1 Dirichlet priors

Riešime nasledovný problém. Je pred nami neznáme vrece, ktoré obsahuje guličky K farieb. N krát sme vybrali guličku, zapísali sme si farbu a vložili sme ju naspäť. Chceli by sme odhadnúť aké je rozdelenie farieb vo vreci.

Prvých niekoľko sekcii je len definovanie potrebných pojmov a označení (multinomická distribúcia, Dirichletova distribúcia a aposteriórna pravdepodobnostná distribúcia). V mieste prvého zavedenia je na okraji riadku zopakované označenie, takže sa dá ľahko vyhľadať.

1.1 Multinomická distribúcia

Máme vrece, v ktorom je nekonečne veľa farebných guličiek. Je K druhov farieb. Rozdelenie farieb vo vreci je dané vektorom θ

$$\theta = (\theta_1, \theta_2, \ldots, \theta_K)$$

$$\sum_{i=1}^{K} \theta_i = 1, 0 \leq \theta_i \leq 1, K \geq 2$$

N krát opakujeme nasledovný postup: Vyberieme guličku, zapíšeme si jej farbu a vrátime ju naspäť. Na konci sa pozrieme, zrátame počet guličiek každej farby, ktoré sme videli. Dostaneme vektor $n = (n_1, n_2, \ldots, n_K), n_i \geq 0, \sum_{i=1}^{K} n_i = N$. N

Rozdelenie, ktoré pre dané n charakterizuje pravdepodobnosti výsledkov tohto pokusu (vektorov n) sa volá multinomické rozdelenie. Je definované nasledovne:

$$\Pr(n|\theta) = \frac{N!}{\prod_{i=1}^{K} n_i!} \prod_{i=1}^{K} \theta_i^{n_i}$$

Všimnite si, že výraz $N!/\prod_{i=1}^{K} n_i!$ nezávisí od θ. My ho budeme označovať $1/M(n)$. $1/M(n)$

S multinomickou distribúciou by sme už mohli zobrať dáta a nájsť vektor θ^{ML}, ktorý maximalizuje $\Pr(n \mid \theta^{ML})$. Tento vektor je $\theta_i^{ML} = n_i/N$. Ak máme málo dát (N je malé), a pravdepodobnosť vytiahnutia farby k je malá, tak sa môže stať, že guličku farby k nevytiahneme a teda $\theta_k = 0$, čo nie je zrovna dobrý odhad.

Vieme to opraviť pridaním fiktívne merania, napríklad každé n_i zvýšime o 1 (alebo o iné, nie nutne celé číslo). Takýmto fiktívnym meraniam budeme hovoriť pseudomerania. Čo ak ale vieme nieco viac o farbách vo vreci? Napriklad ak vieme, že guličiek s farbou číslo 1 je vo vreci dva krát toko ako guličiek s farbou číslo 2?

1.2 Dirichletova distribúcia

Dirichletova distribúcia je definovaná nasledovne.

$$\mathcal{D}(\theta|n) = Z^{-1}(n) \prod_{i=1}^{K} \theta_i^{n_i - 1}$$

$$\theta_i \geq 0, \sum_{i=1}^{K} \theta_i = 1, n_i > 0, K \geq 2$$

kde $Z(n)$ je normalizačný výraz, $Z(n)$

$$Z(n) = \int_{\sum \theta_i = 1, \theta_i \geq 0} \prod_{i=1}^{K} \theta^{n_i - 1} d\theta = \frac{\prod_{i=1}^{K} \Gamma(n_i)}{\Gamma(\sum_{i=1}^{K} n_i)}$$

kde $\Gamma(x)$ je gamma funkcia, ktorá je definovaná ako $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$. Je to zovšeobecnený faktoriál, pre reálne čísla spĺňa nasledovnú vlastnosť, ktorú budeme neskôr potrebovať: $\Gamma(x+1) = x\Gamma(x)$. $\Gamma(x)$ $\Gamma(x+1) = x\Gamma(x)$

Napriek tomu, že výrazy $Z(n)$ a $M(n)$ vyzerajú podobne tak nie sú rovnaké ani keď n_i sú celé čísla.

