Quick review of some useful discrete stuff:

β- distribution

for $x \in [0,1]$

$$P_{\beta}(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{(\alpha-1)} (1-x)^{(\beta-1)}.$$

In this case:

$$E[X] = \frac{\alpha}{\alpha+\beta}$$ (worth remembering)

β is useful for binomial problems because it has nice conjugacy properties.

Eg: I have a coin, $P(\text{heads}) = h$ (Unknown)

I toss, see k heads, $n-k$ tails.

What is $P(h \mid k, n-k)$?

<u>Natural choice</u>: $P(h) = \beta(\alpha_\pi, \beta_\pi)$

$\uparrow$
prior

2

then $P(h \mid k, n-k) \propto P(K, n-k \mid h)\, P(h)$

binomial

$\{\cdot h^{K}(1-h)^{n-k}$

β

$h^{\alpha_\pi - 1} \cdot (1-h)^{\beta_\pi - 1}$

So product

$$\propto h^{(\alpha_\pi + K - 1)} (1-h)^{(n-K+\beta_\pi - 1)}$$

which is $\beta(\alpha_\pi + K,\ \beta_\pi + n - K)$

This is one of those cute things one learns and forgets, but there is a more general point

<u>Dirichlet</u>

assume we have

$x_1 \cdot\cdot x_K$

on a <u>simplex</u>

i.e. $0 < x_i < 1$

$x_1 + \cdot\cdot x_K = 1$

and $\alpha_1 \cdot\cdot \alpha_K > 0$

Then

$$\underset{\text{Dirichlet}}{P}(x_1 \cdots x_K \mid \alpha_1 \cdots \alpha_K) = \frac{1}{B(\alpha)} \prod_i x_i^{(\alpha_i - 1)}$$

(Notice how this generalizes β dist – its a PDF on Prob distributions)

Again, there are neat conjugacy properties

Eg: Assume I roll a K-sided die N times, observing faces $N_1 \cdots N_K$ times.

~~I cont~~ What is $P(p_1 \cdots p_K \mid n_1 \cdots n_K)$?

Bayesian: Prior on $p_1 \cdots p_K$ is $\text{Dir}(\alpha_1 \cdots \alpha_K)$

then $P(p_1 \cdots p_K \mid n_1 \cdots n_K) \propto P(n_1 \cdots n_K \mid p_1 \cdots p_K) P(p_1 \cdots p_K)$

multinomial

$\sim p_1^{n_1} p_2^{n_2} \cdots p_K^{n_K}$

$p_1^{\alpha_1 - 1} \; p_2^{\alpha_2 - 1} \cdots$

This gives

$$P(p_1 \cdot p_K \mid n_1 \cdot n_K) \quad \text{is} \quad Dir(\alpha_1 + n_1, \cdot \cdot \alpha_K + n_K)$$

Sometimes written as

$\alpha = (\alpha_1 \cdot \alpha_K)$ concentration hyperparameter.

prob a face comes up $p \mid \alpha \sim Dir(K, \alpha)$

$X \mid p \sim Cat(K, p)$

multinomial

observed faces

then $c = (c_1 \cdot c_K)$ = counts of cat $1 \cdot K$

$$p \mid X, \alpha \sim Dir(K, \alpha_1 + c_1 \cdot \cdot \alpha_K + c_K).$$

Now, one fair objection to previous methods is that we have no probabilities

- we can <u>bound</u> expected counts below, but not compute expectations.

Consider

$$p \sim Dir(\alpha_1 \cdots \alpha_K)$$

$$E(p) = \frac{(\alpha_1, \cdots \alpha_K)}{\sum_{j=1}^{K} \alpha_j}$$

(which is a discrete probability distribution, as it should be)

Now assume we do not know anything about which face of die is more likely

→ suggests a prior

$Dir(\alpha)$ where $\alpha = \alpha_0 \cdot \underline{1}$

(i.e. all α's the same).

In each case, mean will be uniform pd.

but

$$\mathrm{Dir}(\alpha) \quad \propto \quad x_1^{\alpha-1} \cdots x_K^{\alpha-1}$$

So for very large α, few x_i's can be large (they can't all be smaller than $\frac{1}{K}$!)

- concentration parameter

α big $\Rightarrow$ concentrated dist's
small $\Rightarrow$ more diffuse.

To compute expectations, we need a random model of $CdG(\lambda)$.

<u>Desirable object</u>:

- a way of constructing <u>random</u> probability distributions, so that we can still compute expectations, etc.

There are several such objects

- <u>Dirichlet process</u>:

X – a Polish space

= separable, completely metrizable topological space.

(<u>examples</u>: Real line, Interval, etc)

= space homeomorphic to a complete metric space w/ a countable dense subset:

α . a finite measure.

i.e. $\alpha(\mathcal{X}) < \infty$.

