2/4/14

BN PARAMETER LEARNING

① Setup:

→ Random Variables: $\vec{X} = \{X_1 \ldots X_n\}$

→ "True" Distribution $\sim P^*(X_1 \ldots, X_n)$

→ Fully Observed Dataset of samples from P^*

$$\begin{bmatrix} x_1^{(1)} & \ldots & x_n^{(1)} \\ x_1^{(2)} & \ldots & x_n^{(2)} \\ \vdots & & \\ x_n^{(M)} & \ldots & x_n^{(M)} \end{bmatrix}_{M \times n}$$

→ Known BN structure ie assume P^* factorizes to G.

$$P^*(x_1 \ldots, x_n) = \prod_i P(x_i \mid \underbrace{Pa_{x_i}}_{\text{parents in } G})$$

Goal: estimate CPTs

Tools: → Frequentist: Max-Likelihood

→ Bayesian: Maximum a Posteriori

or full posterior estimation

② Simplest BN: 1 RV: X

CPT $P(X) = \begin{bmatrix} 1-\theta \\ \theta \end{bmatrix} \begin{matrix} x=0 \,/ T \\ x=1 \,/ H \end{matrix}$

Notes on next page from ECE 4984/5984

(2) Max Likelihood Estimation

(Sample Space)

Ω = {Nadal Loses (L), Nadal Wins (W)}

Random Variable

$x \in \{0, 1\}$

L, W

T, H

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

$P(X=1) = \theta$

$P(X=0) = 1-\theta$

Given:

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

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

[e.g if $\theta = 0$, we would never see 1]

Idea: Let's maximize the probability of D under θ

$$\hat{\theta}_{MLE} = \underset{\theta \in [0,1]}{\operatorname{argmax}} \underbrace{P(D \mid \theta)}_{\text{called likelihood function}}$$

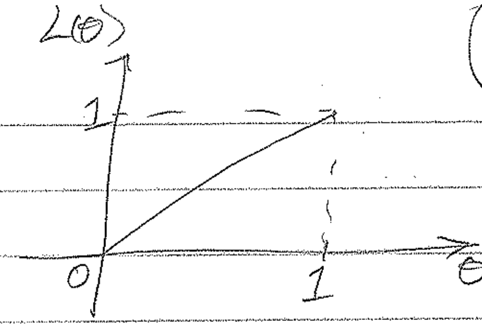
$L(\theta; D) = P(D \mid \theta)$

IMP: Likelihood is a function of θ

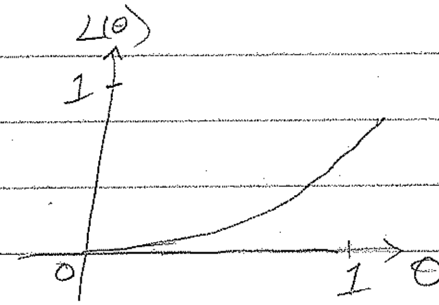
(D is fixed to what we observed)

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

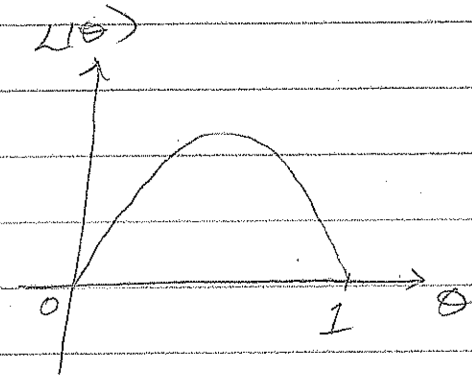
E.g $D=\{1\}$ $\quad L(\theta)=\theta$



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



$D=\{1,0\}$ $\quad L(\theta)=\theta\cdot(1-\theta)$



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

α_H = # Heads/Wins

α_T = # Tails/Losses

$$\hat{\theta}_{MLE} = \underset{\theta\in[0,1]}{\arg\max}\ L(\theta)$$

called log-likelihood

$$= \underset{\theta}{\arg\max}\ \log L(\theta)$$ [Why? ∵ log is a monotone function so preserves argmax]

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

Take 1st derivative, set to zero

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

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

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

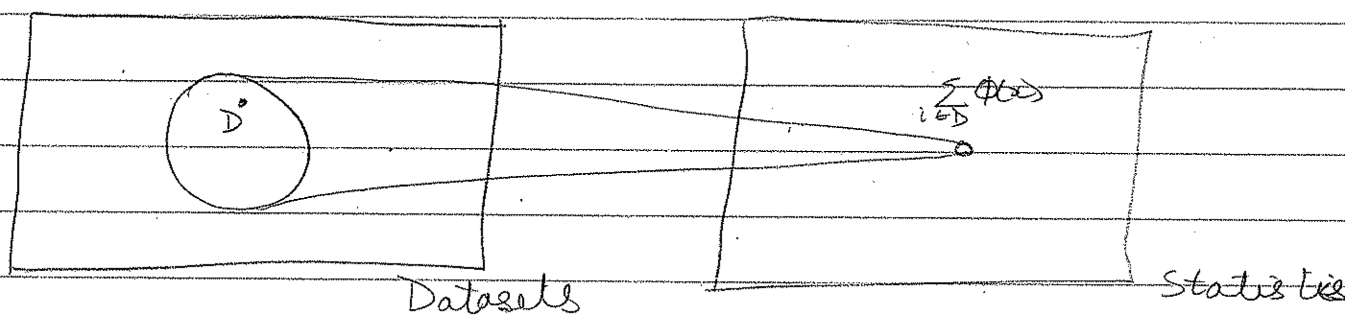
(ignoring boundary conditions) $\Rightarrow$ $\boxed{\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$

③ Sufficient statistics

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

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

Two datasets but look the same to Likelihood.