1.3 Aposteriórna pravdepodobnostná distribúcia

Máme model M, dáta D a parametre modelu θ. Pozorujeme dáta D, ale nevieme parametre modelu θ. Vieme distribúciu $\Pr(D \mid \theta, M)$. Z Bayesovho vzťahu vieme vypočítať distribúciu parametrov:

$$\Pr(\theta \mid D, M) = \frac{\Pr(D \mid \theta, M) \Pr(\theta \mid M)}{\Pr(D \mid M)}$$

Potrebujeme nejako vhodne zvoliť distribúciu $\Pr(\theta \mid M)$, čo je distribúcia parametrov v našom modeli. To je naša apriórna informácia (prior knowledge).

1.4 Dirichletove pseudomerania

Máme vrece s $K \geq 2$ farebnými guličkami. Nevieme aké je rozdelenie farieb $\theta = (\theta_1, \ldots, \theta_K), 0 \leq \theta_i \leq 1, \sum_{i=1}^{K} \theta_i = 1$ vo vreci z ktorého ťaháme, ale máme nejaký „odhad" $\alpha = (\alpha_1, \ldots, \alpha_K), \alpha_i \geq 0$. Z vreca sme vytiahli N guličiek a dostali sme pozorovanie $n = (n_1, \ldots, n_K), N = \sum_{i=1}^{K} n_i$.

Vypočítame aposteriórnu distribúciu k multinomickej distribúcii s apriórnou informáciou, že rozdelenie farieb vo vreci θ sa správa podľa Dirichletovej distribúcie s parametrom α:

$$\Pr(\theta \mid n, \alpha) = \frac{\Pr(n \mid \theta)\mathcal{D}(\theta \mid \alpha)}{\Pr(n \mid \alpha)}$$

$\Pr(n \mid \theta)$ je multinomická distribúcia (α sme odtiaľ vynechali, lebo nemá vplyv na výsledok), $\mathcal{D}$ je Dirichletova distribúcia a $\Pr(n \mid \alpha)$ nám neskôr vypadne. Dosadíme a dostaneme:

$$\begin{aligned}
\Pr(\theta \mid n, \alpha) &= \frac{1}{M(n)} \prod_{i=1}^{K} \theta_i^{n_i} \cdot \frac{1}{Z(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \cdot \frac{1}{\Pr(n \mid \alpha)} \\
&= \frac{\prod_{i=1}^{K} \theta_i^{n_i + \alpha_i - 1}}{\Pr(n \mid \alpha) M(n) Z(\alpha)} \\
&= \frac{1}{\Pr(n \mid \alpha) M(n) Z(\alpha)} \cdot Z(n + \alpha)\mathcal{D}(\theta \mid n + \alpha) \\
&= \mathcal{D}(\theta \mid n + \alpha)
\end{aligned}$$

Posledné dva kroky sú zaujímavé. Prvý je len substitúcia čitateľa za Dirichletovu distribúciu vynásobenú normalizačným výrazom $Z(n + \alpha)$. Druhý krok vyplýva z toho, že $\Pr(\theta \mid n, \alpha)$ aj $\mathcal{D}(\theta \mid n + \alpha)$ sú dobre definované distribúcie nad θ a preto sa ostatné výrazy (ktoré od θ závidia) navzájom vyrušia.

Ako vidíme, zobrali sme multinomickú distribúciu, spravili sme jej aposteriórnu distribúciu s Dirichletovou distribúciou ako apriórnou informáciou a dostali sme Dirichletovu distribúciu. Preto hovoríme, že Dirichletova distribúcia je konjugovaná k multinomickej distribúcii.

