Estimating Probabilities from data

- MLE Recap:

- Seek an estimate of θ that maximizes the probability of the observed data

$$\boxed{f(x;\theta)}$$

A function of x (m.f. / d.f)

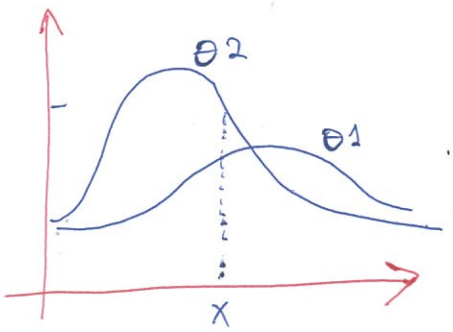
R.V.

fixed

- To do so, we define the likelihood function

$L(\theta) = L(\theta|x) = f(x;\theta)$

- MLE principle find $\theta_{MLE} = \arg\max_{\theta} \log(f(x;\theta))$



Remark: Given that we observe the data, which parameter(s) would make it most likely that we observe what we observe

- Extension to the data set

$L(\theta) = L(\theta|D) = f(D;\theta)$

$= f(x_1;\theta) \cdot f(x_2;\theta) \dots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$

$\Rightarrow \theta_{MLE} = \arg\max_{\theta} \log\left(\prod_{i=1}^{n} f(x_i;\theta)\right)$

$= \arg\max_{\theta} \sum_{i=1}^{n} (\log f(x_i;\theta))$

Summary:

- MLE gives the explanation of the data we observed.
- If n is large and $f(x|\theta)$ is chosen correctly, then MLE will find the true parameter.
- MLE can overfit the data if n is small.
- If $f(x|\theta)$ is chosen incorrectly, then MLE can be terribly wrong.

MAP (Another way of estimating probabilities from data)

MLE: Estimate θ_{MLE} that makes $P(\underbrace{D}_{\text{R.V.}} | \underbrace{\theta}_{\text{parameter (Fixed)}})$ maximized.

- MLE is frequentist statistics, meaning that θ is just some constant.
- In Bayesian statistics, θ can be r.v.

  $\Rightarrow$ There is the distribution $\underbrace{P(\theta)}$ ← Encode your belief of θ

Bayes rule:

likelihood, prior over θ

$$\underbrace{P(\theta|D)}_{\text{Posterior}} = \frac{P(D|\theta)\, P(\theta)}{\underbrace{P(D)}_{\text{normalized term}}}$$

$\Rightarrow P(\theta|D) \propto P(D|\theta)\, P(\theta)$

$P(D|\theta)$ ← which parameter makes our data the most likely

$P(\theta|D)$ ← Given that we have data, what is the most likely parameter

MAP Principle: find $\theta_{MAP} = \arg\max_{\theta} \log P(\theta | D)$.

$$= \arg\max_{\theta} \log P(D|\theta) P(\theta)$$

$$= \arg\max_{\theta} (\log P(D|\theta)) + (\log P(\theta))$$

MAP (Maximum a Posteriori Probability Estimation)

Simple Scenario: coin toss w/ prior knowledge

- Ex, suppose you toss a coin and observe $D = \{H, H, H, H, H\}$

MLE $\Rightarrow$ $\theta_{MLE} = \frac{n_H}{n} = 1$

- If you don't trust your estimate, then you can fix it by

$$\theta = \frac{n_H + m}{n + 2m}$$

- Here, m is the number of imaginary throws that would result in $\frac{m}{2m} = 0.5$ (your hunch is close to 0.5)

- For large n, $\theta \to \theta_{MLE}$.
- For small n, this incorporates your "prior belief" about what θ should be

- Let's formalize this using MAP principle

Natural choice for the prior $P(\theta)$ is the Beta distribution.

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$ constant

$$\cancel{\log} P(\theta) \propto \cancel{\log}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$P(\theta|D) \propto P(D|\theta)\,P(\theta) \propto \binom{n}{n_H} \theta^{n_H} \cdot (1-\theta)^{n-n_H} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$$

$$= \binom{n}{n_H} \theta^{n_H+\alpha-1} (1-\theta)^{n-n_H+\beta-1}$$

$$\Rightarrow \theta_{MAP} = \frac{n_H + (\alpha-1)}{n + (\alpha+\beta-2)}$$