CAP 6617 ADVANCED MACHINE LEARNING

NAME: SANKET KUMAR UF ID: 4513-3352

A BAYESIAN ANALYSIS OF SOME NON-PARAMETRIC PROBLEMS

AIM: This paper presents a class of prior distributions called Dirichlet process priors for which treatment of many nonparametric statistical problems may be carried out, yielding results that are comparable to the classical theory.

Two desirable properties of a prior distribution for nonparametric problems are:

- The support of the prior distribution should be large with respect to some topology on the space of probability distributions on the sample space.
- Posterior distributions given a sample of observations from the true probability distributions should be manageable analytically.

The Dirichlet distribution

The Dirichlet distribution is known to Bayesians as the conjugate prior for the parameters of a multinomial distribution. Unlike normal distributions sampled over real space, Dirichlet distributions are sampled over probability simplex. Gamma distribution G(a,b) has a p.d.f with respect to Lebesgue measure on the real line for a>0 given by:
$F(z|a,b)=(1/\Gamma(a)b^a)e^{-z/b}z^{a-1}I_{(0,\infty)}(z)$ where $I_S(z)$ is the indicator function of set S. Let $Z_j \in G(a_j,1)$ be independent random variables, then the Dirichlet distribution with parameter$(a_1 \ldots a_k)$ denoted by $D(a_1 \ldots a_k)$ is defined as the distribution of $(Y_1 \ldots Y_k)$, where $Y_i=Z_j/\sum_{i=1}^{k} Zi$ for j=1,2,3…k. If $a_j>0$ for all j, the (k-1)-dimensional distribution of $(Y_1 \ldots Y_{k-1})$ is absolutely continuous with density $f(y_1 \ldots y_{k-1}|a_1 \ldots a_k)= (\Gamma(a_1+\ldots+a_k)/ \Gamma(a_1) \ldots \Gamma(a_k))(\prod y_j^{aj-1})(1-\Sigma y_j)^{ak-1} I_S(y_1 \ldots y_k)$ where simplex S is the simplex. For k=2, the density function reduces to Beta distribution denoted by $Be(a_1,a_2)$.

The Dirichlet process

The Dirichlet process is also a distribution over distributions. Let's define random probability P on (X,A), where X is a set and A is sigma field of subsets of X, by defining joint distributions of random variables$(P(a_1),\ldots,P(a_m))$ for every m and every sequence $a_1,\ldots,a_m$ of measurable sets($a_i \in A$). We then verify the Kolmogorov consistency conditions to show there exists a probability P on $([0,1]^A,BF^A)$ yielding these distributions. Given arbitrary measurable sets $a_1,\ldots,a_m$ we form 2^m sets by taking intersections of a_i and their complements: $B_{v1\ldots vm}=\cap a_j^{vj}$. If we are given the joint distribution of $\{P(B_{v1\ldots vm});v_j=0 \text{ or } 1,j=1\ldots m\}$ then we may define joint distributions of $(P(a_1)\ldots P(a_m))$ as $P(a_i)=\Sigma P(B_{v1\ldots vm})$. We note that if $(a_1,\ldots a_m)$ was a measurable partition to start with, then this does not lead to contradictory definitions of the distribution of $(P(a_1)\ldots P(a_m))$ provided $P(\varnothing)$ is degenerate at 0. There is one consistency constraint we would like to have satisfied: If $(B_1'\ldots B'_{k'})$ and $(B_1\ldots B_k)$ are measurable partitions, and if $(B_1'\ldots B'_{k'})$ is a refinement of $(B_1\ldots B_k)$ with $B_1=\cup B_i'$, $B_2=\cup B_i'\ldots B_k=\cup B_i'$ then the distribution of $(\sum_1^{r1} \mathrm{P(Bi')}, \sum_{r1+1}^{r2} \mathrm{P(Bi')},\ldots \sum_{rk-1+1}^{k'} \mathrm{P(Bi')})$ as determined from the joint distributions of $(P(B_1')\ldots P(B'_{k'}))$ is identical to the distribution of $(P(B_1),\ldots P(B_k))$. Thus we say that P is a random probability measure on (X,A) if the above constraint is satisfied, if P(a) takes values only in [0,1] and P(X) is degenerate at 1.
If a system of joint distributions of $(P(B_1),\ldots P(B_k))$ for all k and measurable partitions $(B_1\ldots B_k)$ is defined satisfying the above constraint, and if for arbitrarily measurable sets $a_1\ldots a_m$, then there exists a probability P on $([0,1]^A,BF^A)$ yielding these distributions since $X\cup\varnothing=X$, it follows from the above constraint that $P(\varnothing)$ is degenerate at 0 and thus the distribution of $(P(a_1),\ldots,P(a_m))$ is well defined.
The consistency constraint is the property that $(Y_1\ldots Y_k) \in D(a_1\ldots a_k)$ and $0<r_1<\ldots<r_l=k$ then$(\sum_1^{r1} \mathrm{Yi}\ldots \sum_{1rl-1+1}^{rl} \mathrm{Yi}) \in D(\sum_1^{r1} \mathrm{ai}\ldots \sum_{1rl-1+1}^{rl} \mathrm{ai})$. In addition, since P(X) is degenerate at 1, we call P a random probability measure. Three propositions show a close relationship between properties of P and a:
Let P=Dirichlet process on (X,A) with parameter c, and $a \in A$. 1) If c(a)=0, then P(a)=0 with probability one and if c(a)>0, then P(a)>0 with probability one, since it is seen that P(a) has a beta distribution $Be(c(a),c(a^c))$ 2) If c is sigma-additive, then so is P in the sense that for a fixed decreasing sequence of measurable sets, we have $P(a_n)->0$ with probability one since measurable sets and c is additive, $c(a_n)->0$. 3) Let Q be a fixed probability measure on (X,A) with Q<<c. Then for any + integer m and measurable sets $a_1,\ldots a_m$ and e>0, $P'\{|P(ai)-Q(ai)|<e \text{ for } i=1\ldots m\}>0$.
Let P be a random probability measure on (X,A). We say that $x_1\ldots x_n$ is a sample of size n from P if for any m=1,2,… and measurable sets $a_1\ldots a_n$, $c_1\ldots c_n$. We give a proposition that $P'(x \in a)=c(a)/c(X)$ for $a \in A$ since $P'(x \in a)=EP'(x \in a|P(a))=EP(a)= c(a)/c(X)$.
An important theorem states that the conditional distribution of P given $x_1\ldots x_n$ is a Dirichlet process with parameter $c+\sum_1^n \mathrm{dxi}$. It is sufficient to prove the theorem for n=1 since we can complete the rest of the proof by induction. Let $(B_1\ldots B_k)$ be a measurable partition of X and let $a \in A$. The marginal distributions of a process are identical to the conditional distributions of the marginals. To show that the conditional distribution is $D(y_1\ldots y_k|c(B_1)+d(B_1)\ldots c(B_k)+d(B_k))$, we compute:

