Posterior and MAP Estimate for the Categorical

Weather (W) → Cloudy, Rainy, Sunny } $\underline{\theta}$

observe the weather

$D = \{C, R, S, S, S, R, C, S, \dots\}$

infer $\underline{\theta}$ from D by MLE

→ is there a robuster estimate?

→ can we encode prior knowledge

$W \sim Cat(\underline{\theta})$

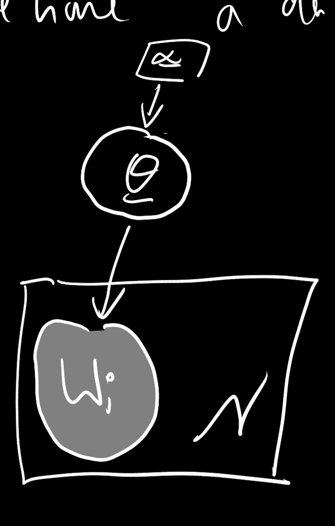
$W \in \{C, R, S\}$

0 1 2

$[1,0,0]^T$ $[0,1,0]^T$ $[0,0,1]^T$

DGM

We have a dataset with N weather obs



$\underline{\theta} \sim Dir(\underline{\theta}; \underline{\alpha})$

$W_i \sim Cat(W_i \mid \underline{\theta})$

joint $p(\underline{\theta}, \underline{W}) = p(\underline{\theta}) \cdot p(\underline{W} \mid \underline{\theta})$

$\overset{i.i.d.}{=} p(\underline{\theta}) \prod_{i=0}^{N-1} p(W_i \mid \underline{\theta})$

$\underbrace{p(\underline{\theta}, \underline{W} = D)}_{p(\underline{\theta}, D)} = \underbrace{p(\underline{\theta})}_{\text{prior}} \cdot \underbrace{\prod_{i=0}^{N-1} p(W_i = w^{[i]} \mid \underline{\theta})}_{p(D \mid \underline{\theta})}$

↳ likelihood

posterior $p(\underline{\theta} \mid D) \overset{\text{Bayes' Rule}}{=} \frac{p(D \mid \underline{\theta})\, p(\underline{\theta})}{p(D)}$

$\sim p(D \mid \underline{\theta}) \cdot p(\underline{\theta})$

$= p(\underline{\theta}) \cdot \prod_{i=0}^{N-1} p(W = w^{[i]} \mid \underline{\theta})$

all W_i are distributed identically

$= Dir(\underline{\theta}; \underline{\alpha}) \cdot \prod_{i=0}^{N-1} Cat(W = w^{[i]} \mid \underline{\theta})$

$= \underbrace{\frac{\Gamma\left(\sum_{d=0}^{D-1} \alpha_d\right)}{\prod_{d=0}^{D-1} \Gamma(\alpha_d)}}_{\text{ignore}} \prod_{d=0}^{D-1} \left(\theta_d^{\alpha_d - 1}\right) \cdot \prod_{i=0}^{N-1} \prod_{d=0}^{D-1} \left(\theta_d^{I(w^{[i]} = d)}\right)$

(we have D states)

$\sim \prod_{d=0}^{D-1} \left(\theta_d^{\alpha_d - 1}\right) \cdot \prod_{d=0}^{D-1} \underbrace{\prod_{i=0}^{N-1} \theta_d^{I(w^{[i]} = d)}}$

e.g. $D = \{1, 0, 0, 2, 3, 0, \dots\}$

$\theta_0^0 \cdot \theta_0^1 \theta_0^1 \theta_0^0 \theta_0^0 \theta_0^1$

$= \prod_{d=0}^{D-1} \left(\theta_d^{\alpha_d - 1} \cdot \theta_d^{\sum_{i=0}^{N-1} I(w^{[i]} = d)}\right)$

$N_d := \sum_{i=0}^{N-1} I(w^{[i]} = d)$

"the number of times we observe the d-th state"

$= \prod_{d=0}^{D-1} \theta_d^{\alpha_d - 1} \cdot \theta_d^{N_d}$

$= \prod_{d=0}^{D-1} \theta_d^{\overbrace{\alpha_d + N_d}^{\alpha_d'} - 1}$

(Can you see why the Dirichlet is the conjugate prior?)

