Naïve Bayes

Naïve Bayes

- Probability basics
- Estimating parameters from data
  - Maximum likelihood (ML)
  - Maximum a posteriori estimation (MAP)
- Naïve Bayes

Contents

- **Probability basics**

- Estimating parameters from data
    - Maximum likelihood (ML)
    - Maximum a posteriori estimation (MAP)

- Naive Bayes

Random Variables

- A random variable *x* takes on a defined set of values with different probabilities.
    - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
    - For example, if you poll people about their voting preferences, the percentage of the sample that responds "Yes on Proposition 100" is a also a random variable (the percentage will be slightly differently every time you poll).
- Roughly, <u>probability</u> is how frequently we expect different outcomes to occur if we repeat the experiment over and over ("frequentist" view)

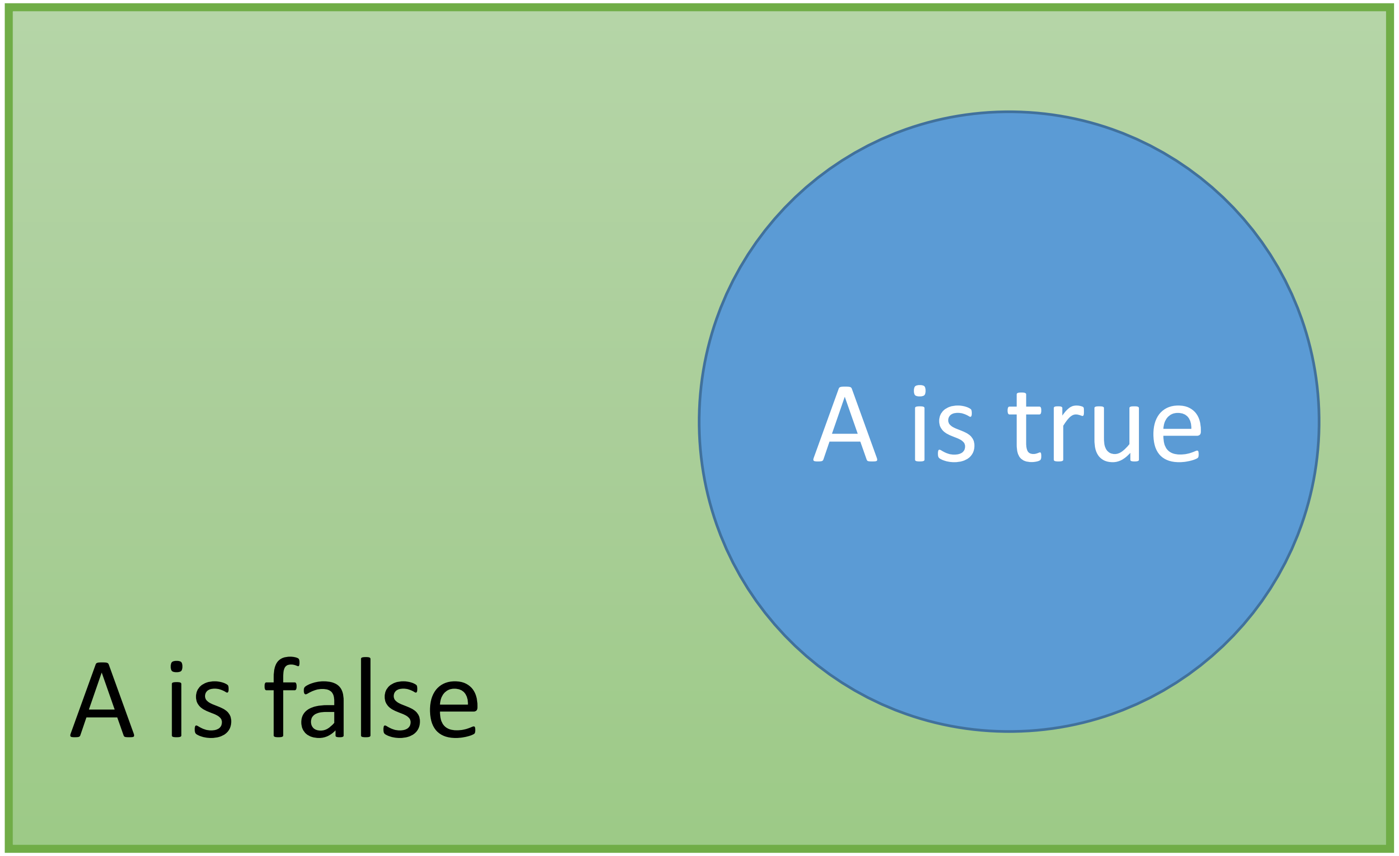
Random variables can be discrete or continuous

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.

Probability functions

- A probability function maps the possible values of *x* against their respective probabilities of occurrence, *p(x)*
- *p(x)* is a number from 0 to 1.0.
- The area under a probability function is always 1.

