CAP 6617 ADVANCED MACHINE LEARNING

NAME: SANKET KUMAR   UF ID: 4513-3352

SOME ASPECTS OF POLYA TREE DISTRIBUTIONS FOR STATISTICAL MODELLING

AIM: Polya tree distributions and mixtures of Polya trees are defined. Predictive and posterior distributions are explained. A canonical construction of Polya tree and choices of Polya tree parameters are discussed.

- Polya trees form a class of distributions for a random probability measure intermediate between Dirichlet process and tailfree processes. Their biggest advantage is that they can be constructed to give probability 1 to the set of absolutely continuous probability measures, whereas Dirichlet process selects a discrete probability distribution with probability 1.
- Let $E^*=\cup$ {0, 1} x {0, 1} x … x{0, 1}. A random probability measure P on a sample space is said to have a Polya tree prior P ~ PT ($\prod$, A = {a_e: e $\in$ E*}) if there exist non-negative numbers A and random variables Y, where $\prod$ is a set of partitions and A is a set of measures on the entire space, such that 1) all random variables in Y are independent. 2) For every $e \in E^*$, Y_e has a beta distribution with parameters a_{e0} and a_{e1}. 3) For every $e \in$ set of infinite sequences of elements of {0, 1}, $P(B_{e1..em})$ is a product of $Y_\varnothing$ and 1-$Y_\varnothing$ where B is a separating binary tree of partitions of the sample space. $Y_\varnothing$ and 1-$Y_\varnothing$ are the probabilities that $\Theta_i \in B_0$ and $\Theta_i \in B_1$. Furthermore, Y_e and 1-Y_e are conditional probabilities that $\Theta_i \in B_{e0}$ and $\Theta_i \in B_{e1}$ given that $\Theta_i \in B_e$.
- In a Polya tree, $Y_\varnothing$ has a Beta distribution. When Θ_1 is observed, so is the truth of $\Theta_1 \in B_0$. Therefore, the conditional distribution of $Y_\varnothing$ given Θ_1 is a Beta distribution in which one of the parameters have been incremented by 1. So, as we traverse down the tree, we add 1 to every ae for the appropriate Θ_i. A Polya tree prior on (0,1] can be constructed using partitions made up of dyadic intervals. The set (0, $1/2^n$] decrease to $\varnothing$. The probability of nth such set is Y x Y_0 x … x $Y_{00\ldots0}$. If this product does not converge to 0 then P will not be continuous and hence not countably additive. If we want P to be countably additive with probability one, then we must choose the a_e's so that the products converge to 0 with probability one.
- For updating predictive densities, we begin with predictive density of Θ_1 and proceed stepwise, always finding predictive density of the next observation given all the previous ones. Because Polya trees are conjugate, it suffices to find the predictive density of Θ_2 given Θ_1.
- The predictive density after one observation is a piecewise rescaled version of the original predictive density. There is a separate rescaling factor for each B_e containing θ_1 where $\theta_1 \in \Theta_1$; this is how $\prod$ plays the role. Although the predictive density exists, it can be discontinuous, with infinitely many discontinuities in every neighborhood of θ_1.
- For canonical construction of Polya tree, we set Θ_1~Q, where Q is specified in advance and continuous. We choose B_0 and B_1 to satisfy $Q(B_0) = Q(B_1) = 1/2$. Then for every $e \in E^*$, choose B_{e0} and B_{e1} to satisfy $Q(B_{e0}|B_e) = Q(B_{e1}|B_e) = 1/2$. Any choice of A satisfying $a_{e0} = a_{e1}$ will satisfy Θ_1~Q. When P initially has a Dirichlet process distribution, although the posterior distribution is a single Polya tree, it is a mixture of Dirichlet processes.
- a_e controls how quickly the updated predictive distribution moves from the prior to sample distribution. If a_e's are large, then the distribution is close to prior, otherwise it is close to the sample. ae also expresses a belief about smoothness of P. Heuristically, if $a_{e0} = a_{e1}$ is large, then $P[B_{e0}|B_e]$ has a distribution tightly concentrated around 1/2. Now the variance of Y_e is $1/(4(2a_{e0}+1))$. Therefore, as long as a_e's increase rapidly with m, P will be absolutely continuous. a_e also controls how closely P is concentrated about its mean.
- P is said to be a mixture of Polya trees if there is a random index variable U, with mixing distribution H, and Polya tree parameters {$\prod_u$, A_u} such that [P|U=u] ~ PT($\prod_u$, A_u). A single Polya tree is a mixture of Polya trees where H is degenerate. When Θ_1 is observed, each A_u must be updated along with H. If the partition elements are different in each $\prod_u$ and if H(u)=0 for all u, then the partition effects get smoothed out and updated predictive densities can be continuous. Mixture of Polya trees can be useful when a standard Bayesian analysis is suspect because the family of sampling densities is not known exactly. The data are modelled as though u is chosen according to h(u); then P is chosen according to PT($\prod_u$, A_u); then θ_1, θ_2,… are chosen according to P.
- **Example cases of Polya trees are:** 1) Estimation of distribution function. 2) Estimation of mean. 3) Lifetimes of spherical pressure vessels.