

SUMMARY of Previous Lect

Joint Probability Distribution

$$\sum_{\text{rows}} (\text{Events})$$

▶ 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 | X) → Bayesian Learning (classification)

▶ Probability Estimations:

↳ Maximum Likelihood Estimation (MLE)

choose θ that maximizes $\text{Prob}(\text{Data} | \theta)$

#Ex: Coin Flip

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} (\text{Prob}(\text{Data} | \theta))$$

Smart Estimations

↳ Maximum A Posteriori (MAP)

choose θ that maximizes $\text{Prob}(\theta | \text{Data})$

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

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\text{argmax}} [\text{Prob}(\theta | \text{data})]$$

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

Annotations:
 - Likelihood: $\text{Prob}(\text{data} | \theta)$
 - Prior: $P(\theta)$
 - Marginal: $P(\text{data})$
 - Max: $\underset{\theta}{\text{argmax}}$

#Ex: Coin flip: $\hat{\theta} = \frac{\alpha_1 + \# \text{PreH}}{(\alpha_1 + \# \text{PreH}) + (\alpha_0 + \# \text{PreT})}$

Annotations:
 - α_1 : α_1
 - α_0 : α_0
 - $\# \text{PreH}$: $\# \text{PreH}$
 - $\# \text{PreT}$: $\# \text{PreT}$

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

Annotations:
 - α_H : α_H
 - α_T : α_T
 - β_H : β_H
 - β_T : β_T
 - $B(\beta_H, \beta_T)$: $B(\beta_H, \beta_T)$
 - $\text{Beta}(\theta, \beta_H, \beta_T)$: $\text{Beta}(\theta, \beta_H, \beta_T)$
 - Conj Prior : Conj Prior

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

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

Annotations:
 - $\alpha_H + (\beta_H - 1)$: $\alpha_H + (\beta_H - 1)$
 - $(\alpha_H + (\beta_H - 1)) + (\alpha_T + (\beta_T - 1))$: $(\alpha_H + (\beta_H - 1)) + (\alpha_T + (\beta_T - 1))$

$$P(\text{data}|\theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \dots \theta_k^{\alpha_k} \left(\sum_{i=1}^k \theta_i = 1 \right)$$



$$P(\theta) = \frac{\theta_1^{\beta_1-1} \theta_2^{\beta_2-1} \dots \theta_k^{\beta_k-1}}{D(\beta_1, \dots, \beta_k)} \sim \text{Dirichlet}(\theta, \beta_1, \dots, \beta_k)$$

$$\hat{\theta}_{MAP}^i = \frac{\alpha_i + (\beta_i - 1)}{\sum_{j=1}^k \alpha_j + (\beta_j - 1)} \quad \checkmark$$

▶ Joint Dist Table → Demerit

↳ Data Sparsity!!

→ MLE : Data Repr.
Sufficient

→ MAP : Prior Knowledge but less Data
Bayes Rule

$$\rightarrow P(Y=0 | x_1, x_2, \dots, x_n) \rightarrow ?$$

$$\#Est = 2^n$$



$$\text{Prob}(y=0 | x^{new}) ?$$

$$P(y | x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n | y) P(y)}{P(x_1, \dots, x_n)}$$

$$2^{n-1} \leftarrow P(x_1, \dots, x_n | y=0)$$

$$2^{n-1} \leftarrow P(x_1, \dots, x_n | y=1)$$

$$P(W | G, HW) = ?$$

$$P(y) \leftarrow 1$$

$$P(x_1, \dots, x_n) \leftarrow$$

$$\left. \begin{matrix} P(y) \leftarrow 1 \\ P(x_1, \dots, x_n) \leftarrow \end{matrix} \right\} \frac{2(2^n - 1) + 1}{2^n}$$

100 attri $\left\{ \begin{matrix} \rightarrow \text{MLE} \\ \rightarrow \text{MAP} \end{matrix} \right\}$

$$P(y=1 | x_1, \dots, x_n)$$

$\rightarrow 2^n$ Training

$P(y=1 | x^{new}) \rightarrow$ classify

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
v1:40.5+	poor	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
v1:40.5+	poor	poor	0.134106
		rich	0.105933

2^{n+1}
 $2-1$

X

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

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

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

assume

$$P(X_1, \dots, X_n | Y)$$

$$= \prod_{i=1}^n P(X_i | Y)$$

$$P(Y | \langle X_1, \dots, X_n \rangle) =$$

$$\frac{2^n + x}{2^n} \quad \text{est} \quad \frac{O(2^n)}{2^n} = \frac{2^n}{2^n} = 1$$

$n = 10^6$

Conditional Independent

Def: $(X \perp\!\!\!\perp Y) | Z$ if

$$P(X | YZ) = P(X | Z)$$

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

$$P(X_1, X_2 | Y) = P(X_1 | X_2 Y) P(X_2 | Y)$$

$$= P(X_1 | Y) P(X_2 | Y)$$

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

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

Naive Bayes Algorithm

Spam filtering →

News Article Separation →

Text Classification →

$$P(Y = y_k | x_1, \dots, x_n)$$

$$= \frac{P(Y = y_k) \prod_{i=1}^n P(x_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_{i=1}^n P(x_i | Y = y_j)}$$

$$P(B) = P(B|\bar{A}) + P(B|A) = P(B|\bar{A})P(\bar{A}) + P(B|A)P(A)$$

x_1	x_2	x_3	Y
good	Read	W	+
bore	Doc.	B	-
bad			

x_1	x_2	x_3	Y
G	R	B	-
G	O	B	-
Bore	R	W	+
Bad	O	B	+
G	R	W	+

$$P(Y = + | B_0 + O + W) = P(Y = +) \frac{P(B_0 | Y = +) P(O | Y = +) P(W | Y = +)}{P(B_0 | Y = +) P(O | Y = +) P(W | Y = +) + \dots}$$

MLE

TE

x_{new}

[B O W]

$y = ?$

W