Probability Overview:

- Conditional Prob, $P(A|B) = \frac{P(A \wedge B)}{P(B)}$
- Bayes Rule, $P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$

Learning Problem:

- $f : X \rightarrow Y$ (unknown)
  - Concept Learning
  - Decision Tree Learning
- Prob $(Y \mid X)$ → Bayesian Learning (Classification)

Probability Estimations:

- Maximum Likelihood Estimation (MLE)
  - choose θ that maximizes Prob $(Data \mid \theta)$
  - $\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \left(Prob(Data \mid \theta) \right)$
  - #Ex: Coin Flip
- Maximum A Posteriori (MAP)
  - choose θ that maximizes Prob $(\theta \mid Data)$

Smart Estimations

SUMMARY of Previous Lect

Joint Probability Distribution

$\sum_{rows}$ (Events)

MAP: choose θ that max. Prob $(\theta \mid \text{data})$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[\text{Prob}(\theta \mid \text{data}) \right]$$

likelihood, Prior, max, marginal

$$= \underset{\theta}{\operatorname{argmax}} \left[\frac{\text{Prob}(\text{data} \mid \theta) \cdot P(\theta)}{P(\text{data})} \right]$$

#Ex: Coin flip:

$$\hat{\theta} = \frac{\alpha_1 + \#\text{Pre}\,H}{(\alpha_1 + \#\text{Pre}\,H) + (\alpha_0 + \#\text{Pre}\,T)} \propto \left(\frac{\alpha_1}{\alpha_1 + \alpha_0} \right)$$

$$P(\text{data} \mid \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

$$P(\theta) = \frac{\theta^{\beta_H - 1} (1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\theta, \beta_H, \beta_T)$$

Conj Prior

$$\frac{\partial}{\partial \theta} \ln \left[\theta^{\alpha_H + \beta_H - 1} (1-\theta)^{\alpha_T + \beta_T - 1} \right] = 0$$

$$\Rightarrow \hat{\theta}_{MAP} = \frac{\alpha_H + (\beta_H - 1)}{(\alpha_H + (\beta_H - 1)) + (\alpha_T + (\beta_T - 1))}$$

$$P(\text{data}|\theta) = \theta_1^{\alpha_1}\,\theta_2^{\alpha_2}\cdots\theta_K^{\alpha_K} \quad \left(\sum_{i=1}^{K}\theta_i = 1\right)$$

$$P(\theta) = \frac{\theta_1^{\beta_1-1}\,\theta_2^{\beta_2-1}\cdots\theta_K^{\beta_K-1}}{D(\beta_1,\ldots,\beta_K)} \sim \text{Dirichlet}(\theta, \beta_1,\ldots,\beta_K)$$

$$\hat{\theta}_{i\,MAP} = \frac{\alpha_i + (\beta_i - 1)}{\sum_{j=1}^{K}\alpha_j + (\beta_0 - 1)}$$

- Joint Dist Table ⟶ Dennril
  - ↳ Data Sparsity!!
- ⟶ MLE : Data Repr. Sufficient
- ⟶ MAP : Prior Knowledge but less Data
  - Bayes Rule

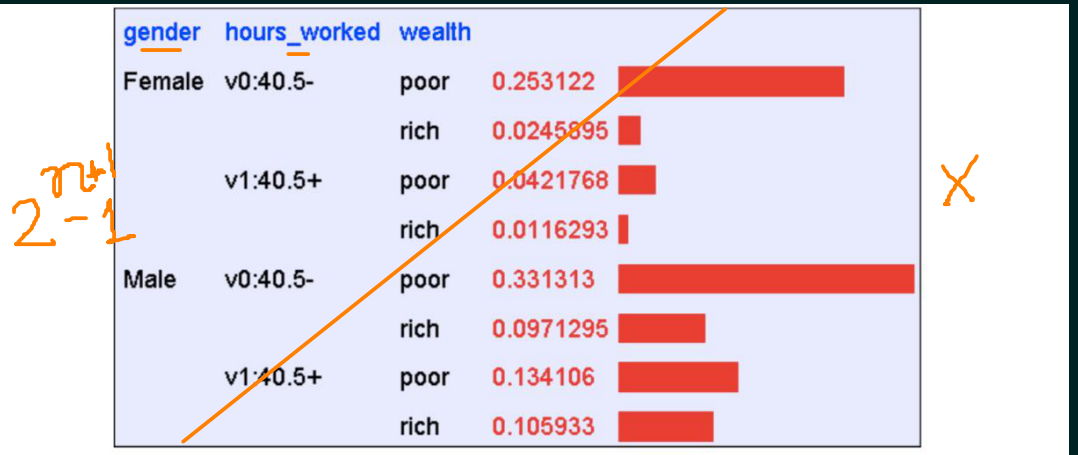
$P(Y=0 \mid X_1 X_2 \cdots X_n) \rightarrow ?$

