown politician who followed the Republican line on every other bill, but voted "yes" on this bill. The classifier will say that there is zero probability of this person being a Republican, since it has never seen the combination (Republican, voted yes) for that bill. It gives that single feature way too much power. To get rid of that, we can use a technique called smoothing, and modify the probabilities a little :

$$\begin{aligned} P(x = k|y) &= \frac{c(x = k, y) + \alpha}{c(y) + K\alpha} \\ k &\in \{1, ..., K\} \end{aligned}$$

Basically we are taking a little bit of the probability mass from things with high probability and giving it to things with otherwise zero probability. (Republicans might veto this technique, since it's like redistribution of wealth!) Note that these probabilities must still sum to 1. This seems great - we've gotten rid of things with zero probability. But doesn't this contradict what we proved earlier? That is, last week we said that we can best infer the probability distribution by solving

$$\begin{aligned} &\underset{\theta}{\operatorname{argmax}} \prod_{n=1}^{N} P_\theta(x_n) \\ &\text{s.t.} \sum_{k=1}^{K} \theta_k = 1 \end{aligned}$$

which results in the count-based distribution

$$\theta_k^* = \frac{c(k)}{N}.$$

How then can we mathematically justify our smoothed probabilities?

1.1 Prior Distributions

We can treat θ as a random variable itself with some probability distribution $P(\theta)$. Recall that θ is a vector of probabilities for each type of event k, so

$$\theta = [\theta_1, \theta_2, \ldots \theta_K]^T$$
$$\text{and } \sum_{k=1}^{K} \theta_k = 1$$

Suppose that we have a coin with two outcomes, heads or tails (K=2). We can picture the θ_1 and θ_2 which we could pick for the probability distribution of these two outcomes. A fair coin has $\theta_1 = 1/2$ and $\theta_2 = 1/2$. An weighted coin might have $\theta_1 = 2/3$ and $\theta_2 = 1/3$. Since we are treating θ as a random variable, its probability $P(\theta)$ is describing the probability that it takes on these values. $P(\theta)$ is called a prior, since it's what we believe about θ before we even have any observations. For example, we might tend to believe that the coin will be pretty fair, so we could have $P(\theta)$ be a normal curve with the peak where $\theta_1 = 1/2$ and $\theta_2 = 1/2$.

1.2 Dirichlet Prior

One useful prior distribution is the Dirichlet Prior :

$$\begin{aligned} P(\theta) &= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \\ &= \frac{1}{Z} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \end{aligned}$$

This is also written as $P(\theta; \alpha)$, $P_\alpha(\theta)$, or $P(\theta|\alpha)$. α is a vector with the same size as θ, and it is known as a "hyperparameter". The choice of α determines the shape of θ's distribution, which you can see by varying it. If α is simply a vector of ones, we just get a uniform distribution; all θs are equally probable. In the case of two variables, we can have α_1=100, and α_2=50 and we see a sharp peak around $2/3$. The larger α_1, the more shaply peaked it gets around $\frac{\alpha_1}{\alpha_1 + \alpha_2}$.

At this point, we are tactfully ignoring that Γ in the Dirichlet distribution. What is that function, and what does it do?

1.3 Gamma Function

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

This function occurs often in difficult, nasty integrals. However, it has the nice property of being equivalent to the factorial function :

$$\Gamma(n) = (n-1)!$$

We can prove this using integration by parts:

$$\begin{aligned}\Gamma(x) &= \int_0^\infty e^{-t}t^{x-1}dt \\ &= \left[-t^{x-1}e^{-t}\right]_0^\infty + \int_0^\infty e^{-t}(x-1)t^{x-2}dt \\ &= 0 + (x-1)\int_0^\infty e^{-t}t^{x-2}dt \\ &= (x-1)\Gamma(x-1)\end{aligned}$$

Further noting that $\Gamma(1) = 1$, we can conclude that $\Gamma(n) = (n-1)!$. This function is used in our Dirichlet prior to guarantee that

$$\Gamma(x) = \int_{\sum_k \theta_k = 1} P(\theta)d\theta = 1.$$

1.4 Justifying the Dirichlet Prior

How can we use this prior to compute probabilities?

$$\begin{aligned}P(x = k|\theta) &= \theta_k \\ P(x = k) &= \int_{\sum_k \theta_k = 1} P(x|\theta)P(\theta)d\theta \\ &= \int \theta_k \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1} d\theta \\ &= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \int \prod_{k'=1}^K \theta_{k'}^{\alpha_{k'} - 1 + I(k'=k)} d\theta \\ &= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\prod_{k'} \Gamma(\alpha_{k'} + I(k'=k))}{\Gamma(\sum_{k'} \alpha_{k'} + I(k'=k))} \\ &= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\Gamma(\sum \alpha_k + 1)} \frac{\Gamma(\alpha_k + 1)}{\Gamma(\alpha_{k'})}\end{aligned}$$

Now we use $\Gamma(x) = (x-1)\Gamma(x-1)$:

$$= \frac{\alpha_k}{\sum_{k'} \alpha_{k'}}$$

Most of the time, all of the α_k's are set to the same number. So, we just showed that

$$P(x) = = \frac{\alpha_k}{\sum_{k'} \alpha_{k'}}$$

But what about

$$\begin{aligned}
P(X_{N+1}|X_1^N) &= \int P(X_{N+1}, \theta|X_1^N)d\theta \\
&= \int P(X_{N+1}|\theta, X_1^N)P(\theta|X_1^N)d\theta \\
&= \int \theta_k \frac{P(X_1^N|\theta)P(\theta)}{P(X_1^N)}d\theta \\
&= \frac{1}{Z}\int \theta_k \prod_n \theta_{X_n} \frac{1}{Z'}\prod_k \theta_k^{\alpha_k - 1}d\theta \\
&\dots \\
&= \frac{c(k) + \alpha_k}{N + \sum_k \alpha_k}
\end{aligned}$$

2 Comparison - Bayesian vs. MLE vs. MAP

2.1 Bayesian

The quantity we just computed is known as the Bayesian:

$$P(X_{N+1}|X_1^N) = \frac{c(k) + \alpha_k}{N + \sum_k \alpha_k}$$

We can compare it to the MLE that we did before:

$$\begin{aligned}
P(x_{N+1}) \text{ follows } \theta^* \\
\theta^* = \underset{\theta}{\operatorname{argmax}} P_\theta(X_1^N)
\end{aligned}$$

And a third alternative is the MAP, or Maximum A Posteriori:

$$\begin{aligned}
P(x_{N+1}) \text{ follows } \theta^* \\
\theta^* = \underset{\theta}{\operatorname{argmax}} P(\theta)P(X_1^N|\theta)
\end{aligned}$$

This is simpler since it does not require an integral. Using the same Lagrange Multipliers technique as we did before:

$$\begin{aligned}
&\operatorname{argmax} \frac{1}{Z}\prod_k \theta_k^{\alpha_k - 1}\prod_k \theta_k^c(k) \\
&\text{s.t.} \sum_k \theta_k = 1
\end{aligned}$$

Then we get the result:

$$\theta_k^* = \frac{c(k) + \alpha_k - 1}{N + (\sum_{k'} \alpha_k') - K}$$

Carolyn Keenan 1/12, DG 1/13, IN 1/13