Teraz potrebujeme odvodiť vhodné θ. Mohli by sme vybrať mohli by sme vybrať tie, ktoré dávajú maximálnu pravdepodobnosť:

$$\theta^{ML} = \arg\max_{\theta} \{\mathcal{D}(\theta \mid n + \alpha)\}$$

Tie parametre nevieme vypočítať exaktne, ale môžeme ich vypočítať pomocou numerických metód. Preto namiesto θ^{ML} vypočítame θ^{PME} (posterior mean estimator), teda strednú hodnotu parametrov θ. Pre niektoré distribúcie sú parametre θ^{ML} a θ^{PME} rovnaké, ale pre Dirichletovu distribúciu to neplatí.

θ^{PME}

$$\begin{aligned}\theta_j^{PME} &= \int \theta_j \mathcal{D}(\theta \mid n+\alpha) d\theta \\ &= Z^{-1}(n+\alpha) \int \theta_j \prod_{i=1}^{K} \theta_i^{n_i+\alpha_i-1} d\theta \\ &= \frac{Z(n+\alpha+\delta_j)}{Z(n+\alpha)}\end{aligned}$$

kde δ_j je vektor, ktorý má j-ty prvok jednotku a ostatné prvku má nulové. Posledná rovnosť bola dosiahnutá preto, lebo ak nám súčin absorbuje θ_j, tak dostaneme presne definíciu normalizačného výrazu z Dirichletovej distribúcie. Rozpísaním výrazu $Z(x)$ a úpravami dostaneme:

$$\begin{aligned}\theta_j^{PME} &= \frac{Z(n+\alpha+\delta_j)}{Z(n+\alpha)} \\ &= \frac{\frac{\prod_{i=1}^{K}\Gamma(n_i+\alpha_i+\delta_{ji})}{\Gamma(\sum_{i=1}^{K} n_i+\alpha_i+\delta_{ji})}}{\frac{\prod_{i=1}^{K}\Gamma(n_i+\alpha_i)}{\Gamma(\sum_{i=1}^{K} n_i+\alpha_i)}} = \frac{\prod_{i=1}^{K}\Gamma(n_i+\alpha_i+\delta_{ji})\Gamma(\sum_{i=1}^{K} n_i+\alpha_i)}{\prod_{i=1}^{K}\Gamma(n_i+\alpha_i)\Gamma(\sum_{i=1}^{K} n_i+\alpha_i+\delta_{ji})} \\ &= \frac{\Gamma(n_j+\alpha_j+1)}{\Gamma(n_j+\alpha_j)} \frac{\Gamma(\sum_{i=1}^{K} n_i+\alpha_i)}{\Gamma(1+\sum_{i+1}^{K} n_i+\alpha_i)} \\ &= \frac{(n_j+\alpha_j)\Gamma(n_j+\alpha_j)}{\Gamma(n_j+\alpha_j)} \frac{\Gamma(\sum_{i=1}^{K} n_i+\alpha_i)}{(\sum_{i+1}^{K} n_i+\alpha_i)\Gamma(\sum_{i=1}^{K} n_i+\alpha_i)} \\ &= \frac{n_j+\alpha_j}{\sum_{i=1}^{K} n_i+\alpha_i}\end{aligned}$$

Pri odvodení sme použili fakt, že j-ty prvok δ_j je 1 a ostatné sú nula a vlastnosť Γ funkcie: $\Gamma(x+1) = x\Gamma(x)$. Ako vidíme, ak máme nejaký odhad α o rozdelení farieb vo vreci, tak ho jednoducho pripočítame k meraniam. Ak máme veľa dát, tento odhad bude zanedbateľný.