$$\begin{aligned}\int_A D(y_1, \cdots, y_k \,|\, \alpha(B_1)+\delta_s(B_1), \cdots, \alpha(B_k)+\delta_s(B_k))\, d\alpha(x)/\alpha(\mathscr{X}) \\ = \sum_{j=1}^{k} \int_{B_j \cap A} D(y_1, \cdots, y_k \,|\, \alpha_1^{(j)}, \cdots, \alpha_k^{(j)})\, d\alpha(x)/\alpha(\mathscr{X}) \\ = \sum_{j=1}^{k} \frac{\alpha(B_j \cap A)}{\alpha(\mathscr{X})} D(y_1, \cdots, y_k \,|\, \alpha_1^{(j)}, \cdots, \alpha_k^{(j)}),\end{aligned}$$

An alternative definition of the Dirichlet process

The aim is to define a random probability measure which is a Dirichlet process on (X,A) with parameter c with probability one is a discrete probability measure on (X,A). Since the Dirichlet distribution is definable as the joint distribution of a set of independent gamma variables divided by the sum, so also should the Dirichlet process be definable as a gamma process with independent increments divided by the sum. Using a representation of the process with independent increments as a sum of a countable number of jumps of random height at a countable number of random points, we may divide by the total heights of the jumps and obtain a discrete probability measure, which should be distributed as a Dirichlet process.
The random probability measure defined by $\mathrm{P(a)}=\sum_1^{r1} \mathrm{P_j d_{vj}(a)}$ is a Dirichlet process on (X,A) with parameter c. $\mathrm{(P(B_1)\ldots P(B_k))}=(1/Z_1) \sum_{j=1}^{\infty} \mathrm{J_j(d_{vj}(B_1), \ldots d_{vj}(Bk))}$ where $M=(B_1\ldots B_k)$ is a measurable partition of X. $d_{vj}(B_1), \ldots d_{vj}(Bk)$are iid random vectors having a multinomial distribution with probability vector $(Q(B_1)\ldots Q(B_k))$. Hence distribution of $\sum_1^{\infty} \mathrm{J_j M_j}$ must be same as of $(Z_{1/k},Z_{2/k}-Z_{1/k}\ldots Z_1-Z_{(k-1)/k})$ where Z_t is a gamma process given by $\mathrm{Z_t}= \sum_{j=1}^{\infty} \mathrm{J_j I_{[0,\,G(t)]}(U_j)}$ where $U_1,U_2\ldots$ are iid variables, uniformly distributed on [0,1] and independent of $J_1,J_2\ldots$, with G(t) chosen so that $G(j/k)-G((j-1)/k)=Q(B_j)$ where j=1…k.

Applications: Certain applications include: 1) Estimation of distribution function 2) Estimation of mean and median 3) Estimation of quantiles 4) Estimation of variance and covariance 5) Estimation of $\int FdG$ for a 2-sample problem.