2/9/15

STATISTICAL LEARNING/ESTIMATION
(MLE, MAP, Bayesian)

① Maximum Likelihood Estimation

→ Real World Phenomenon / Sample Space

$\Omega = \{\text{Nadal Loses } (L), \text{Nadal Wins } (W)\}$

→ Random Variable: $Y = y \in \{0, 1\}$

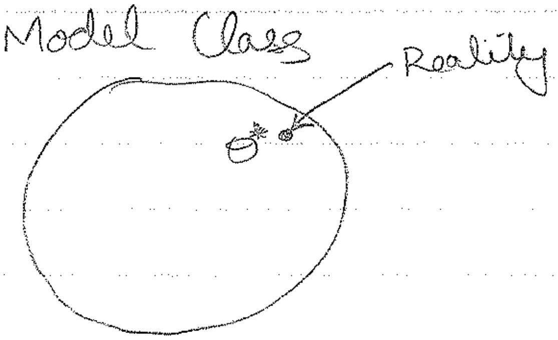
$$\begin{array}{cc} 0, & 1 \\ L, & W \\ H, & T \end{array}$$

→ Hypothesis Class / Model Class or simply "Model"

$$Y \sim \text{Bernoulli}(\theta) \iff \begin{array}{l} P(Y=1) = \theta \\ P(Y=0) = 1-\theta \end{array}$$

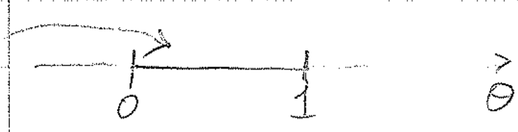
Note: In this example (Nadal W/L), the model is perfect, ie no "approximation/modeling error"

Model Class



[Usually, 'reality' lies outside Model Class]

Or in this case Model Class



→ Learning/Estimation Goal:

Given $D = \{1, 0, 0, 1, 1\}$

estimate $\hat{\theta} \in [0, 1]$

→ How? What's a "good" θ?

$\equiv$ best "explains" the data

$\equiv$ makes it likely for us to have observed D

[e.g if θ were $= 0$, we would never observe a 1]

Formally, let's maximize the prob of D under θ

→ MLE

$$\hat{\theta}_{MLE} = \underset{\theta \in [0\ 1]}{\arg\max} \ \underbrace{P(D \mid \theta)}_{\text{called Likelihood function}}$$

$$L_D(\theta) = P(D \mid \theta)$$

[or sometimes $L(\theta ; D)$]

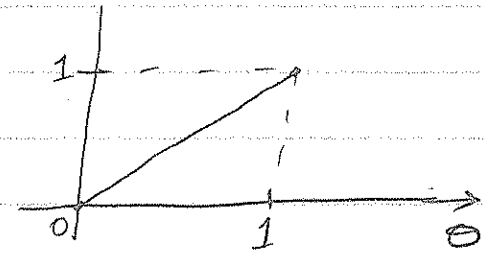
IMP: Likelihood $L_D(\theta)$ or simply $L(\theta)$ is a function of θ.

D is fixed to what we observed.

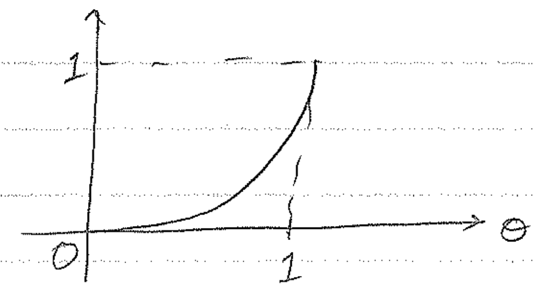
In our running example,

$$L(\theta) = P(D \mid \theta) = \prod_{i=1}^{n} P(y_i \mid \theta) \quad \text{[Why? Hint: IID]}$$

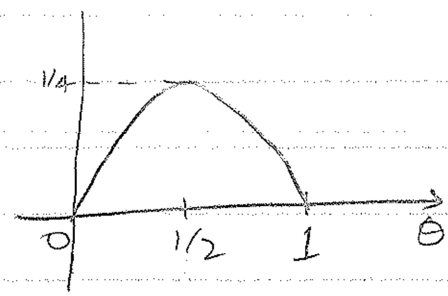
→ e.g. $D = \{1\}$ $\quad L(\theta) = P(Y=1 \mid \theta) = \theta$



