DATASET: $\{x_i\}_{i=1}^{N}$, $x_i \in \{0,1\}$

Example: $\{0,1,0,1,0\} = D$

DATA-DRIVEN APPROACH

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

$$P(x=1) = \frac{2}{5}$$

PRIOR-INDUCED APPROACH

MAXIMUM A POSTERIORI (MAP)

1) Define a set of plausible values for the underlying model parameters (in this example, types of coin)

{fair, 2-headed, coin flips Heads w/ 30%}

prior belief

2) Test each hypothesis.

$$P(\text{fair} \mid \{0,1,0,1,0\}) = \frac{P(\{0,1,0,1,0\} \mid \text{fair}) \cdot P(\text{fair})}{P(\{0,1,0,1,0\})}$$

Bayes' theorem

$$P(\text{fair} \mid D) = \frac{P(D \mid \text{fair}) \cdot P(\text{fair})}{P(D)}$$

$$= \frac{P(0 \mid \text{fair}) \cdot P(1 \mid \text{fair}) \cdot P(0 \mid \text{fair}) \cdot P(1 \mid \text{fair}) \cdot P(0 \mid \text{fair}) \cdot P(\text{fair})}{P(D)}$$

↑ Conditional independence

$$= \frac{(1/2)^3 (1/2)^2 (1/3)}{P(D)}$$

$$P(D) = P(D \mid \text{fair}) \cdot P(\text{fair}) + P(D \mid \text{2-headed}) \cdot P(\text{2-headed}) + P(D \mid \text{flips Heads 30\%}) \cdot P(\text{flips Heads 30\%})$$

$$= (1/2)^3 (1/2)^2 \cdot (1/3) + (0)^3 (1)^2 \cdot (1/3) + (7/10)^3 (3/10)^2 \cdot (1/3)$$

$P(\text{fair} \mid D) = A$

$P(\text{2-headed} \mid D) = B$

$P(\text{coin flip, H w/ 30\%} \mid D) = C$

If A is the largest, then pick prior choice "fair".

2) $P(x=1 \mid \text{fair}) = \frac{1}{2}$

Parameter estimation of MLE and MAP w/ converge to same solution if:

1) Sample size is large enough.

2) prior in MAP is uniformly-distributed.

MLE	MAP
$w = \arg\max_w P(x \mid w)$	$w = \arg\max_w P(x \mid w)\, P(w)$ $= \arg\max_w P(w \mid x)\, P(x)$ $\propto \arg\max_w P(w \mid x)$

$w_{MLE} = w_{MAP}$ if $P(w)$ is a constant.

example Uniform $(-\infty, +\infty)$.

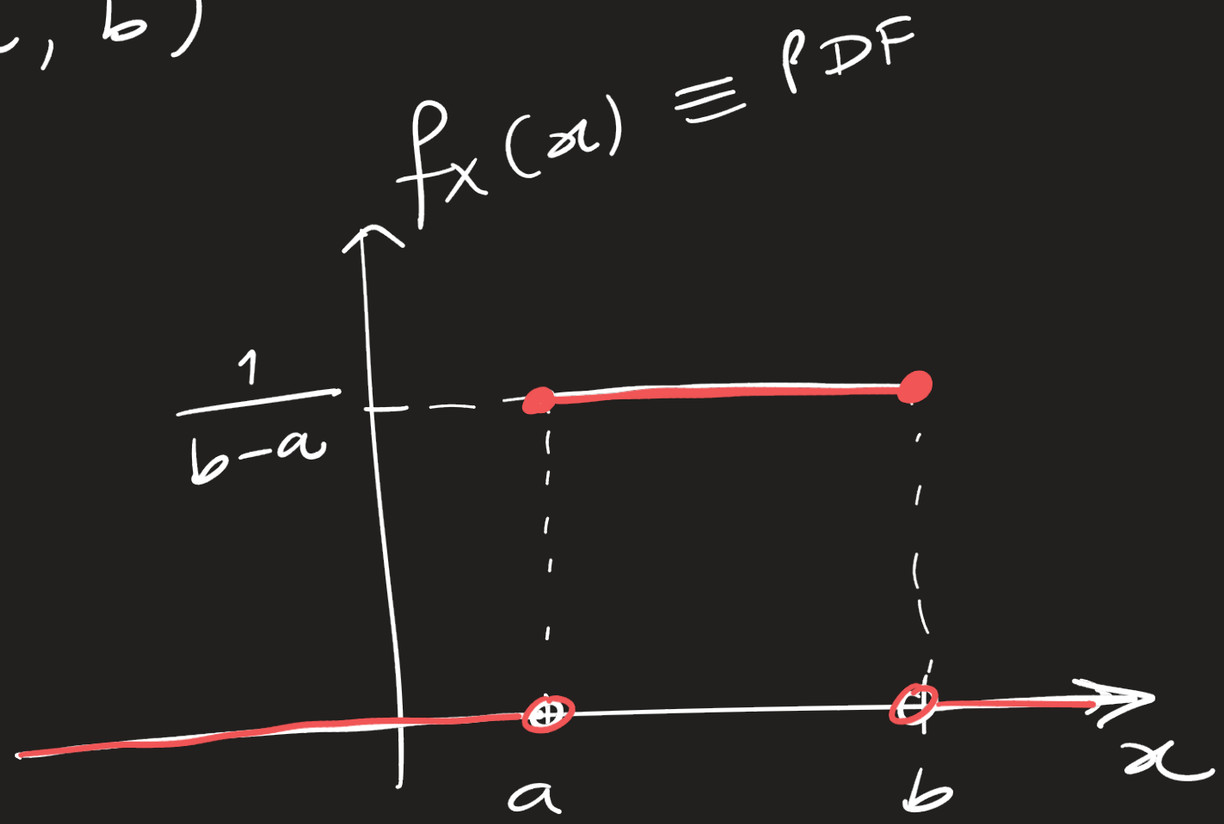
- if $P(w)$ is selected incorrectly, MLE will always outperform MAP.
- But if $P(w)$ is selected correctly then MAP outperforms MLE.

