ML

Supervised learning

$D = \{(\bar{x}_n, y_n)\}_{n=1}^{N}$

input output

$p(y|\bar{x})$

$\bar{x}$ — $f(\bar{x}) \approx y$

$f(\bar{x}) \in [0,1]$

$\approx p(\text{"cat"}|\bar{x})$

Unsupervised learning

$D = \{\bar{x}_n\}_{n=1}^{N}$

clustering

Reinforcement learning

dimensionality reduction

feature extraction

$\bar{x}$ — represent. learning — $\bar{z} \in \mathbb{R}^d$

1 Mpix $3 \cdot 10^6$

$\bar{x} \in \mathbb{R}^{500}$

Clustering

$\bar{x}$ — $\bar{z} \in \mathbb{R}^d$

$d = 0$

project

cats

dogs

Autoencoder

$p(\bar{z})$

$\bar{z} \in \mathbb{R}^d$

$\bar{x} \sim p(\bar{x})$

$\bar{x}$ → Encoder → $\bar{z}$ → Decoder → $\hat{\bar{x}}$

"random noise"

$L = \|\hat{\bar{x}} - \bar{x}\|^2$

$p(x)$

1 - - 6

$p(x)$

$\int_0^1 p(x)\,dx = 1$

0 1

Joint prob

$p(x,y)$

$p(x=a \;\&\; y=b)$

$p(x) = \sum_y p(x,y)$

cond. prob $p(x|y)$ $p(x=a|y=b)$

$p(x,y) \overset{def}{=} p(x|y)p(y)$

$p(x|y) = \frac{p(x,y)}{p(y)}$ def $= p(x)p(y|x)$

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$ Bayes rule formula

independence $p(x,y) = p(x)p(y)$

$p(x|y) = p(x)$

$p(x,y|z) = p(x|z)p(y|z)$ Conditional independence

① θ max

likelihood prior distribution

posterior distribution ② θ max

$$p(\bar{\theta}|D) = \frac{p(D|\bar{\theta})\,p(\bar{\theta})}{p(D)}$$

data model parameters data evidence prob.

max $\bar{\theta}$ $\bar{\theta}$ max

$p(\bar{\theta}|D) \propto p(D|\bar{\theta})p(\bar{\theta})$ proportional

$\log p(\theta|D) = \log p(D|\theta) + \log p(\theta)$ const

$p(D) = \int p(\bar{\theta})p(D|\bar{\theta})\,d\bar{\theta}$

Example: coin tossing

$\theta = p(\text{heads}|\theta)$

$1-\theta = p(\text{tails}|\theta)$

hthhhtt

$p(D|\theta) =$

$= p(d_1|\theta)\,p(d_2|\theta) \cdots p(d_N|\theta)$

$D = htthhhtt$

$p(D|\theta) = p(h|\theta)\,p(t|\theta)\cdots p(t|\theta) = \theta^4 (1-\theta)^3$ | likelihood

1) Maximum likelihood

$\theta_{ML} = \arg\max_\theta p(D|\theta)$ $\quad D = \begin{matrix} n \text{ heads} \\ m \text{ tails} \end{matrix}$

$p(D|\theta) = \theta^n (1-\theta)^m \xrightarrow{\theta} \max$?

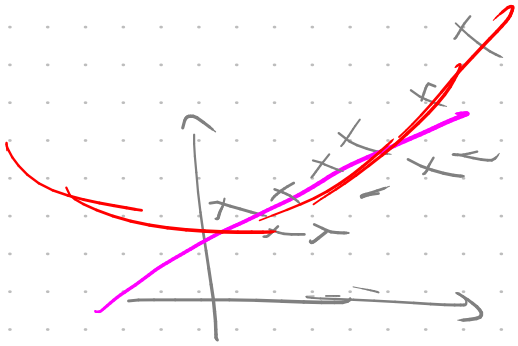
$\frac{\partial p(D|\theta)}{\partial \theta} = n\theta^{n-1}(1-\theta)^m - m\theta^n (1-\theta)^{m-1} = 0$

