Sannolikhet och statistik A4

Johan Jonasson

Maj 2019

Exempel: En slant singlas tre ggr. Det blev krona alla tre gångerna. Vad ska vi tro om sannol att få klave θ? ML-skattningen blir 0. Verkar orimligt. Bayesianskt: antag att prior är likf, dvs $f_\theta(t) = 1$, $0 < t < 1$.

$$f_{\theta|X}(t|0) = Cf_{X|\theta}(0|t)f_\theta(t) = C(1-t)^3.$$

(Man ser snabbt att $C = 4$.) Man kan t.ex. skatta θ med väntevärdet i posterior, dvs

$$\mathbb{E}[\theta|X=0] = 4\int_0^1 t(1-t)^3 dt = \frac{1}{5}.$$

Man kan man känna igen $f_{\theta|X}(t|0)$ som en β-fördelning.
Definition: En sv Y sägs vara β-fördelad med parametrar $a > 0$ och $b > 0$ om

$$f_Y(y) = Cy^{a-1}(1-y)^{b-1}, \; 0 < y < 1.$$

Kortform: $Y \sim \beta(a,b)$.
Obs. att $\beta(1,1)$ är *likf*$(0,1)$.

Ex. ovan är specialfall av följande obs:

Låt $\theta \sim \beta(a, b)$ och sedan $X_1, X_2, \ldots, X_n$ oberoende och

$$\mathbb{P}(X_k = x) = \begin{cases} \theta, & x = 1 \\ 1 - \theta, & x = 2 \end{cases}$$

Låt n_1 vara antal k sådana att $X_k = 1$ och $n_2 = n - n_1$. Då gäller att

$$\begin{aligned} f_{\theta|X_1,\ldots,X_n}(t|x_1,\ldots,x_n) &= Ct^{n_1}(1-t)^{n_2}t^{a-1}(1-t)^{b-1} \\ &= Ct^{a+n_1-1}(1-t)^{b+n_2-1} \end{aligned}$$

$\theta|X_1 = x_1, \ldots, X_n = x_n$ är $\theta \sim \beta(a + n_1, b + n_2)$. Trevligt lätträknat!

Om istället $X \sim Bin(n, \theta)$, dyker det bara upp en faktor $\binom{n}{n_1}$ i räkningarna ovan. Den beror inte på t, så vi får att $\theta | X = n_1 \sim \beta(a + n_1, b + n_2)$ även nu.

Generalisering: Vektorn $Y = (Y_1, Y_2, \ldots, Y_{K-1})$ sägs vara Dirichletfördelad med parametrar $b_1 > 0, b_2 > 0, \ldots, b_K > 0$ om det för alla $y = (y_1, \ldots, y_{K-1})$ sådana att $y_i \geq 0$ för alla i och $y_K = 1 - \sum_{i=1}^{K-1} y_i \geq 0$ gäller att

$$f(y) = C y_1^{b_1 - 1} y_2^{b_2 - 1} \ldots y_K^{b_K - 1}.$$

Betafördelningen är specialfallet $n = 2$. Vi skriver $Y \sim Dir(b_1, \ldots, b_K)$.

Antag att $\theta = (\theta_1, \ldots, \theta_{K-1}) \sim Dir(b_1, \ldots, b_K)$ och sedan $X = (X_1, \ldots, X_n)$ oberoende och sådana att $\mathbb{P}(X_k = j) = \theta_j$, $j = 1, \ldots, K$.

$$\begin{aligned} f_{\theta|X}(t_1, \ldots, t_{K-1}|x) &= C t_1^{n_1} \ldots t_K^{n_k} \cdot t_1^{b_1-1} \ldots t_K^{b_K-1} \\ &= t_1^{b_1+n_1-1} \ldots t_K^{b_K+n_K-1}. \end{aligned}$$

Dvs $\theta|X = x \sim Dir(b_1 + n_1, \ldots, b_K + n_K)$.

Detta är ett exempel på *konjugerande prior*: om data X har en viss fördelning, $f_{X|\theta}$, och det då visar sig att posterior $f_{\theta|X}$ är av samma typ som prior f_θ, säger man att f_θ är en konjugerande prior till $f_{X|\theta}$.

Vi har alltså sett att betafördelningen är en konjugerande prior till binomialfördelningen och mer generellt att Dirichletfördelningen är konjugerande prior till multinomialfördelningen.

Exempel: Låt $\theta \sim N(0,1)$ och sedan $X \sim N(\theta,1)$.

$$f_{\theta|X}(t|x) = Ce^{-\frac{1}{2}(x-t)^2} \cdot e^{-\frac{1}{2}t^2} = Ce^{-(t-x/2)^2}.$$

Dvs $\theta|X = x \sim N(x/2, 1/2)$. Vi ser att normalfördelningen är en konjugerande prior till normalfördelningen med given varians. $\square$

Problem med nämnaren. Latent Dirichlet allocation: Låt $\theta_i \sim Dir(a_1, \ldots, a_K)$, $i = 1, \ldots, D$. Låt $\psi_i \sim Dir(b_1, \ldots, b_n)$, $i = 1, \ldots, K$. Allt oberoende. För alla i, j, $i = 1, \ldots, D$, $j = 1, \ldots, L$, låt $Z_{ij} \sim \theta_i$ och sedan $W_{ij} \sim \psi_{Z_j}$.

Vad är $f_{Z,\psi,\theta|W}(\psi, \theta, z|w)$? Täljaren i Bayes formel hyfsat enkel, men nämnaren är

$$f_W(w) = \sum_z \int \ldots \int f_{W|\psi,\theta,Z}(w|\psi,\theta,z) f(\psi,\theta,z) d\psi d\theta.$$

Detta är K^{DL} termer som alla är integraler i sin tur!

Gibbs sampling: Både $f_{Z|\psi,\theta,W}$, $f_{\psi|\theta,Z,W}$ och $f_{\theta|\psi,Z,W}$ är lätta (de två sistnämnda är Dirichlet och givet allt annat är alla Z_{ij} oberoende med lätt beräknade fördelningar).

Gör så här:

1. Starta med vilka värden som helst på ψ, θ, Z,
2. uppdatera Z enligt $f_{Z|\psi,\theta,W}$,
3. uppdatera ψ enligt $f_{\psi|\theta,Z,W}$,
4. uppdatera θ enligt $f_{\theta|\psi,Z,W}$,
5. upprepa 2-4.

Detta konvergerar mot den rätta fördelningen. Gör många uppdateringar. (Hur många? Intressant forskningsområde.)