STATS205: Introduction to Nonparametric Statistics

Lecturer: Tengyu Ma
Scribe: Benjamin Anderson

Lecture # 10
June 4th, 2021

10.1 Review and overview

Last lecture, we discussed the Gaussian process, which is a Bayesian approach to supervised, nonparametric problems. It can be thought of as a generalization of the mixture of Gaussians model, with an infinite number of Gaussian distributions. Today, we will discuss the Dirichlet process. As with the Gaussian process, our discussion of the Dirichlet process will begin with simpler, parametric mixture models, which we will then build on to understand the more complex Dirichlet process.

Unlike the Gaussian process, the Dirichlet process is used in an unsupervised setting, to model a distribution over some variable X, rather than modeling a conditional distribution of $Y \mid X$. We'll first review parametric mixture models, which are one way to model a probability distribution. Then, we'll discuss how to extend these models to a Bayesian setting, by establishing a prior over the parameters (this will require a tangent to define the Dirichlet distribution). Then, we'll discuss topic modeling, a popular type of unsupervised machine learning model, as an entry point into the Dirichlet process. Finally, we will define the Dirichlet process itself, which can be thought of as a topic model with infinite topics.

10.2 Parametric mixture models & extension to Bayesian setting

10.2.1 Review: Gaussian mixture model

The "mixture of k Gaussians" is a probability distribution with the following generative "story" for how a sample X_i is generated:

1. First, one of k "sources", z_i, is chosen from a discrete (or categorical) distribution π (which can be thought of as simply a non-negative, k-dimensional vector whose components sum to 1).

2. Then, X_i is sampled from a Gaussian distribution whose mean and covariance are conditional on the choice of z_i, i.e. $X_i \mid z_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$

The result is called a mixture of Gaussians because it is as if you combined k Gaussian distributions (each with some weight π_k) into one distribution. You can read more about this model in the previous set of lecture notes.

10.2.2 Dirichlet distribution

In order to extend the mixture of Gaussians to a Bayesian setting, we have to establish priors over the parameters. There are many ways to get priors over μ_{z_i} and Σ_{z_i}, and we won't go into detail on that (a unit normal distribution and chi-squared distribution respectively is one example). However, the choice of parameters for π is unique, as it is constrained: we must have that $\sum_{i=1}^{k} = 1$, so not

just any choice of $\pi_1, \ldots, \pi_k$ is a valid probability distribution. Our distribution over π should only have probability mass on valid choices of π. The Dirichlet distribution (denoted *Dir* for short in mathematical equations), is a natural choice.

In the Bayesian setting, we are interested in studying distributions over the *parameters* of another distribution (which are themselves random variables). A simpler example of this idea is the Beta distribution, which is a probability distribution over the parameter p for a binomial random variable (i.e. a coin flip). The Beta distribution represents the probability distribution over the "true" probability of the coin coming up heads. A Dirichlet distribution is simply a generalization of the Beta distribution to an experiment with more than two outcomes, e.g. a dice roll. So, the Dirichlet distribution could be used to model the distribution over the parameters $\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6$ representing the probability each side of the die has of being rolled.

The Dirichlet distribution is parametrized by $\alpha_1, \ldots, \alpha_k$, which, as with the Beta distribution, can be interpreted as "pseudocounts", i.e. a larger α_i will result in a distribution where larger values of π_i have more density; and moreover, if their relative magnitudes are all held fixed, larger parameters denote more "confidence", resulting in a less uniform distribution. (As a simple example, if you've rolled each number on a die one time, you'd guess that all sides equally likely, but with low confidence. If you've rolled each number on a die 1000 times, you'd guess they're all equally likely, with high confidence.)

The Dirichlet distribution has a few important properties and related intuitions, some of which will be important for our later discussion of the Dirichlet process:

1. The PDF of the Dirichlet distribution over K-dimensional $\vec{\pi}$ is:

$$p(\vec{\pi}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)} \prod_{i=1}^{K} \pi_i^{\alpha_i - 1}$$

   ...where Γ is the Gamma function (too complicated to explain here).

2. $\mathbb{E}[\pi_i] = \frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j}$, which means that the relative magnitudes of the α's determine the expected relative magnitudes of the components of π. (The pseudocounts interpretation helps here: a larger α_i is a larger pseudocount, as if you've already observed that event more, so that should make the parameter for the probability of that event larger.)

3. $\sum_{i=1}^{K} \alpha_i$ controls how "sharp" the distribution is. Again, by the pseudocounts logic, having "seen" (or hallucinated) more data will make us more confident in what we think the distribution is, so this makes sense intuitively.