$\theta^{n-1}(1-\theta)^{m-1}\left(n\cdot(1-\theta) - m\cdot\theta\right) = 0$

$n - (n+m)\cdot\theta$

$\theta \neq 0$, $\theta \neq 1$, $\theta = \frac{n}{n+m}$

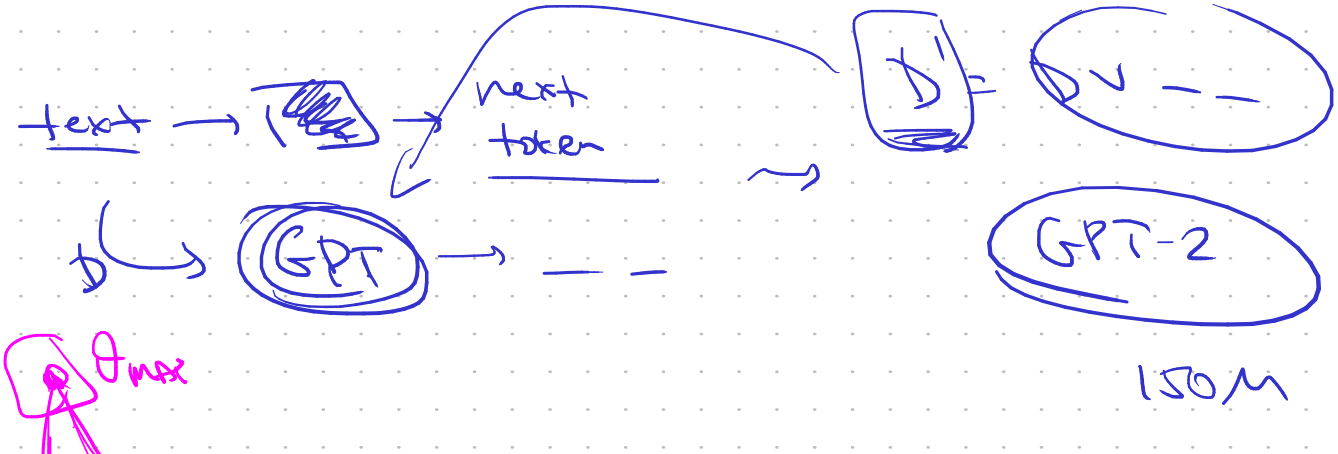
$\theta_{ML} = \frac{n}{n+m} \xrightarrow[n,m\to\infty]{} \text{true } \theta$



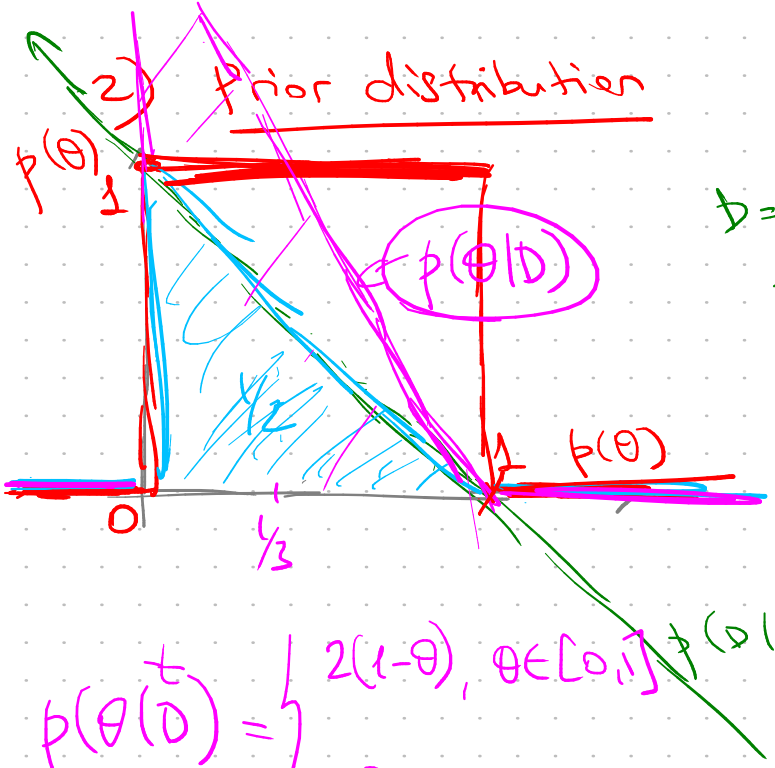
$D = t$ $\quad p(D|\theta) = 1-\theta$

$\theta_{ML} = ?$ $\quad \theta_{ML} \neq -\infty$

$\theta_{ML} = 0$



2) Prior distribution $p(\theta)$



$D = t$

$p(D|\theta) = 1-\theta$

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

$$p(\theta|t) = \begin{cases} 2(1-\theta), & \theta \in [0,1] \\ 0, & \text{--''--} \end{cases} \quad p(D|\theta)$$

$$\theta_{MAP} = \arg\max_\theta p(\theta|D)$$

Max a posteriori

$$\theta_{MAP} = \frac{h}{n+m}$$

$$p(\theta|D) \propto p(D|\theta) p(\theta) = \begin{cases} \theta^n (1-\theta)^m, & [0,1] \\ 0, & \text{--''--} \end{cases}$$

3) Predictive distribution

$$p(\text{heads}|D) = \int p(\text{heads}|\theta)\, p(\theta|D)\, d\theta = E_{p(\theta|D)}\left[p(\text{heads}|\theta)\right]$$

$$= \int_0^1 \theta \frac{\theta^n (1-\theta)^m}{p(D)} d\theta = \frac{\int_0^1 \theta^{n+1} (1-\theta)^m d\theta}{\int_0^1 \theta^n (1-\theta)^m d\theta} = \frac{n+1}{n+m+2}$$