P - a <u>random</u> measure.

i.e. "randomly chosen" and a measure.

then P is a Dirichlet process if for every finite measurable partition $\{B_1 \cdots B_K\}$ of $\mathcal{X}$,

the <u>joint</u> of $\{P(B_1), \cdots P(B_K)\}$ is a K-dimensional Dirichlet dist with params.

$$\alpha(B_1), \quad \cdots \quad \alpha(B_K)$$

α is called the <u>base measure</u>

Notice the property is quite strong; isn't obvious one exists.

Notice also that "big" sets in α ~~get~~ ~~are~~ tend to be "big" in P.

$\boxed{\text{Work on } \mathbb{R}}$

<u>Some properties:</u>

- consider a partition $\{A, A^c \equiv \mathbb{R} - A\}$

$$P(A) \text{ is } \beta(\alpha(A), \alpha(A^c)).$$

random variable — distribution

So $$E[P(A)] = \frac{\alpha(A)}{\alpha(A) + \alpha(A^c)}$$

Now write $\alpha = M \cdot G$

M: $\alpha(\mathbb{R})$ — G: prob measure

$$\therefore E[P(A)] = G(A)$$

Now assume P is a D.P.

and

$$\underbrace{X \mid P \sim P}$$

You can think of P as a prior $P(X \mid P) = P$

then the <u>marginal</u> of X is

G_k ← this is a <u>Distribution</u>

[Show this from the expectations]

Now recall $P(A) \sim \beta(\alpha(A), \alpha(A^c))$.

Variance of $\beta(\alpha, \beta) = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

So

$$\mathrm{Var}\big(P(A)\big) = \frac{\alpha(A)\alpha(A^c)}{M^2(M+1)} = \frac{G(A)G(A^c)}{(M+1)}$$

$\Rightarrow$ bigger M implies "more concentration".

Notice that easy arguments give

$$E\left[\int \Psi \, dP\right] = \int \Psi \, dG$$

from above. This is useful, because it allows us to think about integrals, as above.

P is distributed as D.P. with base measure $\alpha = MG$, write $P \sim DP(M, G)$

<u>Conjugacy</u>:

assume $X_1 \cdots X_n \mid P \sim P$, IID

$$P \sim DP(M, G) = D_\alpha$$

then

$$P \mid X_1 \cdots X_n \sim D_{\alpha + \sum_{i=1}^{n} \delta(X_i)}$$

Interpret

$$D_{\alpha + \sum_i \delta(x_i)} \quad \text{as another DP,}$$

with base measure $\alpha + \sum_i \delta(X_i)$

δ functions at X_i

This must be true, because:
(weaker than proof!)

Consider partition $(A_1 \cdots A_r)$

and counts $n_1 \cdots n_r$

$$n_j = \sum_i \mathbb{1}(X_i \in A_j)$$

$P(A_1) \cdots P(A_r)$ is Dirichlet; prior is ~~$\alpha(A_1)$~~

$$Dir(r, \alpha(A_1) \cdots \alpha(A_r))$$

$\therefore$ posterior must be

$$Dir(r, \alpha(A_1)+n_1 \cdots \alpha(A_r)+n_r)$$

And this must work for all partitions

In turn,

$$E(P \mid X_1 \cdots X_n) = \frac{M}{M+n} G + \frac{n}{M+n} \cdot \mathbb{P}_n$$

empirical dist, 1δ at each data point.

It can be proven that:

(a) samples from the DP are discrete measures, with prob 1

(b) Dirichlet processes have nice self similarity properties

let A be some set st
$0 < G(A) < 1$

$P|_A$ is the restriction of P to A

i.e. $P|_A(B) = \frac{P(A \cap B)}{P(A)}$

Now $P|_A$ is $DP(MG(A), G|_A)$

$P|_{A^c}$ is $DP(MG(A^c), G|_{A^c})$

(You can see this by thinking about marginals.)

and they are independent.

<u>AND</u>

$P|_A$ is independent of $P(A)$

So this means that, for any given set A

- how mass is dist. in A
- " " " " in A^c
- the total mass of A

are independent.

Predictive distributions from DP.

consider $\quad P \sim DP$

$$X_1 \cdots X_n \sim P \quad \text{IID}$$

1) $X_1 \sim G$ (~~by exp~~ marginal, from expectation result.)

2) $X_2 \mid P, X_1 \sim P$

and $$P \mid X_1 \sim DP\left(M+1, \frac{M}{M+1} G + \frac{1}{M+1} \delta X_1\right)$$

(from result above on DP and data)

Now Take expectation OVER P :

$$X_2 \mid X_1 \sim \frac{M}{M+1} G + \frac{1}{M+1} \delta X_1$$

i.e.