4. **Relationship to Gamma distribution:** If $\eta_k \overset{\text{iid}}{\sim} Gamma(\alpha_k, 1)$, for $i \in \{1, \ldots, K\}$, and $\pi_i = \frac{\eta_i}{\sum_j \pi_j}$, then $(\pi_1, \ldots, \pi_K) \sim Dir(\alpha_1, \ldots, \alpha_K)$.

5. **Merging rule:** If $(\pi_1, \ldots, \pi_K) \sim Dir(\alpha_1, \ldots, \alpha_K)$, then we can "merge" π's by summing them. Doing so will create a new Dirichlet distribution with fewer components, parametrized by new α's obtained by summing the α_j's corresponding to the π_j's that were combined. For example:

$$(\pi_1 + \pi_2, \pi_3 + \pi_4, \ldots) \sim Dir(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4, \ldots)$$

6. **Expanding rule:** Reverse merging rule; you can also obtain a new Dirichlet distribution from an existing one by "splitting" components; for example:

$$(\pi_1\theta, \pi_1(1-\theta), \pi_2, \ldots, \pi_K) \sim Dir(\alpha_1 b, \alpha_1(1-b), \alpha_2, \ldots, \alpha_K)$$

   ...where $\theta \sim Beta(\alpha_1 b, \alpha_1(1-b))$ for $0 < b < 1$.

7. **Renormalizing property:** Given $(\pi_1, \ldots, \pi_K) \sim Dir(\alpha_1, \ldots, \alpha_K)$, then if we discard one π_i and its associated α_i, and renormalize, we get another Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_K$.

10.2.3 Bayesian Gaussian mixture model

Now, if we wanted to extend the mixture of Gaussians to the Bayesian setting, we have the tools to do so. The only change from the frequentist version of the generative story is that the parameters themselves (μ's, Σ's, and π) are drawn from a prior distribution; *then* the latent variables z_i are drawn from $z \mid \pi \sim Categorical(\pi_1, \ldots, \pi_k)$; and finally $X_i \mid \mu, \Sigma, z_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$.

10.2.4 Dirichlet topic model

Topic modeling is a common unsupervised technique in natural language processing, which models a distribution over documents (collections of words) by grouping them into clusters (topics). This is a mixture model just like the mixture of Gaussians: each topic is a "source", and then the conditional on the document being associated with that source (topic), there's a set of probabilities associated with each word appearing in the document.

More formally, we set up with a vocabulary $\mathcal{V}$ with W words. Each document is represented as a vector in $\mathbb{R}^W$, with the i-th component equal to the number of times word i appears in it. (This is called a "bag of words" representation.) For simplicity, we assume each document is length n. Then, we aim to model the distribution over these document vectors with a mixture model. Unlike the Gaussian mixture, these document vectors can only take on non-negative integer values in each entry, so rather than a Gaussian conditioned on the source, we model a document as a multinomial distribution conditioned on its topic. (A multinomial distribution is just n trials of a categorical distribution.) So, we have parameters π for the choice of topic, and then θ_{z_i} for the multinomial distribution over words for each topic.

Then, the generative story is:

1. First, select a topic, $z_i \sim Categorical(\pi)$, where as before, the parameter π is a vector whose components sum to 1.

2. Generate n words with $Multinomial(n, \theta_{z_i})$, this produces the document X.

In order to make this Bayesian, which gets us to the Dirichlet Topic Model, we only need a prior over the parameters, π and θ_{z_i} for $i \in \{1, \ldots, W\}$, since the number of words n is fixed. Both can be Dirichlet priors: one for π with the number of parameters equal to the number of topics; and the other for each θ_{z_i} with the number of parameters equal to the number of words in the vocabulary. Then, to generate a document, we first sample π and all θ's, then follow the generative process above.

10.3 Dirichlet process

10.3.1 Overview

One way to think about the Dirichlet process is as a topic model whose number of topics is not fixed, but rather can grow with the number of data points grows. Rather than a fixed number of topics in advance, we allow choosing the number of topics to be "part of the model", in a sense. To do this, we need a prior that can generate probability vectors of any dimension, not just a fixed K, in other words, a distribution over $\bigcup_{K=1}^{\infty} \Delta_K$. We can think of the Dirichlet process as doing exactly this. Let's take some abstractions from the parametric model to generalize to the Dirichlet process setting.