Visualizing probability $P(A)$



A is true

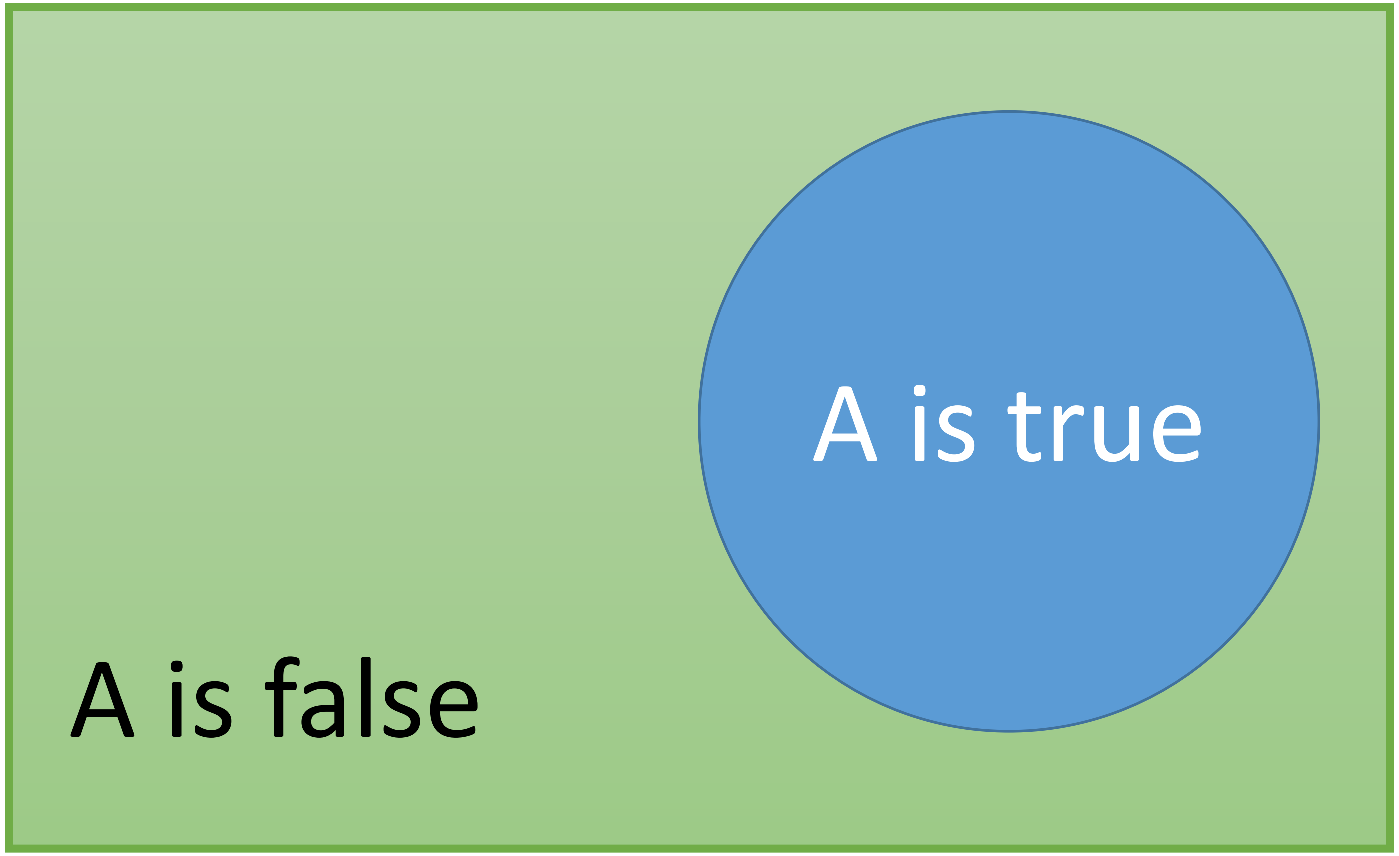
A is false

Sample space

Area = 1

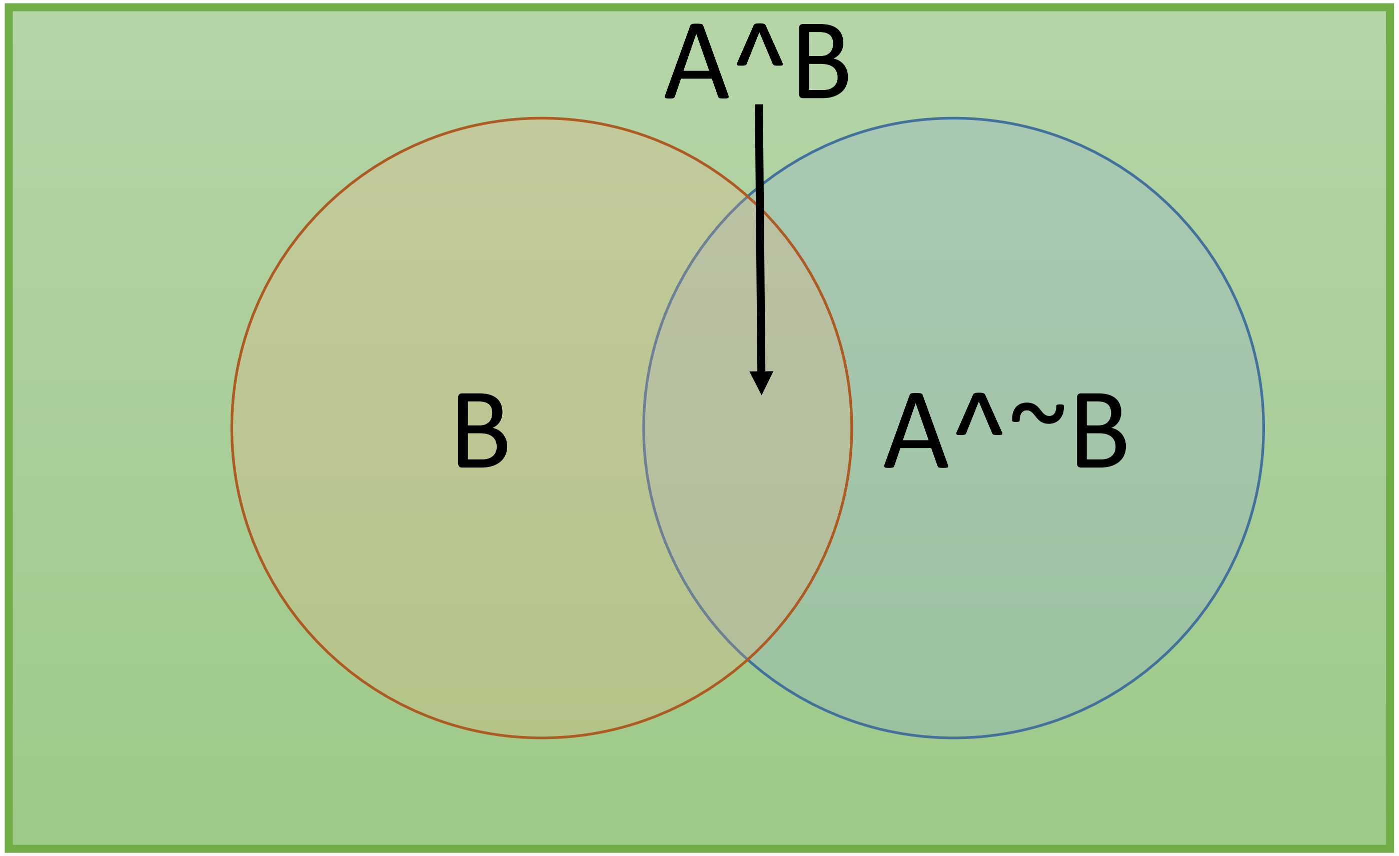
