Processo Dirichlet

Paulo C. Marques F.

Seminário relâmpago ministrado no Insper

8 de Abril de 2016

Mistura de normais

- Considere um modelo de mistura de normais em que

$$x_1, \ldots, x_n \in \mathbb{R}$$

são condicionalmente independentes e identicamente distribuídas, dados (w, μ, σ^2), tais que

$$f(x \mid w, \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i \mid w, \mu, \sigma^2) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j \, \phi(x_i \mid \mu_j, \sigma_j^2),$$

em que

$$\phi(x_i \mid \mu_j, \sigma_j^2) = \frac{e^{-(x_i^2 - \mu_j)^2 / 2\sigma_j^2}}{\sqrt{2\pi}\sigma_j}.$$

- Note que a função de verossimilhança $L_x(w, \mu, \sigma^2) = f(x \mid w, \mu, \sigma^2)$ tem k^n termos.

Variáveis latentes (1)

- Neste modelo de mistura é possível "aumentar os dados" (Tanner e Wong) introduzindo n variáveis latentes $z_i \in \{1, \ldots, k\}$ que dizem a qual componente da mistura cada um dos x_i's pertence.
- Formalmente, se postularmos que $f(x_i \mid z_i, \mu, \sigma^2) = \phi(x_i \mid \mu_{z_i}, \sigma^2_{z_i})$ e $f(z_i \mid w) = w_{z_i}$, usando o teorema da probabilidade total e a regra do produto, temos que

$$
\begin{aligned}
f(x \mid w, \mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i \mid w, \mu, \sigma^2) = \prod_{i=1}^{n} \sum_{z_i=1}^{k} f(x_i, z_i \mid w, \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \sum_{z_i=1}^{k} f(x_i \mid z_i, w, \mu, \sigma^2) f(z_i \mid w, \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \sum_{z_i=1}^{k} w_{z_i} \, \phi(x_i \mid \mu_{z_i}, \sigma^2_{z_i}) = L_x(w, \mu, \sigma^2).
\end{aligned}
$$

Infinitas componentes (1)

- Esta nova representação do modelo possibilita, sob certas condições, o cálculo das condicionais completas necessárias para a construção de um Gibbs sampler.
- As marginais *a posteriori* dos z_i's permitem uma análise de clusters no contexto deste modelo de mistura em que k é conhecido.
- É possível estender o modelo para o caso em que temos infinitas componentes:

$$f(x_i \mid w, \mu, \sigma^2) = \sum_{j=1}^{\infty} w_j \, \phi(x_i \mid \mu_j, \sigma_j^2).$$

- Para criar alguma intuição do caso infinito, suponha que alteramos os rótulos das componentes da mistura de modo a ordenar os pesos $w_1 \geq w_2 \geq w_3 \geq \ldots$.

Infinitas componentes (2)

- Lembrando que $w_j \geq 0$ e que $\sum_{j=1}^{\infty} w_j = 1$, a partir de um certo k os componentes da mistura passam a ser "irrelevantes" (não têm peso apreciável).
- Informalmente, para o referido k, teríamos que $\sum_{j=1}^{k} w_j \approx 1$.
- Formalmente, no modelo em que os rótulos das componentes da mistura foram redefinidos de modo a termos os pesos em ordem crescente, para todo $\epsilon > 0$, existe um $k = k(\epsilon) \geq 1$, tal que

$$\int_{\mathbb{R}} \left| \sum_{j=1}^{\infty} w_j \, \phi(t \mid \mu_j, \sigma_j^2) - \sum_{j=1}^{k} w_j \, \phi(t \mid \mu_j, \sigma_j^2) \right| dt < \epsilon.$$

- Deste modo, um modelo de mistura com infinitas componentes pode ser representado mentalmente pelo modelo finito que o aproxima com qualquer precisão desejada.

Processo Dirichlet

- Um Processo Dirichlet $G \sim \mathrm{DP}(c, G_0)$ é uma distribuição de probabilidade aleatória, em que $c > 0$ é a constante de concentração do processo e G_0 é a distribuição em que o processo está centrado, no sentido de que $\mathrm{E}[G] = G_0$.
- Do mesmo modo que uma variável aleatória $X \sim \mathrm{N}(0, 1)$ representa nossa incerteza a respeito de um número real, um Processo Dirichlet representa nossa incerteza sobre uma distribuição de probabilidade.
- Se $x_1, \ldots, x_n$ são condicionalmente IID, dado G, com distribuição G, então *a posteriori*

$$G \mid x_1, \ldots, x_n \sim \mathrm{DP}\left(c + n, \frac{n}{c+n}\hat{F}_n + \frac{c}{c+n}G_0\right).$$

- Note os compartamentos quando $c \downarrow 0$ e quando $n \to \infty$ (Bayesian Bootstrap).
- Podemos definir hierarquias $G \mid H \sim \mathrm{DP}(c, H)$, $H \sim \mathrm{DP}(d, H_0)$ etc.