In the parametric mixture models we've been discussing, you sample some parameters θ_k^* for each source (or topic) from some distribution H; sample π from a Dirichlet distribution; sample the latent z from $Categorical(\pi)$, and finally sample X from some distribution parametrized by θ_z. The important thing to take away here is that, for a given sample X_i, once you've fixed π and all the θ_k^*'s, then your choice of z_i completely determines θ_{z_i}, i.e. the set of parameters you'll use to select X_i.

Let's say we are modeling n examples, $X_1, \ldots, X_n$. For each, we can think about its corresponding θ_{z_i} itself as a random variable, drawn from a distribution G. G is fixed given a choice of π and all θ_k^*'s; which means that the prior for G is determined by the choice of α and H. A realization of G is a choice of θ_i (i.e. the θ used to sample X_i), which is one of the k possible choices $\theta_1^*, \ldots, \theta_K^*$. G is basically a discrete distribution with point masses on all the locations defined by $\theta_1^*, \ldots, \theta_K^*$, with the caveat that the magnitude of K is not fixed. The goal is to construct a prior over G, which in turn gives a distribution over θ_i, which in turn parametrizes a distribution over X_i.

There are two approaches for designing a prior for G. One is to directly construct it. (We'll do that later.) The other is to model the joint distribution over $\theta_1, \ldots, \theta_n$ (i.e. the choices of parameters for each of the n examples), which then implicitly defines G. We will start with this approach. This will require a few theoretical building blocks, which will occupy the next few sections.

10.3.2 Exchangeability & de Finett's theorem

Exchangeability is a fundamental concept in Bayesian statistics. Given a sequence of random variables $X_1, \ldots, X_n$, we say they are *exchangeable* if their joint distribution p is permutation-invariant. That is, if $p(X_1 = k_1, \ldots, X_n = k_n) = c$, then if we scramble up all the k's, the joint probability would still be c, no matter what order the k's are in. Furthermore, we say a sequence of random variables is *infinitely exchangeable* if any length-n prefix of the sequence is exchangeable for all $n \geq 1$.

Theorem 10.1. *De Finett's Theorem: If $\theta_1, \ldots, \theta_n$ are infinitely exchangeable, then there exists a random variable G such that $p(\theta_1, \ldots, \theta_n) = \int p(G) \prod_{i=1}^{n} p(\theta_i \mid G) dG$.*

In other words, there exists some G such that joint distribution over all n θ's "factors" and is equivalent to the distribution obtained by first sampling G from $p(G)$, then sampling θ_i from the distribution defined by G. The implication is that we don't have to define G directly; we can instead describe $\theta_1, \ldots, \theta_n$ (the "effect" of G) and this is sufficient (since by this theorem, G is guaranteed to exist, and we can do inference tasks using just the θ_i's).

10.3.3 The Chinese restaurant process

To define the joint distribution over $\theta_1, \ldots, \theta_n$, we first have to explain something called the Chinese restaurant process, which provides intuition for this distribution. Imagine a restaurant with infinitely many tables, and n customers. Customers enter one at a time, and sit at a table according to the following rules:

1. Customer 1 sits at table 1 with probability 1.
2. For $i > 1$, customer i sits at (occupied) table k with probability $\frac{n_k}{\alpha+i-1}$ (where n_k is the number of previous customers at that table), or else sits down and starts a new table with probability $\frac{\alpha}{\alpha+i-1}$.

Because all the n_k's add up to $i-1$ (the number of previous customers) we can quite easily confirm this setup makes sense (i.e. the probabilities of the customer's choices sum to 1). What does this thought process have to do with Dirichlet processes though? Well, let the latent variable z_i be the table number of the i-th customer. Then, if each "table" is assigned some $\theta_k^* \sim H$, then this gives us a way of picking θ_i's, by simply letting θ_i be the value assigned to the table where the i-th customer sits. This is also known as the Blackwell-MacQueen urn scheme.

This provides a joint distribution over $\theta_1, \ldots, \theta_n$. Moreover, it is exchangeable (possible to verify, but we won't do it here). Intuitively, it will result in some outcomes that "could be" IID draws from some discrete distribution G (informally speaking). Formally, applying de Finetti's theorem, because exchangeability holds, we know that there exists a G such that $\theta_1, \ldots, \theta_n$ chosen according to this scheme are equivalent to first sampling $G \sim DP(\alpha, H)$, then sampling each $\theta_i \mid G \overset{\text{iid}}{\sim} G$. We don't know what G is, just that it exists; and we can do all the interesting probabilistic inference without it (using just the θ_i's).