$P(A) =$ Area of the blue circle

Visualizing probability $P(A) + \mathrm{P}(\sim\mathrm{A})$



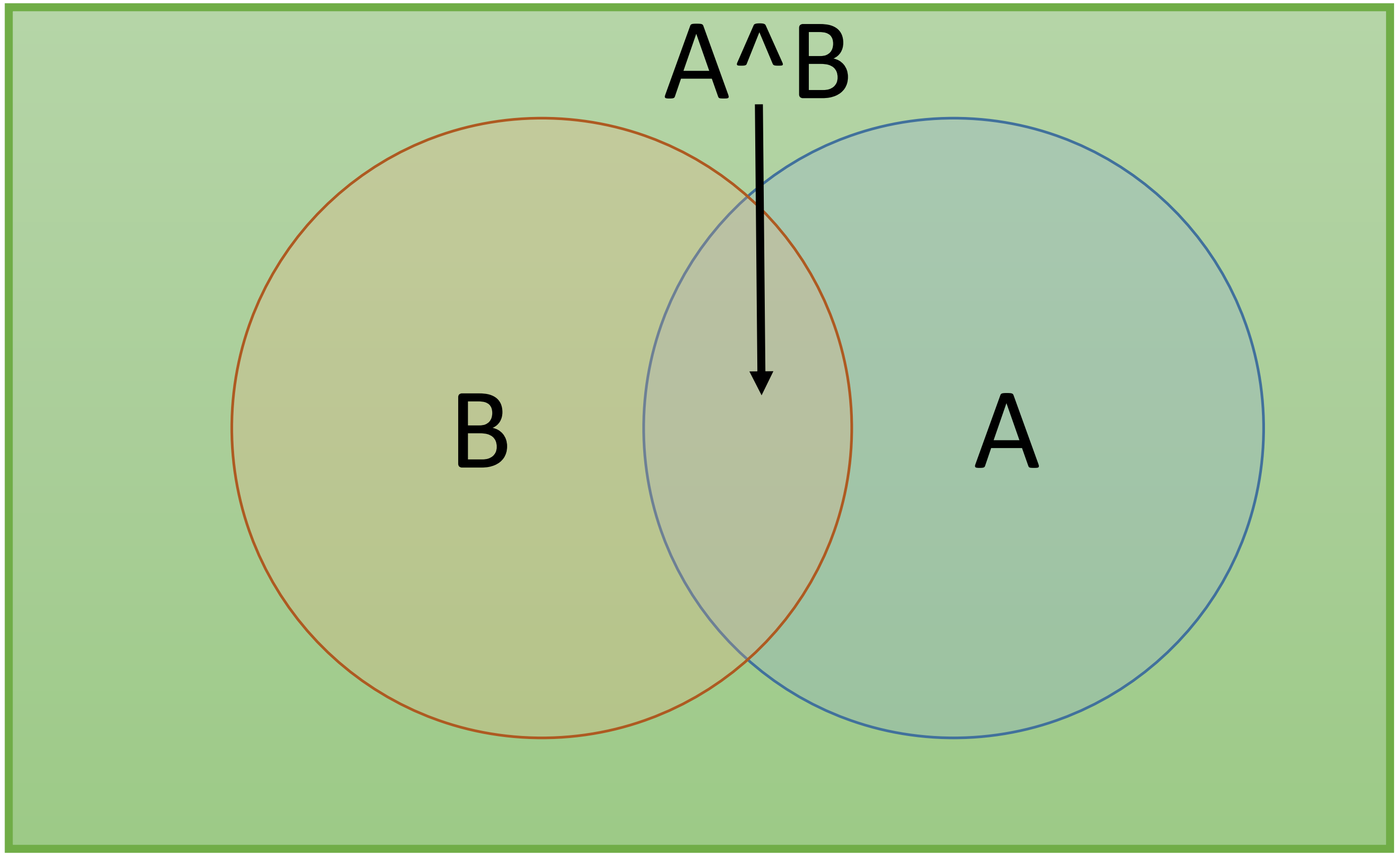
$$P(A) + \mathrm{P}(\sim\mathrm{A}) = 1$$