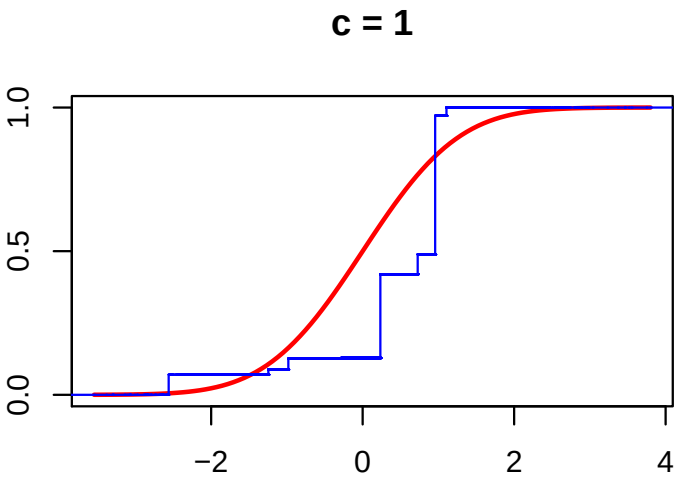
“Stick breaking” (1)

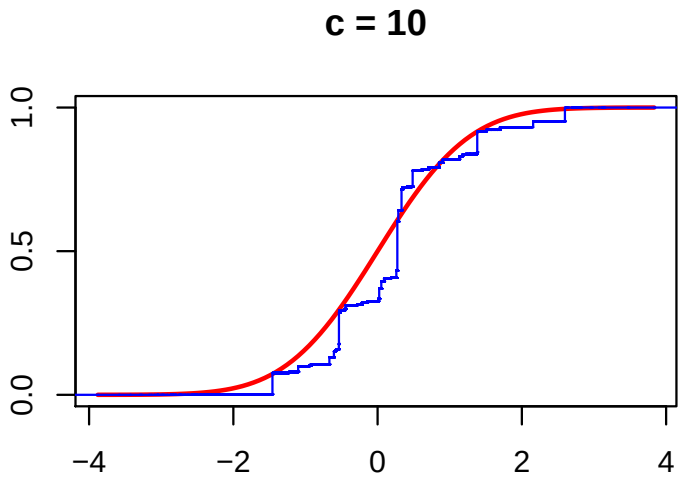
- Se pudéssemos simular um Processo Dirichlet, como seriam suas realizações?
- Blackwell: com probabilidade 1, as realizações de um Processo Dirichlet são distribuições de probabilidade discretas com suporte infinito.
- Sethuraman descobriu uma construção explícita do Processo Dirichlet, denominada “stick breaking”.
- Dadas $\{\beta_i\}_{i=1}^{\infty}$ IID com distribuição Beta$(1, c)$.
- Defina $w_1 = \beta_1$ e $w_i = \beta_i \prod_{j=1}^{i-1}(1 - \beta_j)$, para $j \geq 2$.
- Sejam os átomos $\{y_i\}_{i=1}^{\infty}$ IID com distibuição G_0.
- Supondo que o espaço amostral é a reta real, uma realização de G seria

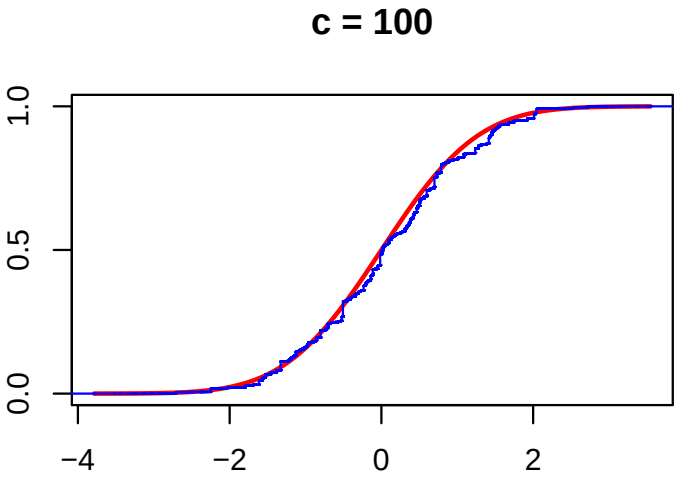
$$G(t) = \sum_{i=1}^{\infty} w_i \, I_{(-\infty, y_i]}(t).$$