Under small-ish sample sizes.

→ it needs a lot more data to catch up with MLE.

$$X \sim \text{Uniform}(a, b)$$

$$f_X(x) = \frac{1}{b-a}$$



$\mu \equiv$ learnable parameter (represents prob. of flipping Heads)

$x_i \equiv$ sample $= \{0, 1\}$, $\{x_i\}_{i=1}^{N}$ i.i.d.

$$P(x=0 \mid \mu) = 1 - \mu$$

$$P(x=1 \mid \mu) = \mu$$

In general, we can write the data likelihood for sample x as:

$$P(x \mid \mu) = \mu^{x} \cdot (1-\mu)^{1-x}$$

Bernoulli Distribution

$$X \sim \text{Bernoulli}(\mu)$$

Maximum Likelihood Estimation (MLE)

① Set up the observed data likelihood

for dataset $\{x_i\}_{i=1}^{N}$ i.i.d.

$$\mathcal{L}^{\circ} = P(\{x_1, x_2, \dots, x_N\} \mid \mu)$$

$$= \prod_{i=1}^{N} P(x_i \mid \mu)$$

↙ i.i.d. samples

$$= \prod_{i=1}^{N} \mu^{x_i} \cdot (1-\mu)^{1-x_i}$$

$$= \mu^{\sum_{i=1}^{N} x_i} \cdot (1-\mu)^{\sum_{i=1}^{N} (1-x_i)}$$

$$= \mu^{\sum_{i=1}^{N} x_i} \cdot (1-\mu)^{N - \sum_{i=1}^{N} x_i}$$

Illustration:

$\mu \cdot (1-\mu) \cdot \mu$

$= \mu^2 \cdot (1-\mu)$

② Log-likelihood:

$$\mathcal{L} = \ln \mathcal{L}^{\circ}$$

$$= \left(\sum_{i=1}^{N} x_i\right) \cdot \ln(\mu) + \left(N - \sum_{i=1}^{N} x_i\right) \cdot \ln(1-\mu)$$

③ DERIVE w.r.t. μ:

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0 \Leftrightarrow \left(\sum_{i=1}^{N} x_i\right) \cdot \frac{1}{\mu} + \left(N - \sum_{i=1}^{N} x_i\right) \cdot \left(\frac{-1}{1-\mu}\right) = 0$$

$$\Leftrightarrow (1-\mu) \cdot \left(\sum_{i=1}^{N} x_i\right) - \mu \cdot \left(N - \sum_{i=1}^{N} x_i\right) = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} x_i - \mu \cdot \sum_{i=1}^{N} x_i - \mu \cdot N + \mu \cdot \sum_{i=1}^{N} x_i = 0$$

$$\Leftrightarrow \boxed{\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}}$$

Maximum A Posteriori (MAP)

It starts by encoding prior beliefs for μ in the form of a probabilistic model.

$$\mu \sim Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \mu^{\alpha-1} \cdot (1-\mu)^{\beta-1}$$

// constant for some α and β.