Visualizing probability $P(A)$



$$P(A) = \mathrm{P(A\hat{}B)} + \mathrm{P(A\hat{}}{\sim}B)$$

Visualizing conditional probability

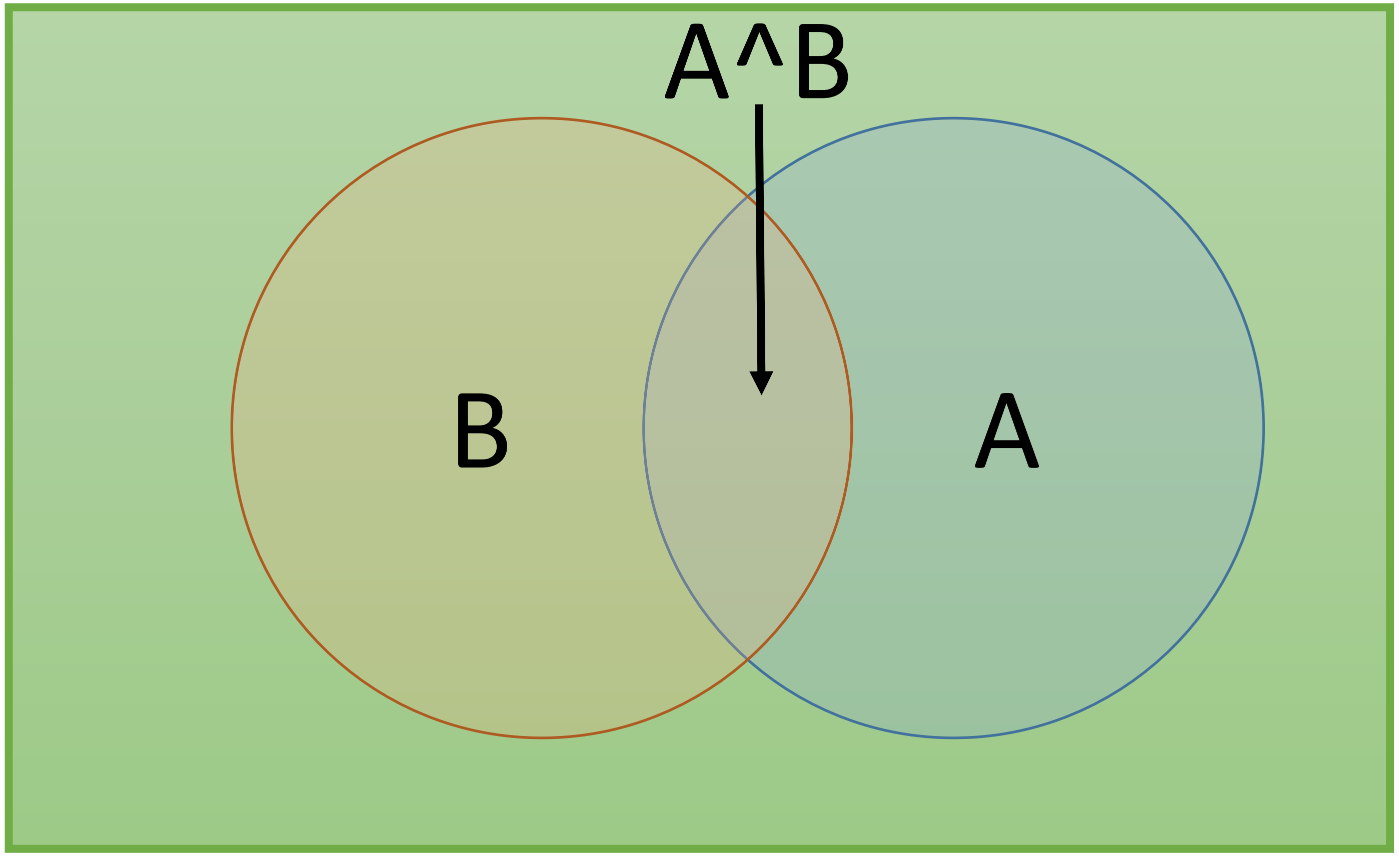


$$P(A|B) = P(A\hat{}B)/P(B)$$

Corollary: The chain rule

$$P(A,B) = P(A|B)P(B)$$

Bayes rule



Thomas Bayes

$$P(A|B) = \frac{P(A,B)}{P(B)}$$
$$= \frac{P(B|A)P(A)}{P(B)}$$

Corollary: The chain rule

$$P(A,B) = P(A|B)P(B) = \mathrm{P(B)P(A|B)}$$

Other forms of Bayes rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B,X) = \frac{P(B|A,X)P(A,X)}{P(B,X)}$$

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

Applying Bayes rule

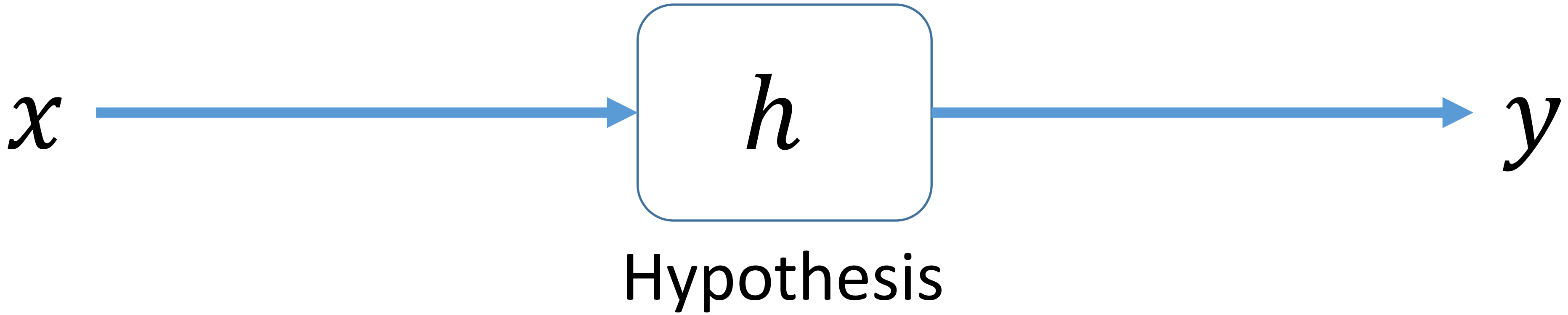
$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|{\sim}A)P({\sim}A)}$$

- A = you have the flu
  B = you just coughed
- Assume:
  - $P(A) = 0.05$
  - $P(B|A) = 0.8$
  - $P(B|{\sim}A) = 0.2$
- What is P(flu | cough) = P(A|B)?
- = 0.17

$$P(A|B) = \frac{0.8 \times 0.05}{0.8 \times 0.05 + 0.2 \times 0.95} \sim 0.17$$

Slide credit: Tom Mitchell

Why we are learning this?



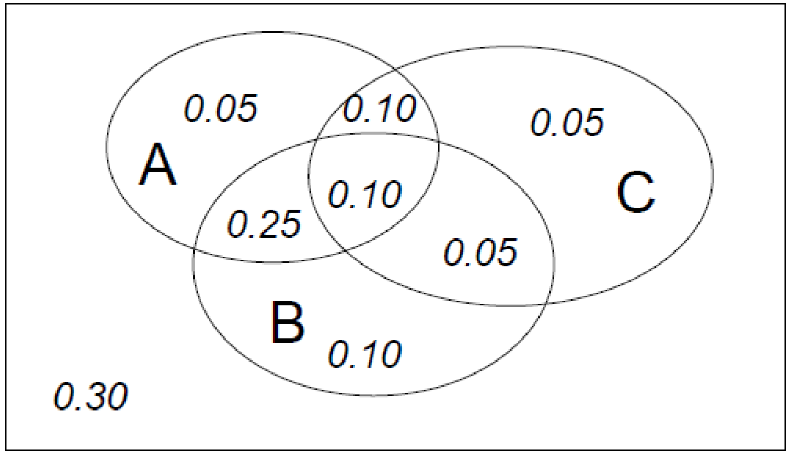
Learn $P(Y|X)$

Joint distribution

- Making a joint distribution of M variables

1. Make a truth table listing all combinations
2. For each combination of values, say how probable it is
3. Probability must sum to 1