1.5 Zmiešané dirichletove distribúcie

Niekedy vieme obsah vreca z ktorého ťaháme rozdeliť rozdeliť do m podvriec (očíslované od 1 po m). Ak ťaháme z vreca, tak s pravdepodobnosťou q_t ťaháme z podvreca t. Pre každé podvrece t máme „odhad" $\alpha^t = (\alpha_1^t, \alpha_2^t, \ldots, \alpha_K^t)$ na rozloženie farieb. Chceme zistiť, aké je rozloženie farieb α^t vo vreci a použiť pri tom tieto odhady ako pseudomerania.

Dostali sme zmiešanú distribúciu (mixed distribution):

$$\Pr(\theta \mid \alpha^1, \alpha^2, \ldots, \alpha^m) = \sum_{t=1}^{m} q_t \mathcal{D}(\theta \mid \alpha_t)$$

Dostali sme vektor dát n, ideme vypočítať $\Pr(\theta \mid n)$. Z definície podmienenej pravdepodobnosti dostaneme

$$\Pr(\theta \mid n) = \sum_{t=1}^{m} \Pr(\theta \mid \alpha^t, n) \Pr(\alpha^t \mid n) = \sum_{t=1}^{m} \Pr(\alpha^t \mid n) \mathcal{D}(\theta \mid n+\alpha_t)$$

lebo $\Pr(\theta \mid \alpha^t, n) = \mathcal{D}(\theta \mid n+\alpha_t)$. $\Pr(\alpha^t \mid n)$ si odložíme na neskôr. Potrebujeme si vypočítať strednú hodnotu θ_j^{PME}.

$$\theta_j^{PME} = \int \theta_j \sum_{t=1}^{m} \Pr(a^t \mid n) \mathcal{D}(\theta \mid n+\alpha^t) d\theta = \sum_{t=1}^{m} \Pr(\alpha^t \mid n) \frac{n_j+\alpha_j^t}{\sum_{i=1}^{K} n_i+\alpha_i^t}$$

Mohli by sme už skončiť, ale stále nepoznáme hodnotu $\Pr(\alpha^t|n)$. Musíme si ju odvodiť. Skôr než budeme pokračovať, pripomenieme si nasledovný vzťah:

$$\frac{1}{\Pr(n \mid \alpha)M(n)Z(\alpha)} \cdot Z(n+\alpha)\mathcal{D}(\theta \mid n+\alpha) = \mathcal{D}(\theta \mid n+\alpha)$$

Jednoduchou úpravou z neho dostaneme nasledovnú rovnosť.

$$\Pr(n \mid \alpha) = \frac{Z(n+\alpha)}{Z(\alpha)M(n)}$$

Chceme vypočítať $\Pr(\alpha^t \mid n)$. Z Bayesovho vzorca vieme:

$$\begin{aligned}
\Pr(\alpha^t \mid n) &= \frac{\Pr(\alpha^t)\Pr(n \mid \alpha^t)}{\sum_{i=1}^{m} \Pr(\alpha^i)\Pr(n \mid \alpha^i)} \\
&= \frac{q_t \frac{Z(n+\alpha^t)}{Z(\alpha^t)M(n)}}{\sum_{i=1}^{m} q_i \frac{Z(n+\alpha^i)}{Z(\alpha^i)M(n)}} \\
&= \frac{q_t Z(n+\alpha^t)/Z(\alpha^t)}{\sum_{i=1}^{m} q_i Z(n+\alpha^i)/Z(\alpha^i)}
\end{aligned}$$

Zhrnutie: Ak máme odhady $\alpha^t = (\alpha_1^t, \ldots, \alpha_K^t)$ s pravdepodobnosťou q_t a meranie $n = (n_1, \ldots, n_K)$, tak farbe j dáme v našom modeli pravdepodobnosť

$$\theta_j^{PME} = \sum_{t=1}^{m} \left(\frac{q_t Z(n+\alpha^t)/Z(\alpha^t)}{\sum_{i=1}^{m} q_i Z(n+\alpha^i)/Z(\alpha^i)} \cdot \frac{n_j + \alpha_j^t}{\sum_{i=1}^{K} n_i + \alpha_i^t} \right)$$