DS 4400
Instructor: Ehsan Elhamifar

Last lecture, we discussed the problem of Max Likelihood estimation, where given a dataset D, and a prob model whose parameters are specified by θ, we want to find the best parameter which makes observations D as likely /probable as possible

$$L(\theta) \triangleq p(D|\theta) \qquad \max_{\theta} \overset{\text{likelihood}}{L(\theta)}$$

$$\text{log-likelihood} \leftarrow \ell(\theta) = \log p(\overset{\{x_1, \dots, x_N\}}{D}|\theta) \overset{x_i\text{'s}}{\underset{\text{iid}}{=}} \sum_{i=1}^{N} \log p(x_i|\theta)$$

Distributions over parameters

In the MLE, we were looking for best θ that maximizes probability of observations, D
what if we have some distribution over θ (coming from prior information, e.g, the coin is fair)?

We want to find the most likely θ, given observations, D i.e, want to maximize

$$\underset{\theta}{\text{argmax}}\; p(\theta|D) \qquad \text{Maximum A Posteriori (MAP) estimation}$$

This is really a Bayesian framework, where data and parameters have distributions

To do MAP, $\quad p(\theta|D) = \dfrac{p(D|\theta)\,p(\theta)}{p(D)} \propto p(D|\theta)\,p(\theta)$

$$\underset{\theta}{\text{argmax}}\; p(\theta|D) = \underset{\theta}{\text{argmax}}\; p(D|\theta)\,p(\theta)$$

$$p(\theta|D) \propto p(D|\theta)\, p(\theta)$$

posterior likelihood prior

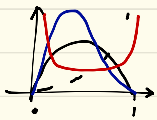
IE, we have some belief about θ ($p(\theta)$), we observe data D, how our belief on θ changes, given D ($p(\theta|D)$)

Let's assume θ has a Beta distribution (good dist for variable $\in [0,1]$)

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \quad \theta \in [0,1]$$

$$B(\alpha,\beta) \rightarrow \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\begin{cases} \Gamma(\alpha) \triangleq \int_0^\infty t^{\alpha-1} e^{-t}\, dt \\ \Gamma(n) = (n-1)! \end{cases}$$



$E[\theta] = \frac{\alpha}{\alpha+\beta}$

$\alpha, \beta > 0$

Mode $= \frac{\alpha-1}{\alpha+\beta-2}$ $\quad \alpha,\beta > 1$

We know $p(D|\theta) = \theta^{\sum_n x_n} (1-\theta)^{N-\sum_n x_n}$

$$\rightarrow p(\theta|D) \propto \theta^{\alpha+\sum_n x_n - 1} (1-\theta)^{\beta+N-\sum_n x_n - 1} \propto f_{Beta}\left(\alpha+\sum_n x_n,\ \beta+N-\sum_n x_n\right)$$

$\rightarrow$ Also a Beta distribution with mode

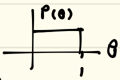
$$\underset{\theta}{\operatorname{argmax}}\ p(\theta|D) = \hat{\theta}_{MAP} = \frac{\sum_n x_n + \alpha - 1}{N+\alpha+\beta-2}$$

$$= \left(\frac{N}{N+\alpha+\beta-2}\right)\frac{\sum x_n}{N} + \left(\frac{\alpha+\beta-2}{N+\alpha+\beta-2}\right)\frac{\alpha-1}{\alpha+\beta-2}$$

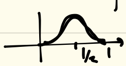
$$= \text{Convex comb}\left(\frac{\sum_n x_n}{N},\ \frac{\alpha-1}{\alpha+\beta-2}\right)$$

$N \rightarrow 0 \Rightarrow$ only prior

$N \rightarrow \infty \Rightarrow$ MLE (prior forgotten as N increases)

If $p(\theta) =$ uniform 

$\rightarrow \alpha,\beta = 1 \rightarrow$ MAP $=$ MLE

If $p(\theta)$ 

fair coin $\rightarrow \alpha=\beta=2 \rightarrow \frac{1}{N+2} + \frac{\sum_n x_n}{N+2}$