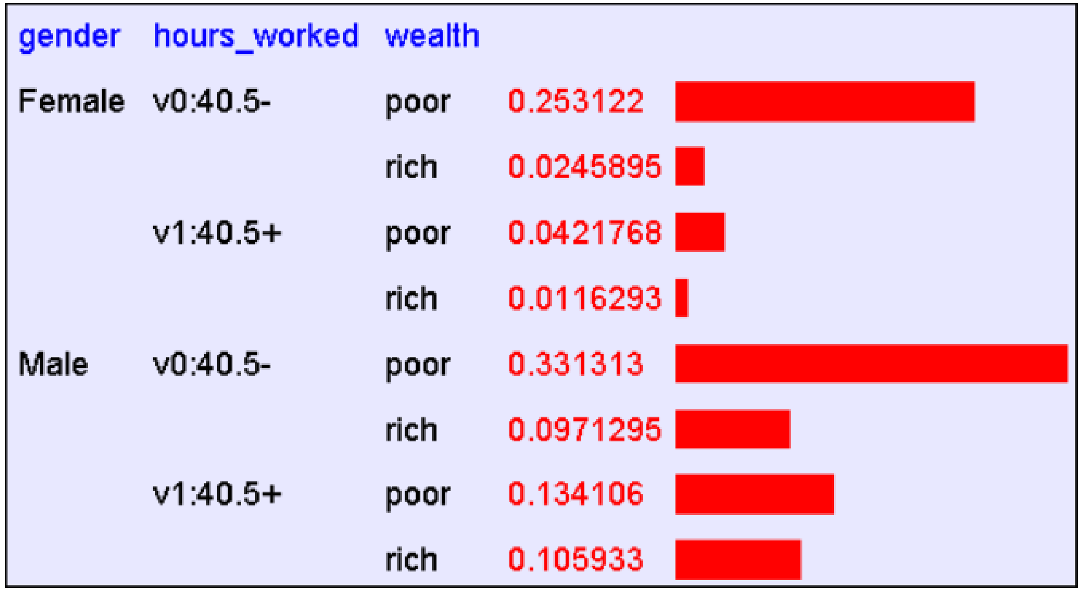
A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Slide credit: Tom Mitchell

Using joint distribution

- Once you have the JD you can ask for the probability of **any** logical expression involving these variables



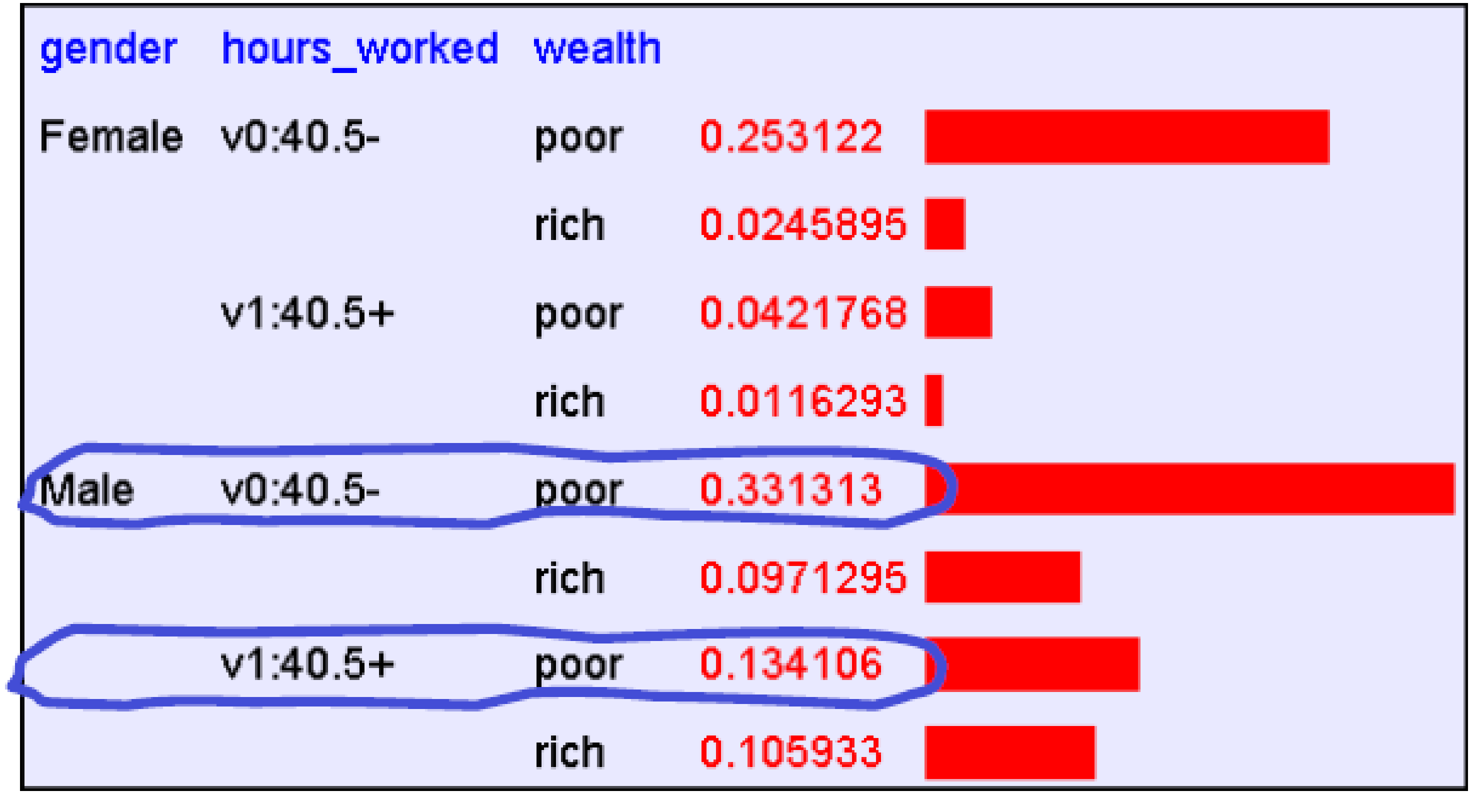
gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

- $P(E) = \sum_{\text{rows matching E}} P(\text{row})$

- $P(E_1|E_2) = \dfrac{\sum_{\text{rows matching E}_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$