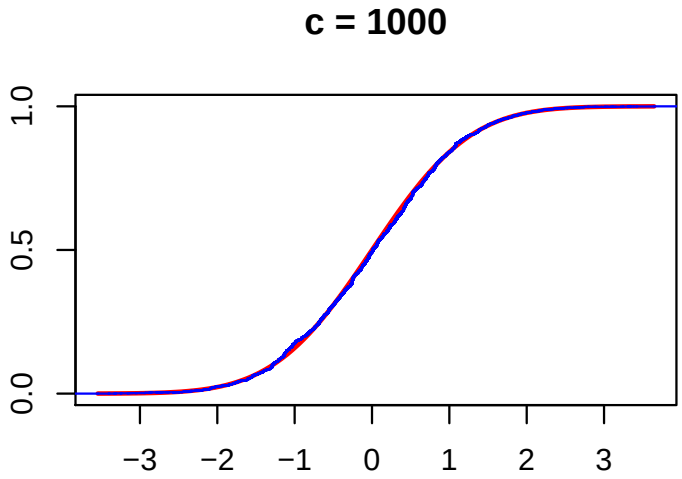
- Para representar a realização no computador, truncamos a série.

“Stick breaking” (2)









Restaurante chinês

- Há probabilidade positiva de que duas observações geradas a partir de uma realização de um Processo Dirichlet sejam exatamente iguais.
- É possível provar que

$$x_{n+1} \mid x_1, \ldots x_n \sim \frac{n}{c+n}\hat{F}_n + \frac{c}{c+n}G_0.$$

- Interpretação metafórica: um novo cliente chega ao "restaurante chinês".
- Com probabilidade $c/(c+n)$ ele "se senta" sozinho em uma "nova mesa" (um novo cluster é criado).
- Com probabilidade $(n/(c+n)) \times (n_k/n) = n_k/(c+n)$ ele "se senta" na k-ésima "mesa existente", na qual já tínhamos n_k clientes.
- Portanto, é mais provável que ele "se sente" em uma mesa em que já há muitos clientes.
- Esta é a "clustering property" do Processo Dirichlet.

Mistura via Processo Dirichlet (1)

- Um modelo de mistura via Processo Dirichlet é definido pela seguinte hierarquia.
- $G \sim \mathrm{DP}(c, G_0)$.
- $\mu_1, \ldots, \mu_n$ são condicionalmente IID, dado G, com distribuição G.
- $x_1, \ldots, x_n$ são condicionalmente independentes, dado $\mu = (\mu_1, \ldots, \mu_n)$, tais que $x_i \mid \mu_i \sim \mathrm{N}(\mu_i, \sigma_0^2 \, \mathbb{I})$.

Mistura via Processo Dirichlet (2)

- Usando a construção "stick breaking" do Processo Dirichlet e introduzindo variáveis latentes de alocação, podemos reescrever o modelo de mistura via Processo Dirichlet como um modelo de mistura de normais com infinitas componentes.
- $\beta_1, \beta_2, \ldots$ são IID com distribuição Beta$(1, c)$.
- $w_1 = \beta_1$ e $w_j = \beta_j \prod_{\ell=1}^{j-1}(1 - \beta_\ell)$, para $j \geq 2$.
- $\mu_1, \mu_2, \ldots$ são IID com distribuição G_0.
- $z_1, \ldots, z_n$ são condicionalmente IID, dado $w = \{w_j\}_{j=1}^{\infty}$, com distribuição $\Pr\{z_i = k \mid w\} = w_k$, para $k \geq 1$.
- $x_1, \ldots, x_n$ são condicionalmente independentes, dados $z = (z_1, \ldots, z_n)$ e $\mu = \{\mu_j\}_{j=1}^{\infty}$, tais que $x_i \mid z_i, \mu \sim \mathrm{N}(\mu_{z_i}, \sigma_0^2\, \mathbb{I})$, para $i = 1, \ldots, n$.
- Esta hierarquia torna possível a construção de um Gibbs sampler análogo ao utilizado no modelo de mistura tradicional.
- Os clusters são obtidos a partir das marginais *a posteriori* dos z_i's.
- Biblioteca do `R`: `dpmixsim()`.

E passamos a bola para o palestrante de hoje...