$\#Est = 2^n$

$Y=0 \leftarrow X^{new} = \langle \; \rangle$

$Y=1$

$Prob(Y=0 \mid X^{new})$?



Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

2^2

$P(Y \mid X_1 \cdots X_n) = \frac{P(X_1 \cdots X_n \mid Y) P(Y)}{P(X_1 \cdots X_n)}$

$P(W \mid G, HW) = ?$

$2^n - 1 \leftarrow P(X_1 \cdots X_n \mid Y=0)$

$2^n - 1 \leftarrow P(X_1 \cdots X_n \mid Y=1)$

$P(Y) \leftarrow 1$

$P(X_1 \cdots X_n)$ ←

$2(2^n - 1) + 1$

100 attri → MLE, MAP

$P(Y=1 \mid X_1 \cdots X_n)$ → 2^n Training

$P(Y=? \mid X^{new}) \rightarrow$ classify

$$P(Y=1 \mid X_1 \dots X_n) = ?$$

$$= \frac{P(X_1 \dots X_n \mid Y)\,P(Y)}{P(X_1 \dots X_n)}$$

assume

$$P(X_1 \dots X_n \mid Y) = \prod_{i=1}^{n} P(X_i \mid Y)$$

$$P(Y \mid \langle X_1 \dots X_n \rangle) = \frac{P(\langle X_1 \dots X_n \rangle \mid Y)\,P(Y)}{P(X_1 \dots X_n)}$$

$$= \frac{P(Y)\cdot \prod_{i=1}^{n} P(X_i \mid Y)}{P(X_1 \dots X_n)}$$

$2n + 1$ est $O(2n)$ (2^n)

$n = 10^4$

Conditional Independence

def: $X \perp Y \mid Z$ if

$$P(X \mid YZ) = P(X \mid Z)$$

Ex: $P(T \mid R \wedge L) = P(T \mid L)$

$$P(X_1 X_2 \mid Y) = P(X_1 \mid X_2 Y)\,P(X_2 \mid Y)$$

$$= P(X_1 \mid Y)\,P(X_2 \mid Y)$$

Naive Bayes Algorithm

- Spam filtering
- News Article Separation
- Text Classification

$$P(Y=y_k \mid x_1 \dots x_n) = \frac{P(Y=y_k)\prod_{i=1}^{n} P(x_i \mid Y=y_k)}{\sum_j P(Y=y_j)\prod_{i=1}^{n} P(x_i \mid Y=y_k)}$$

$P(B)$
$= P(B \mid \bar{A}) + P(B \cap A)$
$= P(B \mid \bar{A})\,P(\bar{A}) + P(B \mid A).P(A)$

x_1	x_2	x_3	Y
good	Red	W	+
Bore	Ou.	B	−
bad			

$$P(Y=+ \mid B_o + O + W) = \frac{P(Y=+)\left[P(B_o \mid Y=+)\,P(O \mid Y=+)\,P(W \mid Y=+)\right]}{\cdots}$$

X_1	X_2	X_3	Y
G	R	B	−
G	O	B	−
Bore	R	W	+
Bad	O	B	+
G	R	W	+

1/3

MLE

TE

X^{new}

$[B_o\ O\ W]$

$Y = ?$