④ MLE is OPT if model class is correct.

Consider $\frac{1}{N} LL(\theta) = \frac{1}{N} \sum_{i=1}^{n} \log P(X_i \mid \theta)$

$$= \frac{1}{N}\left[\#(X=1)\, P(X=1 \mid \theta) + \#(X=2)\, P(X=2 \mid \theta) + \dots \right]$$

(counting argument)

As data becomes infinite

$$\lim_{N\to\infty} \frac{\#(X=1)}{N} = P(X=1 \mid \theta^*)$$ ← true parameter (unknown)

Shorthand $\left\{ \begin{array}{l} P^*(x) = P(X \mid \theta^*) \\ P_\theta(x) = P(X \mid \theta) \end{array} \right\}$

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

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

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

$$= \sum_{i=1}^{k} P^*(x) \log P^*(x) - \sum_{i=1}^{k} P^*(x) \log \frac{P^*(x)}{P_\theta(x)}$$

$$\frac{1}{N} LL(\theta) = -H(P^*) - KL(P^* \,\|\, P_\theta)$$

so $\max \frac{1}{N} LL(\theta) \equiv \min_\theta KL(P^* \,\|\, P_\theta)$ $\left[\because H(P^*) \text{ is a constant} \right]$

POWERFUL RESULT: → We did not specify $P(X \mid \theta)$
→ Any distribution

Caveat: → Inf data
→ We must know the family P_θ, which we usually don't (e.g. is life Gaussian?)

(5) MAP + Bayesian Estimation

Let's think of θ as a random quantity & apply Bayes Rule: $P(\theta | D) = \dfrac{P(D|\theta)\, P(\theta)}{P(D)}$

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

$\theta \in [0,1]$ So we need a distribution over parameters of our distribution

Beta distribution:

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

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

Important Facts:

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

$$\text{mode of distribution} = \frac{\beta_H - 1}{\beta_H + \beta_T - 2}$$

Maximum A Posteriori (MAP) Estimation

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta | D)$$

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

constant w.r.t θ

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

$$= \arg\max_{\theta} \theta^{\alpha_H} (1-\theta)^{\alpha_T} \times \theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}$$

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H + \beta_H - 1} (1-\theta)^{\alpha_T + \beta_T - 1}}_{Beta(\alpha_H + \beta_H,\ \alpha_T + \beta_T)}$$

= mode of Beta

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

Very Nice, so β_H, β_T act as pseudo-flips/coun

→ Phantom experiments not contained in our datase

Special Cases: $\beta_H = \beta_T = 1$

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

When $\beta_H = \beta_T = 1$

$$\text{Beta} \equiv \underbrace{\theta^0 (1-\theta)^0}_{\text{uniform distribution}}$$

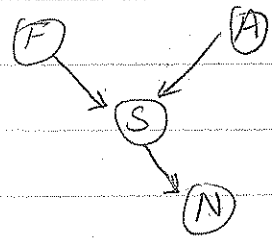
No prior $\Rightarrow$ posterior = prior

When $N \to \infty$ effect of β_H, β_T is forgotten.

Back to ECE 6504 Notes.

③ All ideas extend to general BN CPT estimation

example: G:



Vars: F, A, S, N (all binary)

CPTs: $P(F) = \begin{bmatrix} 1-\theta_F \\ \theta_F \end{bmatrix}$ $P(A) = \begin{bmatrix} 1-\theta_A \\ \theta_A \end{bmatrix}$

$P(S \mid F, A) =$

S \ F, A	00	01	10	11
0	1-(.)	1-(.)	1-(.)	1-(.)
1	θ_{S100}	θ_{S101}	θ_{S110}	θ_{S111}

$P(N \mid S) =$

N \ S		
	1-(.)	1-(.)
	θ_{N10}	θ_{N11}

Parameters to be estimated: $\vec{\theta} = \{ \theta_F, \theta_A, \theta_{S100}, \theta_{S101}, \theta_{S110}, \theta_{S111}, \theta_{N10}, \theta_{N11} \}$

MLE: $\vec{\theta}_{MLE} = \underset{\vec{\theta}}{\text{argmax}} \log P(D \mid \vec{\theta}, G)$

$$= \underset{\vec{\theta}}{\text{argmax}} \sum_{j=1}^{M} \log P(F=f^{(j)}, A=a^{(j)}, S=s^{(j)}, N=n^{(j)} \mid \vec{\theta}, G)$$

$$= \underset{\vec{\theta}}{\text{argmax}} \sum_{j=1}^{M} \underbrace{\Big[\log P(F=f^{(j)} \mid \vec{\theta}, G) + \log P(A=a^{(j)} \mid \vec{\theta}, G) + \log P(S=s^{(j)} \mid F=f^{(j)}, A=a^{(j)}, \vec{\theta}, G) + \log P(N=n^{(j)} \mid S=s^{(j)}, \vec{\theta}, G)}_{\text{independent subproblems}}$$

independent subproblems
each term depends only on its own θ

Same argument as for a single variable

In general

$$\theta_{X_i=a \mid Pa_{X_i}=\vec{b}} = \frac{Count(X_i=a,\ Pa_{X_i}=\vec{b})}{Count(Pa_{X_i}=\vec{b})}$$

[For multi-label variables $X_i \in \{1, \ldots, k\}$, need just a bit more math (Lagrange Multipliers)
$\because \sum_a \theta_{X_i=a \mid Pa_{X_i}=\vec{b}} = 1$ constrained optimization; But same result as binary vars]

For MAP estimation

assume $P(\theta_{X_i \mid Pa_{X_i}=\vec{b}}) \sim Dir(\vec{\alpha}_{X_i \mid Pa_{X_i}=\vec{b}})$

Each CPT column has it's own prior given by hyperparameters