10.3.4 Explicitly constructing G (informal)

We don't have to specify G indirectly in this way—it can also be directly defined and constructed. First, a slightly-incorrect, informal treatment. We *basically* want an infinite-dimensional Dirichlet distribution, $\lim_{k\to\infty} Dir(\alpha/k, \ldots, \alpha/k)$. Then, we would just select some $\theta_k^* \sim H$ for each of these components, and have an infinite mixture. G could then be defined as a set of "point masses" with some density on each of the θ_k^*'s, i.e.:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

...where δ denotes the direct measure.

This is all slightly imprecise and incorrect, but it gets at the basic idea. To formalize it, we use a variate of the **merging rule** (10.2.2). Rather than summing pairs of π's, as discussed there, imagine *partitioning* π's into groups. By the same rule, we get a new Dirichlet distribution with a component for each partition, whose α parameters for each partition are the sum of the α_k's in that partition. The Dirichlet process is a bit different, because we will partition an infinite space, not a finite list of π's, but this is exactly the idea.

Recall that G is a distribution over the space of Θ, the set of all possible θ_k^*. (A discrete distribution, with point masses on certain possible values θ_k^*.) Consider the partition of Θ into $A_1, \ldots, A_m$. Then, $G(A_i)$ is basically the total mass of G that's contained in the A_i segment of the partition; i.e. $G(A_i) = \Pr[\theta \in A_i]$. This is deterministic for fixed G (but of course random otherwise, since G is a random variable itself). The claim is that $G(A_1), \ldots, G(A_m) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_m))$, i.e. that a partition of Θ into some finite number of segments results in a Dirichlet distribution.

We have $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$, but we can write $G(A_i)$ as the sum over only the mass in partition A_i, i.e $\sum_{k=1}^{\infty} \pi_k \mathbf{1}\{\theta_k^* \in A_i\}$. Or, letting I_j be the set of j such that $\theta_k^* \in A_j$, we can also write $G(A_j) = \sum_{k \in I_j} \pi_k$. We can then write:

$$(G(A_1), \ldots, G(A_m)) = \left(\sum_{k \in I_1} \pi_k, \ldots, \sum_{k \in I_m} \pi_k \right) \sim$$

$$Dir(\sum_{k \in I_1} \alpha_k, \ldots, \sum_{k \in I_m} \alpha_k)$$

This is close to showing the claim, but it's still not quite right because I is not fixed, it is a random variable. But, we can intuitively explain why the I's aren't that important. Remember, our goal is to get something like an infinite-dimensional Dirichlet distribution, with parameters α/K. Because of the idea of this uniform prior, we can say that $\sum_{\alpha_k \in I_j} \alpha_k = \sum_{k \in I_j} \alpha/K = |I_j|\alpha/K$, i.e. α multiplied by the fraction the probability mass in partition defined by I_j, which is just $\alpha H(A_j)$, which is what we wanted to show.

10.3.5 Explicitly constructing G (formal)

The formal definition of a Dirichlet process: A unique distribution over distributions on Θ such that for any partition $A_1, \ldots, A_m$ of Θ, we have that when $G \sim DP(\alpha, H)$, then $G(A_1), \ldots, G(A_m) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_m))$. We can explicitly construct such a distribution with the "stick-breaking construction."

1. Sample $\theta_k^* \overset{\text{iid}}{\sim} H$ for $k = 1, 2, \ldots, \infty$.
2. Choose $\beta_k \overset{\text{iid}}{\sim} Beta(1, \alpha)$ for $k = 1, 2, \ldots, \infty$.
3. Set $\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i)$.
4. Then, $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$.

It's called the "stick-breaking construction" because the intuition is that you begin with a stick of length 1, and then at the k-th step, break off the fraction β_k of what's left, and choose that

value for π_k. So first, the stick is length 1, you break off β_1, and so $\pi_1 = \beta_1$. Then, there's $(1 - \beta_1)$ left; you break off β_2 of that, so then $\pi_2 = (1 - \beta_1)\beta_2$. So on and so forth. This gives a formal construction of G for the Dirichlet process.

Then, all that remains is to do inference, which is typically done with Markov-Chain Monte Carlo (e.g. Gibbs Sampling). This is tractable, since the conditional distributions of the θ_i's have nice properties.

10.4 Summary

Since this is the last class, let's look back at what we've learned.

1. Non-parametric regression, including the kernel estimator, local polynomial/linear regression, splines, and using cross-validation to select a model and tune hyperparameters.
2. The kernel method, and its connection to splines and wide two-layer neural networks.
3. Neural networks, transfer learning, and few-shot learning.
4. Density estimation, for CDF and PDF.
5. Bayesian nonparametric models (Gaussian and Dirichlet processes).