- with prob $\frac{1}{M+1}$, X_2 is duplicate of X_1
- otherwise, it's New.

We can extend this argument

$X_n | X_1 \cdots X_{n-1}$

(we will have n_i copies of θ_i where θ_i is one of the <u>distinct</u> values)

So

$$X_n | X_1, \cdots X_{n-1} \sim \begin{cases} \delta\theta_i & \text{prob } \frac{n_i}{M+n-1} \\ G & \text{prob } \frac{M}{M+n-1} \end{cases}$$

Now p(new value) at Step 1 is 1

2 $\frac{M}{M+1}$

3 $\frac{M}{M+2}$

n $\frac{M}{M+n-1}$

So write K_n for number of <u>distinct</u> values

$$E(K_n) = \sum_{i=1}^{n} \frac{M}{M+i-1}$$

This looks like $M \log \frac{n}{M}$ as $n \to \infty$

(i.e. few new things.
- <u>great as a clustering prior</u>.
- not like words, objects)

<u>Stick breaking repn.</u>

Consider $\theta_i \sim G$ IID.
$Y_i \sim \beta(1, M)$ IID

write

$$P = \sum_{i=1}^{\infty} V_i \, \delta\theta_i$$

where $V_1 = Y_1$, $V_2 = (1-Y_1)\, Y_2$, $V_3 = (1-Y_1)(1-Y_2)\, Y_3$

$$V_i = Y_i \prod_{j=1}^{i-1} (1-Y_j)$$

Stick breaking because Y_1 breaks off part of stick, Y_2 breaks off part of remains $[(1-Y_1)]$, etc.

notice this has a <u>recursive</u> form

$$P =_d Y_1 \delta\theta_1 + (1-Y_1) P$$

sample from DP(M, G)   Sample from DP(M, G)

This makes it quite easy to prove some things

1) $E\left(\int \psi \, dP\right) = \int \psi \, dG$

$$\int \psi \, dP = Y_1 . \psi(\theta_1) + (1-Y_1) \int \psi \, dP$$

so $E\left(\int \psi \, dP\right) = E\left(Y_1 \psi(\theta_1)\right) + E\left((1-Y_1) \int \psi \, dP\right)$

<u>But</u>: Y_1, θ_1 are <u>indep</u>

$(1-Y_1), \int \psi \, dP$ are <u>indep</u>.

So

$$E\left(\int \psi \, dP\right) = E(Y_1)\, E\left(\psi_t(\theta_1)\right) + E(1-Y_1) \,.\, E\left(\int \psi \, dP\right).$$

$Y_1 \sim \beta(1, M)$ so $E(Y_1) = \dfrac{1}{M+1}$

$$E(1-Y_1) = \frac{M}{M+1}$$

So

$$\left(1 - \frac{M}{M+1}\right) E\left(\int \psi \, dP\right) = \frac{1}{M+1} \, E\left(\psi_1(\theta_1)\right)$$

$\uparrow$ $\int \psi \, dG$ by weak law of large nums.

So $E\left(\int \psi \, dP\right) = \cancel{k} \int \psi \, dG$

a similar argument yields $\mathrm{Var}\left(\int \psi \, dP\right)$.

Now assume we have

$$G \sim DP(\alpha, H) = DP_{\alpha H}$$

αH is base measure $H(A) = 1$ is prob. dist.

we know $G | X_1 \ldots X_n \sim DP\left(\alpha + n, \dfrac{\alpha H + \sum_i \delta X_i}{\alpha + n}\right)$

from earlier arguments.

<u>Another representation</u>:

consider $G|_{X_1 \ldots X_n} = P$

~~$P(X_1)$~~ there are K unique X_i, $K < n$

call these $X_1^* \ldots X_K^*$

then $P(X_1^*), P(X_2^*), \ldots, P(X_K^*), P(A - \sum_i X_i^*)$

$$\sim Dir\left(n_1, \ldots n_K, \left[\frac{\alpha H(A - \sum_i X_i^*)}{\alpha + n}\right] \alpha + n\right)$$

$$= Dir(n_1, \ldots n_K, \alpha).$$

This means we can represent

$$G | X_1 \cdots X_n \quad \text{as}$$

$$\sum_{i=1}^{K} p_i \, \delta X_i^* + p_{K+1} G^*$$

where

$$(p_1 \quad \cdots \quad p_{K+1}) \sim \mathrm{Dir}(n_1 \cdots n_K, \alpha)$$

and $G' \sim DP(\alpha, H)$.

a DP won't really do the work for us.

