MLE for the Categorical

Weather (w) → Cloudy, Rainy, Sunny

$$p(w)=\prod_{d=0}^{D-1}\theta_d^{I(w=d)}$$

?

$$w\in\{\overset{0}{C}, \overset{1}{R}, \overset{2}{S}\}$$

given observations

$$D=\{0,0,1,1,2,0,2,2,2,1,0,\dots\}$$

what is $\underline{\theta}$?

Likelihood

$$\mathcal{L}(D;\underline{\theta})\overset{i.i.d.}{=}\prod_{i=0}^{N-1} p(W=w^{[i]}) = \prod_{i=0}^{N-1}\prod_{d=0}^{D-1}\theta_d^{I(w^{[i]}=d)}$$

Log-Likelihood

$$\ell(D;\underline{\theta}) = \log(\mathcal{L}(D;\underline{\theta})) = \sum_{i=0}^{N-1}\sum_{d=0}^{D-1} I(w^{[i]}=d)\log(\theta_d)$$

Maximum Likelihood Estimate (MLE)

$$\underline{\theta}^* = \underset{\substack{\underline{\theta}\in[0,1]^D \\ \sum_{d=0}^{D-1}\theta_d = 1}}{\arg\max}\,(\ell(D;\underline{\theta}))$$

constrained optimization

→ incorporate constraints into target function

→ Lagrange Multiplier

constraint function

$$\underbrace{\sum_{d=0}^{D-1}\theta_d - 1}_{g(\underline{\theta})} = 0$$

$$g(\underline{\theta}) = 1-\sum_{d=0}^{D-1}\theta_d$$

Augmented target

$$\hat{\ell}(D;\underline{\theta},\lambda) = \ell(D;\underline{\theta}) + \lambda g(\underline{\theta}) = \sum_{i=0}^{N-1}\sum_{d=0}^{D-1}\left(I(w^{[i]}=d)\log(\theta_d)\right) + \lambda\left(1-\sum_{d=0}^{D-1}\theta_d\right)$$

→ Take derivative and set to 0

$$\frac{\partial\hat{\ell}}{\partial\theta_e} = \sum_{i=0}^{N-1}\left(I(w^{[i]}=e)\frac{1}{\theta_e}\right) - \lambda \overset{!}{=} 0$$

θ_e ... one component of $\underline{\theta}$

$$\Leftrightarrow \lambda = \frac{1}{\theta_e}\underbrace{\sum_{i=0}^{N-1} I(w^{[i]}=e)}_{N_e}$$

N_e ... the # of times the weather is 'e'

$$\boxed{\lambda = \frac{N_e}{\theta_e}}$$

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

How to get rid of the λ?

$$\left(\Leftrightarrow 1 = \sum_{d=0}^{D-1}\theta_d\right)$$

Some trick:

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

$$\lambda = \frac{N_e}{\theta_e} = \text{const.} = \frac{N_{e-1}}{\theta_{e-1}} = \frac{N_{e+1}}{\theta_{e+1}} = \frac{N_d}{\theta_d}$$

$$\lambda = \sum_{d=0}^{D-1}\frac{N_d}{\cancel{\theta_d}}\cdot\cancel{\theta_d} = \sum_{d=0}^{D-1}N_d = \boxed{N = \lambda}$$

$$N = \frac{N_e}{\theta_e} \Leftrightarrow \boxed{\theta_e = \frac{N_e}{N}}$$