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(heads) $= h$ (unknown)

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

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

Natural choice: $\underset{\text{prior}}{\underset{\uparrow}{P(h)}} = \beta(\alpha_\pi, \beta_\pi)$

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

Dirichlet

assume we have

$$x_1 \cdot\cdot\ x_K$$

on a simplex

i.e. $0 < x_i < 1$

$x_1 + \cdot\ \cdot\ x_K = 1$

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

Then

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

(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 dice N times, observing faces $N_1 \cdots N_K$ times.

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

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

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

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

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

This gives

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

Sometimes written as

$$\alpha = (\alpha_1 \cdot \alpha_K) \quad \text{concentration hyperparameter.}$$

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

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

multinomial

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

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

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

$\rightarrow$ 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>:

$\mathcal{X}$ – a Polish space

= separable, completely metrizable topological space.

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

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

α - a finite measure.

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

P - a random measure.

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

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

the joint of $\{P(B_1), \ldots P(B_K)\}$ is a K-dimensional Dirichlet dist with params.

$$\alpha(B_1), \ldots \alpha(B_K)$$

α is called the base measure

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.

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

Some properties:

- 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 $\alpha = M . G$

prob measure

$\alpha(\mathbb{R})$

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

Now assume P is a D.P.

and

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

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

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

G ← 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

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

⇒ bigger M implies "more concentration".