Laplace's Rule of Succession

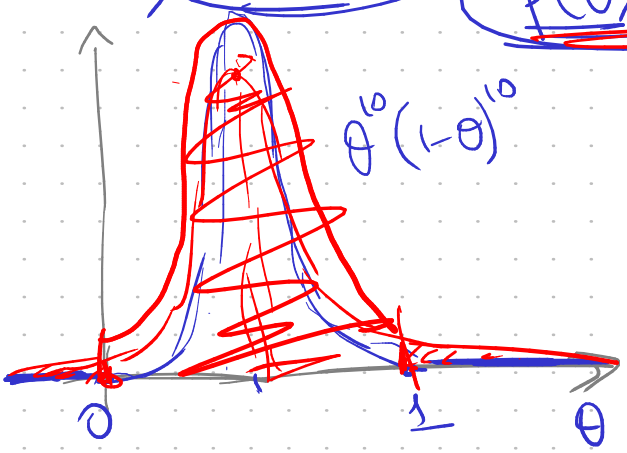
$$p(\bar\theta|D) = \frac{p(D|\bar\theta)\, p(\bar\theta)}{p(D)}$$

likelihood: $p(\bar x|\bar\theta)$ posterior: $p(\bar\theta|D)$

Pred. distr. $p(\bar x|D) = \int p(\bar x, \bar\theta|D) d\bar\theta = \int p(\bar x|\bar\theta)\, p(\bar\theta|D) d\bar\theta$

4) Prior ?

$p(\theta) = ?$



$$p(\theta) = \begin{cases} c \cdot e^{-\frac{1}{2\sigma^2}(\theta-\frac{1}{2})^2} & [0,1] \\ 0, \text{ — " — } \end{cases}$$

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

$$= c \cdot e^{-\frac{1}{2\sigma^2}(\theta-\frac{1}{2})^2} \cdot \theta^n (1-\theta)^m$$

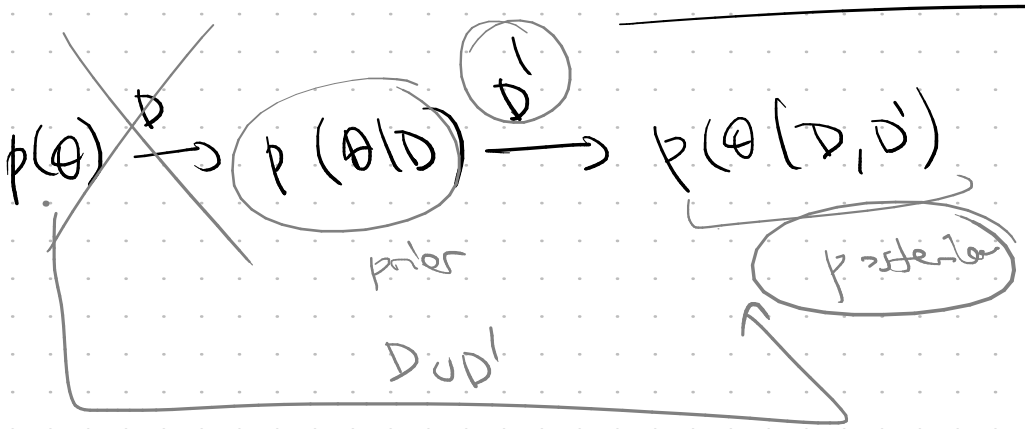
max ←

likelihood | prior | posterior

$$\theta^n \cdot (1-\theta)^m \times c \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{(\alpha+n)-1}(1-\theta)^{(\beta+m)-1}$$

conjugate prior

$$D=(n,m) \times Beta(\theta|\alpha,\beta) \propto Beta(\theta|\alpha+n, \beta+m)$$



$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Laplace's rule

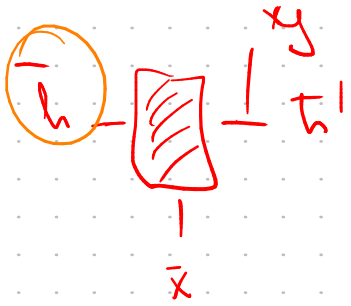
$$p(heads|n,m) = \frac{n+1}{n+m+2}$$

$$p(D|\theta) \xrightarrow{\theta} max \qquad \theta_{ML} = \frac{n}{n+m}$$

$$p(\theta|D) \xrightarrow{\theta} max \qquad \theta_{MAP} = \frac{n}{n+m}$$

uniform $p(\theta)$

$$p(x|D) = \int p(x|\bar{\theta}) p(\bar{\theta}|D) d\bar{\theta} = E_{p(\bar{\theta}|D)}[p(x|\bar{\theta})]$$



$$D = \{(\bar{x}_n, y_n)\}$$