→ $D = \{1, 1\} \Rightarrow L(\theta) = \theta \cdot \theta = \theta^2$



→ $D = \{1, 0\} \Rightarrow L(\theta) = P(Y=1 \mid \theta) \cdot P(Y=0 \mid \theta)$
$\qquad = \theta \cdot (1-\theta)$



→ In general, $L(\theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$

α_H = # Heads / Wins
α_T = # Tails / Losses

$$\hat{\theta}_{MLE} = \underset{\theta \in [0\ 1]}{\text{argmax}}\ L(\theta)$$

$$= \underset{\theta \in [0\ 1]}{\text{argmax}}\ \overbrace{\log L(\theta)}$$

log-likelihood or $LL(\theta)$

[Why? ∵ log is a monotone function, so preserves argmax.]

How do we find argmax of $LL(\theta)$?

Elementary, my dear Calculus!

Take 1st deriv; set to zero

$$\frac{\partial LL(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left[\alpha_H \log \theta + \alpha_T \log(1-\theta) \right]$$

$$= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} \qquad \left[\text{assuming } \begin{matrix} \theta \neq 0 \\ \theta \neq 1 \end{matrix}\right]$$

$$= \frac{\alpha_H - \alpha_H \theta - \alpha_T \theta}{\theta(1-\theta)} = 0$$

$$\Rightarrow \boxed{\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

② Sufficient Statistic

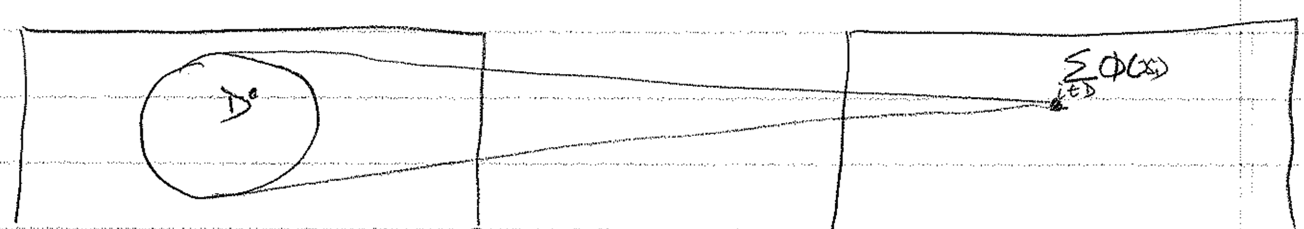
$\theta(X)$ is a sufficient statistic iff

$$\underbrace{\sum_{i \in D_1} \phi(x_i) = \sum_{i \in D_2} \phi(x_i)}_{\text{Same statistics}} \Rightarrow \underbrace{L(\theta; D_1) = L(\theta; D_2)}_{\text{Datasets appear "equivalent" to likelihood fn.}}$$

e.g. $D_1 = \{1, 1, 1, 0, 0, 0\}$ $\quad \alpha_H = \alpha_T = 3$

$D_2 = \{1, 0, 1, 0, 1, 0\}$ $\quad \alpha_H = \alpha_T = 3$

$\Rightarrow \phi(Y) = [\![Y = 1]\!]$



Datasets

Statistics

(4) Why MLE? MLE is OPT if model class is correct!

→ Note: Strong Statement. Strong Assumption.

Consider, Y discrete (but argument generalizes)

$$\frac{1}{N} LL(\theta) = \frac{1}{N} \sum_{i=1}^{N} \log P(Y = y_i \mid \theta)$$

$$= \frac{1}{N} \left[\underbrace{\#(Y=1)}_{\text{count in dataset } D} \cdot \log P(Y=1 \mid \theta) + \#(Y=2) \cdot \log P(Y=2 \mid \theta) + \cdots \right]$$

[∵ Counting argument]

As data becomes infinite,

$$\lim_{N \to \infty} \frac{\#(Y=1)}{N} = P(Y = 1 \mid \theta^*)$$

← True parameter or unknown "reality"

Let's use shorthand $\left\{ \begin{array}{l} P^*(y) = P(Y=y \mid \theta^*) \\ P_\theta(y) = P(Y=y \mid \theta) \end{array} \right\}$