Slide credit: Tom Mitchell

Using the Joint

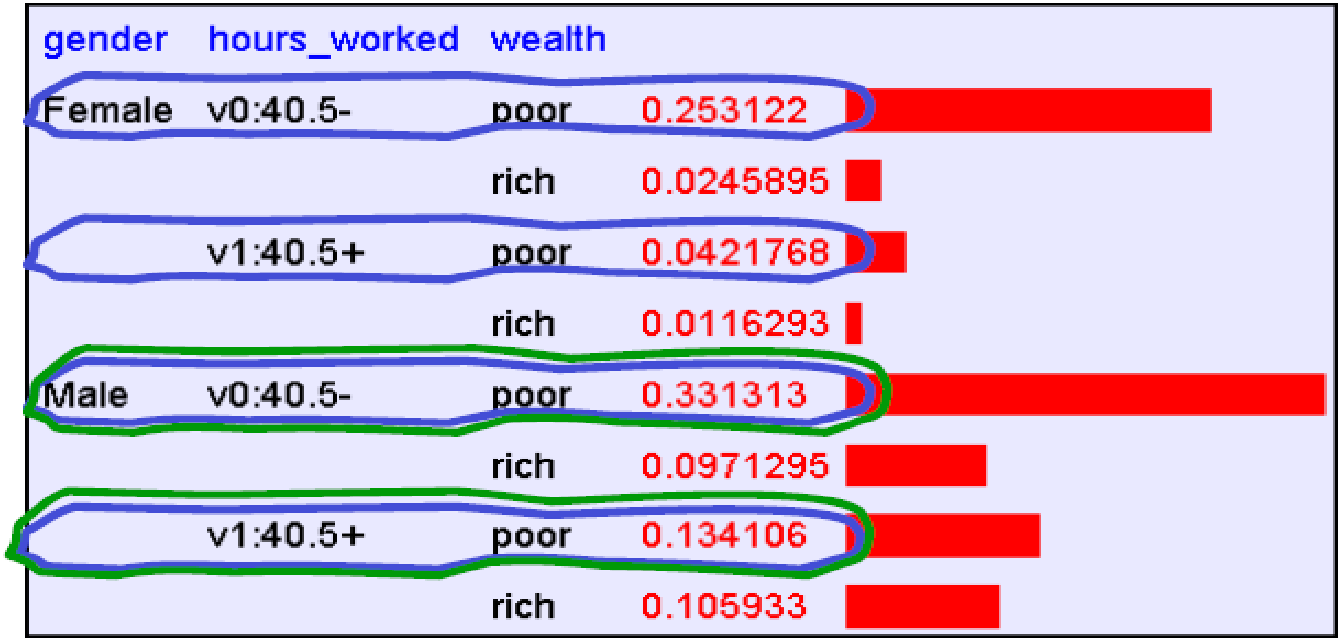


gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

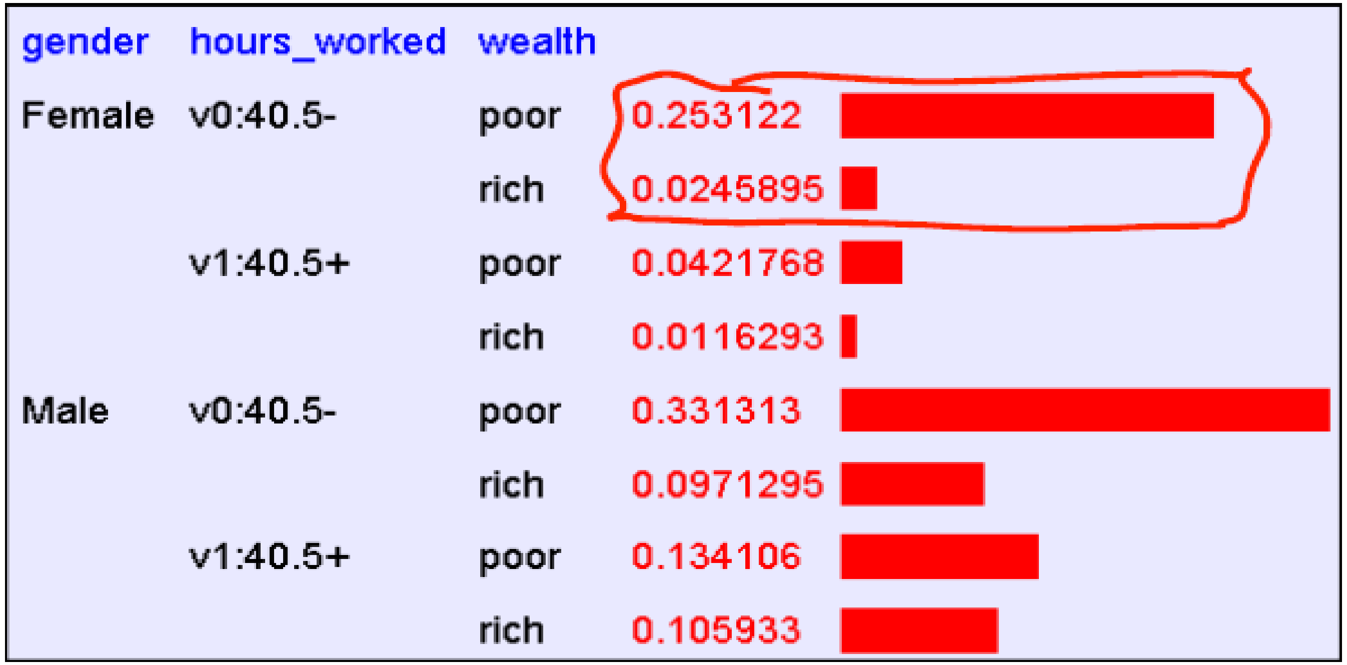


gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum\limits_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum\limits_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Learning and the Joint Distribution



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

Suppose we want to learn the function $f: \langle G, H \rangle \rightarrow W$

Equivalently, $P(W \mid G, H)$

Solution: learn joint distribution from data, calculate $P(W \mid G, H)$

e.g., $P(W=\text{rich} \mid G = \text{female}, H = 40.5-) =$

The solution to learn $P(Y|X)$?

- Main problem: learning $P(Y|X)$ may require more data than we have
- Say, learning a joint distribution with 100 attributes
- # of rows in this table? $2^{100} \geq 10^{30}$
- # of people on earth? 10^{9}

Slide credit: Tom Mitchell

What should we do?

1. Be smart about
   how we estimate probabilities from sparse data
   - Maximum likelihood estimates (ML)
   - Maximum a posteriori estimates (MAP)

2. Be smart about
   how to represent joint distributions
   - Bayes network, graphical models

Slide credit: Tom Mitchell

- Probability basics

- **Estimating parameters from data**
  - **Maximum likelihood (ML)**
  - **Maximum a posteriori (MAP)**

- Naive Bayes

Estimating the probability

$X = 1$ X =0

- Flip the coin repeatedly, observing
  - It turns heads α_1 times
  - It turns tails α_0 times
- Your estimate for $P(X = 1)$ is?