Continuous R.V.

$\alpha, \beta > 0$

$\Gamma(x) = (x-1)!$

① OBSERVED DATA LIKELIHOOD × PRIOR:

$$\mathcal{L}^{\circ} = P(\{x_1, \ldots, x_N\} | \mu) \cdot P(\mu)$$

$$= \left[\prod_{i=1}^{N} P(x_i | \mu) \right] \cdot \left[P(\mu) \right]$$

↙ independent

$$= \left[\mu^{\sum_{i=1}^{N} x_i} \cdot (1-\mu)^{N - \sum_{i=1}^{N} x_i} \right] \cdot \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \mu^{\alpha-1} \cdot (1-\mu)^{\beta-1} \right]$$

↙ identically distributed

constant

$$\propto \mu^{\sum_{i=1}^{N} x_i} \cdot (1-\mu)^{N - \sum_{i=1}^{N} x_i} \cdot \mu^{\alpha-1} \cdot (1-\mu)^{\beta-1}$$

$$= \mu^{\sum_{i=1}^{N} x_i + \alpha - 1} \cdot (1-\mu)^{N - \sum_{i=1}^{N} x_i + \beta - 1}$$

② DATA Log-likelihood:

$$\mathcal{L} = \ln \mathcal{L}^{\circ}$$

$$= \left(\sum_{i=1}^{N} x_i + \alpha - 1\right) \cdot \ln(\mu) + \left(N - \sum_{i=1}^{N} x_i + \beta - 1\right) \cdot \ln(1-\mu)$$

③ Derive w.r.t. μ:

$$\frac{d\mathcal{L}}{d\mu} = 0 \iff \boxed{\mu_{MAP} = \frac{\sum_{i=1}^{N} x_i + \alpha - 1}{N + \alpha + \beta - 2}}$$

In cases where the resulting posterior has the same probabilistic shape as the prior

prob. (up to a constant factor).

$\Longrightarrow$ they are said to have a **conjugate prior** relationship.

DATA Likelihood $\times$ PRIOR $\propto$ POSTERIOR

MAP

$$\mathcal{L}^{\circ} = \left[\mu^{\sum_{i=1}^{N} x_i + \alpha - 1} \cdot (1-\mu)^{N - \sum_{i=1}^{N} x_i + \beta - 1} \right] \sim \text{Beta}\left(\alpha + \sum_{i=1}^{N} x_i,\ \beta + N - \sum_{i=1}^{N} x_i \right)$$

and prior

$$P(\mu) = \underbrace{c}_{\text{constant}} \cdot \left[\mu^{\alpha - 1} \cdot (1-\mu)^{\beta - 1} \right] \sim \text{Beta}(\alpha, \beta)$$

Bernoulli x Beta $\propto$ Beta

This forms a conjugate prior relationship

online update of prior

$$\begin{cases} \alpha^{(t+1)} \longleftarrow \alpha^{(t)} + \sum_{i=1}^{N} x_i \\ \beta^{(t+1)} \longleftarrow \beta^{(t)} + N - \sum_{i=1}^{N} x_i \end{cases} \quad \text{(RECURSIVE update)}$$

Pseudo-code for Online Prior Update:

$t = 0$ (iteration)

1) START w/ initial guess for prior parameters.

$$\alpha^{(t)}, \quad \beta^{(t)}$$

2) Calculate the solution for parameter w/ current prior values:

$$\mu_{MAP} = \frac{\sum_{i=1}^{N} x_i + \alpha^{(t)} - 1}{N + \alpha^{(t)} + \beta^{(t)} - 2}$$

3) Update the prior parameters

$$\alpha^{(t+1)} \leftarrow \alpha^{(t)} + \sum_{i=1}^{N} x_i$$

$$\beta^{(t+1)} \leftarrow \beta^{(t)} + N - \sum_{i=1}^{N} x_i$$

4) $t \leftarrow t + 1$, go back to step (2).

Before Next class ...

1) Review	Univariate	Gaussian distributions.
	Bivariate	
	Multivariate (in general)	