4 Estimating Probabilities of Data: MLE

Assumption: There is a distribution for data $p(x,y)$

If we have the distribution, then we can predict label using $p(y|x)$

Two ways to obtain $p(x,y)$

$p(x,y) = p(y|x)\,p(x)$ Discriminative learning (Logistic Regression, SVM)

$\quad\;\; = p(x|y)\,p(y)$ Generative learning (VAE, GAN)

Maximum Likelihood Estimation

Binomial Distr: H, T, T, H, H, H, T, T, T, T

What is $\mathbb{P}\{H\}$? $\mathbb{P}\{H\} \approx \frac{n_H}{n_H + n_T}$

MLE: $\mathbb{P}\{D;\theta\}$ Let $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathbb{P}\{D;\theta\}$

D is data. Want to find θ that maximizes the prob of observing the data

$$\mathbb{P}\{D;\theta\} = \binom{n_H + n_T}{n_H} \theta^{n_H} (1-\theta)^{n_T}$$

Find log-likelihood log monotonic increasing

$$\log(\mathbb{P}\{D;\theta\}) = \log\binom{n_H + n_T}{n_H} + n_H \log(\theta) + n_T \log(1-\theta)$$

Maximize the log by taking derivative

$$\frac{\partial}{\partial\theta} \log(\mathbb{P}\{D;\theta\}) = \frac{n_H}{\theta} + \frac{n_T}{1-\theta} = 0$$

$$\theta = \frac{n_H}{n_H + n_T}$$

Shortfalls: With small sample size, estimate unstable

With 1 toss, predict H/T all the time

Fix: Smoothing. Hallucinate samples of α H in β toss

$$\frac{n_H + \alpha}{n_H + n_T + \beta}$$

Bayesian vs. Frequentist

Frequentist $\mathbb{P}\{D;\theta\}$ θ is an unknown constant

$\mathbb{P}\{\theta\}$ ill defined

MLE maximizes $\mathbb{P}\{D;\theta\}$

Bayesian $\mathbb{P}\{D|\theta\}$ θ is a distribution

$\mathbb{P}\{\theta\}$ encodes your belief of what θ should be

Baye's Rule

$\mathbb{P}\{D|\theta\}$ Likelihood

$\mathbb{P}\{\theta\}$ Prior

$\mathbb{P}\{\theta|D\}$ Posterior

$$\mathbb{P}\{\theta|D\} = \frac{\mathbb{P}\{D|\theta\}\,\mathbb{P}\{\theta\}}{\mathbb{P}\{D\}}$$

← Likelihood × Prior

← Normalization

Back to coin toss example

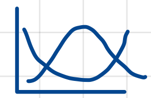
Let prior be beta distribution

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

$B(\alpha,\beta)$ is normalizing constant

Depends on, α, β,

Beta distr looks like



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

$$= \binom{n_H+n_T}{n_H}\theta^{n_H}(1-\theta)^{n_T}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

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

If you were to do MLE on $P(\theta|D)$

$$\hat{\theta}_{MLE} = \frac{n_H+\alpha-1}{n_H+n_T+\alpha+\beta-1}$$

Same as smoothing