- Case A: 100 flips: 51 Heads ($X = 1$), 49 Tails ($X = 0$)

  $P(X = 1) = ?$

- Case B: 3 flips: 2 Heads ($X = 1$), 1 Tails ($X = 0$)

  $P(X = 1) = ?$

Slide credit: Tom Mitchell

Two principles for estimating parameters

- **Maximum Likelihood Estimate (MLE)**
  Choose θ that maximizes probability of observed data

$$\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}} = \operatorname*{argmax}_{\theta} P(Data|\theta)$$

- **Maximum a posteriori estimation (MAP)**
  Choose θ that is most probable given prior probability and data

$$\widehat{\boldsymbol{\theta}}^{\mathrm{MAP}} = \operatorname*{argmax}_{\theta} P(\theta|D) = \operatorname*{argmax}_{\theta} \frac{P(Data|\theta)P(\theta)}{P(Data)}$$

Slide credit: Tom Mitchell

Two principles for estimating parameters

- **Maximum Likelihood Estimate (MLE)**
  Choose θ that maximizes $P(Data|\theta)$

$$\widehat{\boldsymbol{\theta}}^{\text{MLE}} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

- **Maximum a posteriori estimation (MAP)**
  Choose θ that maximize $P(\theta|Data)$

$$\widehat{\boldsymbol{\theta}}^{\text{MAP}} = \frac{(\alpha_1 + \#\text{halluciated 1s})}{(\alpha_1 + \#\text{halluciated } 1s) + (\alpha_0 + \#\text{halluciated 0s})}$$

Slide credit: Tom Mitchell

Maximum likelihood estimate

$X = 1$ X =0

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

- Each flip yields Boolean value for X

$$X \sim \text{Bernoulli}: P(X) = \theta^X (1 - \theta)^{1-X}$$

- Data set D of independent, identically distributed (iid) flips, produces α_1 ones, α_0 zeros

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1 - \theta)^{\alpha_0}$$

$$\widehat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\, P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Slide credit: Tom Mitchell

Maximum Likelihood Estimation

- Goal : Find the parameter p that maximizes the likelihood of seeing all training samples.
- Example: 6H, 4T
- P(H) = p, P(T) = 1-p

- $L(p) = p^{6}(1-p)^{4}$ ---------→ Find the total likelihood
- $\log L(p) = 6 \log(p) + 4 \log(1-p)$ ----------- > Take log likelihood
- $d(\log L(p))/dp = 6/p - 4/(1-p) = 0$ ---------→ Take derivative
- $P = 6/10$ ---------→ Solve

Classification by likelihood

- Suppose we have two classes C_1 and C_2.
- Compute the likelihoods $P(D|C_1)$ and $P(D|C_2)$.
- To classify test data D’ assign it to class C_1 if $P(D|C_1)$ is greater than $P(D|C_2)$ and C_2 otherwise.

Gaussian models

- Assume that class likelihood is represented by a Gaussian distribution with parameters μ (mean) and σ (standard deviation)

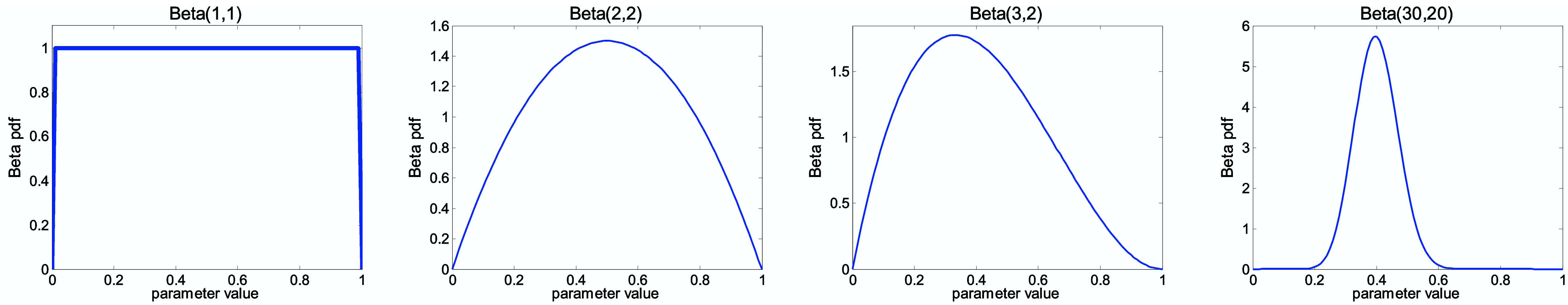
$$P(x \mid C_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \qquad P(x \mid C_2) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

- We find the model (in other words mean and variance) that maximize the likelihood (or equivalently the log likelihood). Suppose we are given training points *x_1,x_2,…,x_{n1}* from class *C_1*. Assuming that each datapoint is drawn independently from *C_1* the sample log likelihood is

$$P(x_1, x_2 \ldots, x_{n1} \mid C_1) = P(x_1 \mid C_1) P(x_2 \mid C_1) \ldots P(x_{n1} \mid C_1) = \frac{1}{\sqrt[n1]{2\pi}\sigma_1} e^{-\frac{\sum_{i=1}^{n1} (x_i-\mu_1)^2}{2\sigma_1^2}}$$

Beta prior distribution $P(\theta)$

- $P(\theta) = Beta(\beta_1, \beta_0) = \frac{1}{B(\beta_1,\beta_0)} \theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}$
- β makes the probability integrated up to one and a well formed distribution function(a constant and a kind of normalization)



Slide credit: Tom Mitchell

Maximum A Posteriori estimate

$X = 1$ $X = 0$

- Data set D of iid flips, produces α_1 ones, α_0 zeros

$$P(Data|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

- Assume prior (Conjugate prior: Closed form representation of posterior)
  - Conjugate prior: P(θ) is the conjugate prior for likelihood function P((Data|θ)) if the forms of P(θ) and P(θ|Data) are the same

$$P(\theta) = Beta(\beta_1, \beta_0) = \frac{1}{B(\beta_1, \beta_0)}\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}$$

$$\widehat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)\,\mathrm{P}(\theta) = \frac{\alpha_1 + \beta_1 - 1}{(\alpha_1 + \beta_1 - 1) + (\alpha_0 + \beta_0 - 1)}$$

Slide credit: Tom Mitchell

Some terminology

- **Likelihood function** $P(Data|\theta)$
- **Prior** $P(\theta)$
- **Posterior** $P(\theta|Data)$

- Conjugate prior:

Prior $P(\theta)$ is the conjugate prior for a **likelihood function**$P(Data|\theta)$ if the **prior** $P(\theta)$ and the **posterior** $P(\theta|Data)$ have the same form.

- Example (coin flip problem)
  - **Prior** $P(\theta)$: $Beta(\beta_1, \beta_0)$ **Likelihood** $P(Data|\theta)$: Binomial $\theta^{\alpha_1}(1-\theta)^{\alpha_0}$
  - **Posterior** $P(\theta|Data)$: $Beta(\alpha_1 + \beta_1, \alpha_0 + \beta_0)$

Slide credit: Tom Mitchell

How many parameters?

