Chapter 25

stevenjin8

July 11, 2021

Comments

My clustering experience has been very limited to non-probabilistic methods such as k-means and hierarchical clustering, so I really wanted to get a good understanding of Dirichlet Processes (DP). However, I found the notation somewhat confusing. Hopefully, if my future self ever sees DP, this little blurb will serve as a good reminder.

The confusion starts in section 25.2.2 where the author defines a Dirichlet Process as a "distribution over probability measures $G : \Theta \to \mathbb{R}^+$, where we require $G(\theta) \geq 0$ and $\int_\Theta G(\theta) = 0$." He then goes on to say that $(G(T_1), ..., G(T_K))$ has a join Dirichlet distribution." This made little sense to me because $T_i \subseteq \Theta$, not $T_i \in \Theta$. It made even less sense in equation 25.22 in the usage of the Dirac delta, since it is only relevant at one point.

What really helped me understand was learning what a measure is, and the motivation behind measures in probability theory. Following [1], we see that valid probability distributions can be quite clunky to express with pdf's and cdf's. Let A be a random variable with a support of $\{0, 1\}$ and uniform probabilities. Now let B be a random variable such that $B = 1$ if $A = 1$, but $B|A = 0 \sim \text{Unif}[0, 1]$. The marginal cdf of B has a discontinuity at 1. Thus, the pdf does not exist. In other words, despite the marginal of B being a valid random variable, its distribution cannot be expressed in terms of a pdf (very cleanly).

What we really want is an abstract function that gives us a probability for subsets of its support. More formally we want a function $G : \mathcal{A} \to [0, 1]$ such that

1. $G(\Theta) = 1$,
2. $G(S) + G(T) = G(S \cup T), S \cap T = \emptyset$.

where $\mathcal{A}$ is an algebra of Θ (or a σ-algebra if Θ is continuous). An algebra of Θ is a set of sets that contains Θ, and is closed under unions and complements. A σ-algebra is like an algebra, but it is also closed under countably finite unions (not too sure in what circumstances an algebra would not be a σ-algebra).

Applying this to section 25.2.2, I think it would be more appropriate to say that G is a probability measure over Θ. If we let I be the posterior probability measure over Θ given some observations $\overline{\theta}_1, ..., \overline{\theta}_N$, with distinct values $\theta_1, ..., \theta_K$, then we can partition Θ into $K + 1$ partitions $\{\theta_1\}, ..., \{\theta_K\}, \Theta \setminus \{\theta_1, ..., \theta_N\}$. Rewriting equation 25.27, we have

$$\begin{aligned} I(\{\theta_k\}) &= \frac{N_k}{\alpha + N} \\ I(\Theta \setminus \{\theta_1, ..., \theta_K\}) &= \frac{\alpha}{\alpha + N}, \end{aligned}$$

where N_k is the amount of times θ_k occurs in our samples.

The moral of the story is that, unlike pdf's, probability measures allow us to assign non-zero probabilities to sets of measure 0.

My final note is that in the context of sets, $\delta_x(T) = \mathbb{I}(x \in T)$.

References

[1] Evans Lawrence. *Mini Lecture #1 - Why use measure theory for probability?*. `https://www.youtube.com/watch?v=RjPXfUT7Odo`