~~· 2 Parameter~~

· Richer family of random processes

· 2 parameter Poisson - Dirichlet process

OR. Pitman · Yor process

<u>Choose</u>:

Base PDF G —

$0 < \sigma < 1$

$\theta > -\sigma$

] - parameters

Now choose $x_i \overset{IID}{\sim} G$

$$V_1 \sim \beta(1-\sigma, \theta + \sigma)$$

$$V_i \sim \beta(1-\sigma, \theta + i\sigma)$$

$$P = \sum_{i=1}^{\infty} \tilde{p}\, \delta(x_i) \qquad \text{where } \tilde{p} = V_1$$

$$\tilde{p}_i = V_i \prod_{j=1}^{i-1} (1 - V_j) \qquad i \geq 2$$

This object P is a sample from a Pitman-Yor process.

Notice $\tilde{P}_i$ is a series

$$1 \geq \tilde{P}_i \geq 0 \quad , \quad \sum_{i=1}^{\infty} \tilde{P}_i = 1 \quad a.s$$

Proving facts about P involves manipulating PDF's on such sequences (which is hard).

- State results, but do not prove them.

$$\cancel{X_{n+1} \mid X_1 \cdots X_n =} \Bigg\{$$

- assume we see $X_1 \cdots X_n$
- notice we should have duplicates
- write $X^*_{n,j}$ for jth distinct of k species
- $n_{n,j}$ for # of this type

Then

a new one

$$X_{n+1} \mid X_1 \cdots X_n = \begin{cases} x_i & \text{with prob } \dfrac{(\theta + \sigma k_n)}{(\theta + n)} \\ X^*_{n,j} & \text{" } \dfrac{(n_{n,j} - \sigma)}{(\theta + n)} \end{cases}$$

- Notice
  - rich get richer
    i.e. if you've seen many $X^*_{n,j}$, you'll see more.
  - prob of seeing a new species is high if k_n similar to n

(Not sure how to prove this - ~~proof~~ is associated w/ Pitman + Yor)

Notice also this gives us a coverage result.

Coverage

$= \%$ of species rep.

$= 1 -$ (prob next one will be new)

$$= 1 - \left(\frac{\theta + \sigma k_n}{\theta + n} \right)$$

This could be quite high eg: Shakespeare Data.

$$\frac{\theta + \sigma(37k)}{\theta + \cancel{\sigma}(885k)}$$

$\theta = 1$ ~ 0.05

$\theta = 10^6$ ~ 0.5

θ matters!

22

Now consider drawing n values from

$$p \sim PD(\sigma, \theta).$$

what is <u>posterior</u> on P ?

assume $k \leq n$ distinct values

$X_1^* \cdot \quad X_k$

with $n_1 \cdot \cdot \quad n_k$ ~~freqs~~. counts.

Then

$$p \mid X \overset{d}{=} \sum_{j=1}^{k} p_j^* \delta_{X_j^*} + \left(1 - \sum_{j=1}^{k} p_j^*\right) \tilde{p}^{(k)}$$

there are errors here!

see 22 a

where $\cancel{p_1^* \cdots p_j^* \sim \mathrm{Dir}\left(\frac{n_1-\sigma}{\theta+k\sigma}, \cdots \frac{n_k-\sigma}{\theta+k\sigma}\right)}$

and $\cancel{\tilde{p}^{(k)} \text{ is } PD(\sigma, \theta+k\sigma)}$

$$P \mid X \stackrel{d}{=} \sum_{j=1}^{k} P_j^* \delta_{X_j^*} + R^* \tilde{P}^{(K)}$$

where $P_1^*, \ldots, P_K^*, R^* \sim \mathrm{Dir}(n_1 - \sigma, \ldots, n_K - \sigma, \theta + K\sigma)$

and $\tilde{P}^K$ is $PD(\sigma, \theta + K\sigma)$.

One way to view this model is as a probability distribution on partitions.

Partition of $n = \sum_{i=1}^{K_n} N_{i,n}$ things into K_n partitions each containging $N_{i,n}$ things

$$\underset{N_{1,n}}{(\qquad)}\underset{N_{2,n}}{(\qquad)}\cdots\underset{N_{K_n,n}}{(\quad)}$$

Notice previous results give ~~P(new par~~ expressions for where a new element goes, given an existing partition.

From these, we can derive

$$P\left(K_n = k,\ N_{1n} = n_1 \cdots N_{K_n n} = n_K\right)$$

$$= \frac{\prod_{i=1}^{K-1} (\theta + i\sigma)}{(\theta+1)_{n-1}} \cdot \prod_{j=1}^{K} (1-\sigma)_{n_j - 1}$$

Using the notation

$$a_b = a \cdot (a+1) \cdots (a+b-1)$$

and.

$$a_0 = 1$$

This yields a strategy for estimating θ, σ

$$\max P\left(K_n = k, N_{1n} = n_1 \cdots N_{K_n} = n_K \,\middle|\, \theta, \sigma\right)$$

(max likelihood)

→ should give large θ for many small categories