Now $\frac{1}{N} LL(\theta)$ as $N \to \infty$

$$= \sum_{y=1}^{k} P^*(y) \log P_\theta(y)$$

$$= \sum_{y=1}^{k} P^*(y) \log \left[\frac{P_\theta(y)}{P^*(y)} \cdot \frac{P^*(y)}{1} \right]$$

$$= \underbrace{\sum_y P^*(y) \log P^*(y)}_{\text{entropy}} - \underbrace{\sum_y P^*(y) \log \frac{P^*(y)}{P_\theta(y)}}_{\text{KL-Divergence}}$$

$$\frac{1}{N} LL(\theta) = \underbrace{-H(p^*)}_{\substack{\text{entropy/} \\ \text{uncertainty} \\ \text{in } p^* \\ \equiv \text{Constant} \\ \text{wrt } \theta}} \underbrace{- KL(p^* \,||\, p_\theta)}_{\text{How far is } p_\theta \text{ from } p^*} \quad (\text{as } N \to \infty)$$

$$\Rightarrow \underset{\theta}{\operatorname{argmax}} \; LL(\theta) = \operatorname{argmin} \; KL(p^* \,||\, p_\theta)$$

Very cool!

POWERFUL RESULT

→ We did not specify $P(Y=y|\theta)$ or the "Model Class"

→ Result valid for any model!

Caveat

→ Inf dat

→ We must know the "true" model P_θ, which we usually don't (e.g. Is life Gaussian?)

5 MAP + Bayesian Estimation

Key intuition: Let's think of θ as a random quantity & apply Bayes Rule

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

Note: In Frequentist statistics, θ is unknown but not a random quantity
[e.g there is only 1 world with 1 Nadal $\Rightarrow \theta$ is fixed]
So can't talk about $P(\theta)$, but in Bayesian Stats:

$P(\theta)$ = Prior Belief
= what do we believe about θ without any data

In our running example $\theta \in [0, 1]$
$\Rightarrow$ need a continuous distribution / density function
$\Rightarrow$ a distribution over parameter of another distribution

$\longrightarrow$ Meet the Beta distribution

$$P(\theta | \beta_H, \beta_T) = \frac{\theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}}{\text{constant}}$$

hyper-parameters: parameters of distribution over parameter (θ)

Important Facts:

$\longrightarrow$ constant = $\underbrace{\int_0^1 \theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1} \, d\theta}$

Because pdf integrates to 1

$\rightarrow$ mode of pdf

$$\hat{\theta}_{mode} = \frac{\beta_H - 1}{\beta_H + \beta_T - 2}$$

Maximum A Posteriori (MAP) Estimation

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \; P(\theta \mid D)$$ [Let's pick the one value we "believe" in the most]

$$= \underset{\theta}{\operatorname{argmax}} \; \frac{P(D \mid \theta) \, P(\theta)}{P(D)}$$

constant w.r.t θ

$$= \underset{\theta}{\operatorname{argmax}} \; P(D \mid \theta) \, P(\theta)$$

Special Case: $P(\theta) = $ constant [all θs are equally likely]

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \; P(D \mid \theta)$$

$$= \hat{\theta}_{MLE}$$ [Very nice. So frequentists are just Bayesians with no priors!]

In our setup,

$$P(\theta|D) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T} \times \underbrace{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}_{\text{constant}}$$

$$\propto \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}$$

$$\propto \underbrace{\text{Beta}(\alpha_H+\beta_H, \alpha_T+\beta_T)}$$

Very Nice! Conjugate Priors make math easy!

$\Rightarrow$ $\hat{\theta}_{MAP}$ = mode of posterior Beta

$$= \frac{\alpha_H+\beta_H-1}{\alpha_H+\beta_H+\alpha_T+\beta_T-2}$$

So basically β_H, β_T act as "pseudo-flips"
→ experiments/dat not contained in our dataset

Special Case : $\alpha_H = \beta_H = 1$

$$\hat{\theta}_{MAP} = \frac{\alpha_H+1-1}{\alpha_H+1+\alpha_T+1-2} = \hat{\theta}_{MLE}$$

Why? $\because$ $P(\theta) \propto \theta^{1-1}(1-\theta)^{1-1}$ = constant / uniform prior