$=: \mathcal{M}(D; \underline{\theta})$

"posterior likelihood"

Full posterior

→ full distribution

$p(\underline{\theta} \mid D) \sim \mathcal{M}(D; \underline{\theta})$

$\alpha_d' := \alpha_d + N_d$

$p(\underline{\theta} \mid D) \sim \prod_{d=0}^{D-1} \theta_d^{\alpha_d' - 1}$

isn't it just a Dirichlet?

$p(\underline{\theta} \mid D) = Dir(\underline{\theta}; \underline{\alpha}')$ — different from α

$$p(\underline{\theta} \mid D) = \frac{\Gamma\left(\sum_{d=0}^{D-1} \alpha_d'\right)}{\prod_{d=0}^{D-1} \Gamma(\alpha_d')} \cdot \prod_{d=0}^{D-1} \theta_d^{\alpha_d' - 1}$$

MAP estimate

Maximum A Posteriori

→ point estimate

$\underline{\theta}^*_{MAP} = \underset{\underline{\theta}}{\arg\max} \left(\mathcal{M}(D; \underline{\theta})\right)$ s.t. $\sum_{d=0}^{D-1} \theta_d = 1$

→ "log posterior likelihood"

$m(D; \underline{\theta}) = \log(\mathcal{M}(D; \underline{\theta}))$

$= \sum_{d=0}^{D-1} (\alpha_d + N_d - 1) \cdot \log(\theta_d)$

Build Lagrangian to include constraint

$\hat{m}(D; \underline{\theta}, \lambda) = m(D; \underline{\theta}) + \lambda \left(1 - \sum_{d=0}^{D-1} \theta_d\right)$

$= \sum_{d=0}^{D-1} \left((\alpha_d + N_d - 1) \cdot \log(\theta_d)\right) + \lambda \cdot \left(1 - \sum_{d=0}^{D-1} \theta_d\right)$

↳ Take derivative and set to 0

$\frac{\partial \hat{m}}{\partial \theta_e} = (\alpha_e + N_e - 1) \cdot \frac{1}{\theta_e} - \lambda \overset{!}{=} 0$

e-th component of $\underline{\theta}$

$\Leftrightarrow \lambda = \frac{\alpha_e + N_e - 1}{\theta_e}$ (*)

(also holds for $(e-1), (e+1)$: $\lambda = \frac{\alpha_d + N_d - 1}{\theta_d}$)

$\frac{\partial \hat{m}}{\partial \lambda} = 1 - \sum_{d=0}^{D-1} \theta_d \overset{!}{=} 0$

$\Leftrightarrow \sum_{d=0}^{D-1} \theta_d = 1$

How to eliminate λ

$\lambda = \lambda = \lambda \cdot 1 = \lambda \cdot \sum_{d=0}^{D-1} \theta_d = \sum_{d=0}^{D-1} \lambda \theta_d$

$= \sum_{d=0}^{D-1} \frac{\alpha_d + N_d - 1}{\theta_d} \cdot \theta_d$

$= \sum_{d=0}^{D-1} (\alpha_d + N_d - 1)$

$= \sum_{d=0}^{D-1} (\alpha_d) + \underbrace{\sum_{d=0}^{D-1} (N_d)}_{N} - \underbrace{\sum_{d=0}^{D-1} (1)}_{D}$

$= N - D + \sum_{d=0}^{D-1} \alpha_d = \lambda$

plug into (*)

$N - D + \sum_{d=0}^{D-1} \alpha_d = \frac{\alpha_e + N_e - 1}{\theta_e}$

MAP

$$\theta_e = \frac{N_e + \alpha_e - 1}{N - D + \sum_{d=0}^{D-1} \alpha_d}$$

"pseudo-counts"

with $N_e = \sum_{i=0}^{N-1} I(w^{[i]} = e)$

(for One-Hot-Categorical $N_e = \sum_{i=0}^{N-1} w_e^{[i]}$)