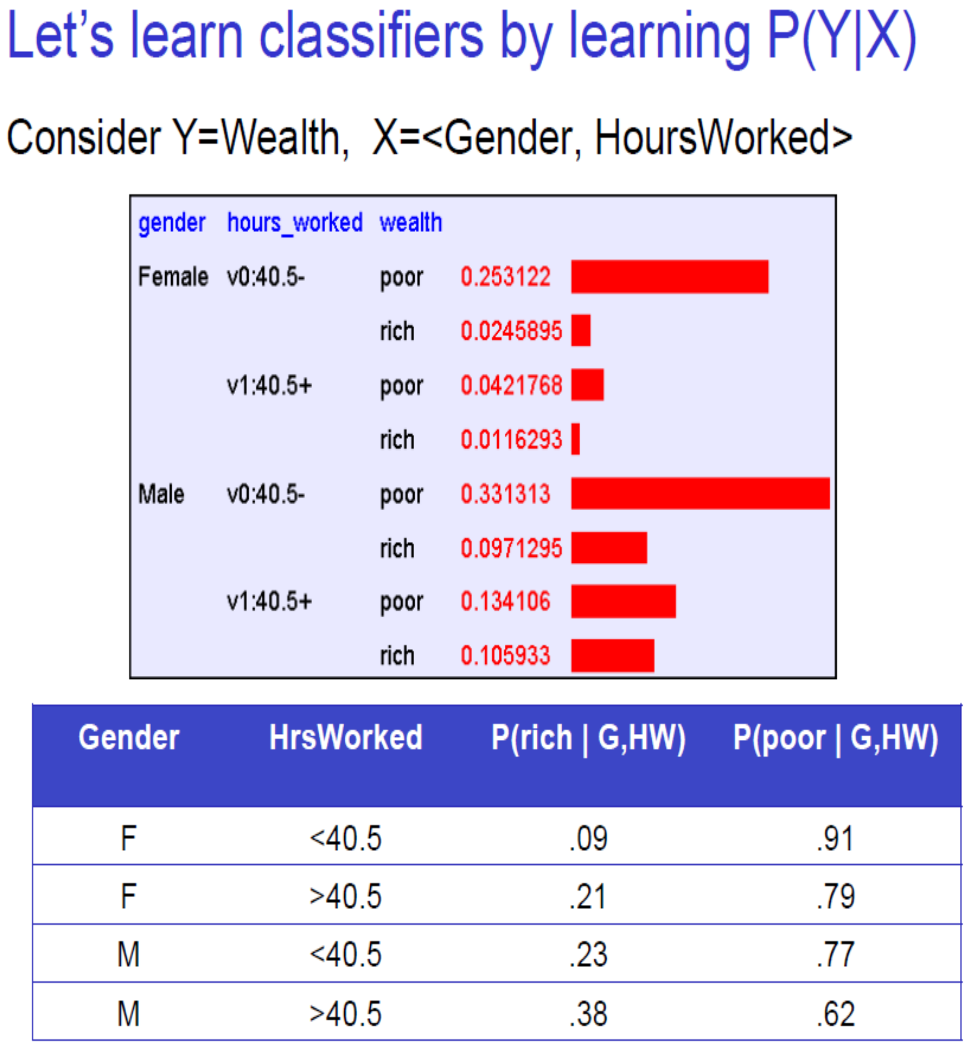
- Suppose $X = [X_1, \cdots, X_n]$, where

X_i and Y are Boolean random variables

To estimate $P(Y|X_1, \cdots, X_n)$

When $n = 2$ (Gender, Hours-worked)?

When $n = 30$?



Let's learn classifiers by learning P(Y|X)

Consider Y=Wealth, X=<Gender, HoursWorked>

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

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

Slide credit: Tom Mitchell

Can we reduce paras using Bayes rule?

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- How many parameters for $P(X_1, \cdots, X_n|Y)$?

$$(2^n - 1) \times 2$$

- How many parameters for $P(Y)$?

$$1$$

Slide credit: Tom Mitchell

Contents

- Probability basics

- Estimating parameters from data
  - Maximum likelihood (ML)
  - Maximum a posteriori estimation (MAP)

- **Naive Bayes**

Naïve Bayes

- Assumption:

$$P(X_1, \cdots, X_n | Y) = \prod_{j=1}^{n} P(X_j | Y)$$

- i.e., X_i and X_j are <u>conditionally independent</u> given Y for $i \neq j$

Slide credit: Tom Mitchell

Conditional independence

- **Definition**: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k)\ P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z_k)$$

$$P(X | Y, Z) = P(X | Z)$$

Example:

$$P(\text{Thunder}|\text{Rain, Lightning}) = P(\text{Thunder}|\text{Lightning})$$

Slide credit: Tom Mitchell

Applying conditional independence

- Naïve Bayes assumes X_i are conditionally independent given Y

e.g., $P(X_1|X_2, Y) = P(X_1|Y)$

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \text{ (chain rule)} \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

General form: $P(X_1, \cdots, X_n|Y) = \prod_{j=1}^{n} P(X_j|Y)$

How many parameters to describe $P(X_1, \cdots, X_n|Y)$? $P(\mathrm{Y})$?

- Without conditional indep assumption? $2(2^n-1)+1$
- With conditional indep assumption? 2n+1

Slide credit: Tom Mitchell

Naïve Bayes classifier

- Bayes rule:

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) P(X_1, \cdots, X_n | Y = y_k)}{\sum_j P\left(Y = y_j\right) P\left(X_1, \cdots, X_n \middle| Y = y_j\right)}$$

- Assume conditional independence among X_i's:

$$P(Y = y_k | X_1, \cdots, X_n) = \frac{P(Y = y_k) \Pi_i P(X_i | Y = y_k)}{\sum_j P\left(Y = y_j\right) \Pi_i P(X_i | Y = y_j)}$$

- Pick the most probable Y

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Slide credit: Tom Mitchell

Naïve Bayes algorithm – discrete $\mathrm{X_i}$

- For each value $\mathrm{y_k}$

    Estimate $\pi_k = P(Y = y_k)$ (Prior Prob.)

    For each value $\mathrm{x_{ij}}$ of each attribute $\mathrm{X_i}$

        Estimate $\theta_{ijk} = P(X_i = x_{ijk}|Y = y_k)$

- Classify $\mathrm{X^{test}}$

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k)\Pi_i P\left(X_i^{\text{test}}\middle|Y = y_k\right)$$

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} \pi_k \, \Pi_i \theta_{ijk}$$

Estimating parameters: discrete Y, X_i

- Maximum likelihood estimates (MLE)

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

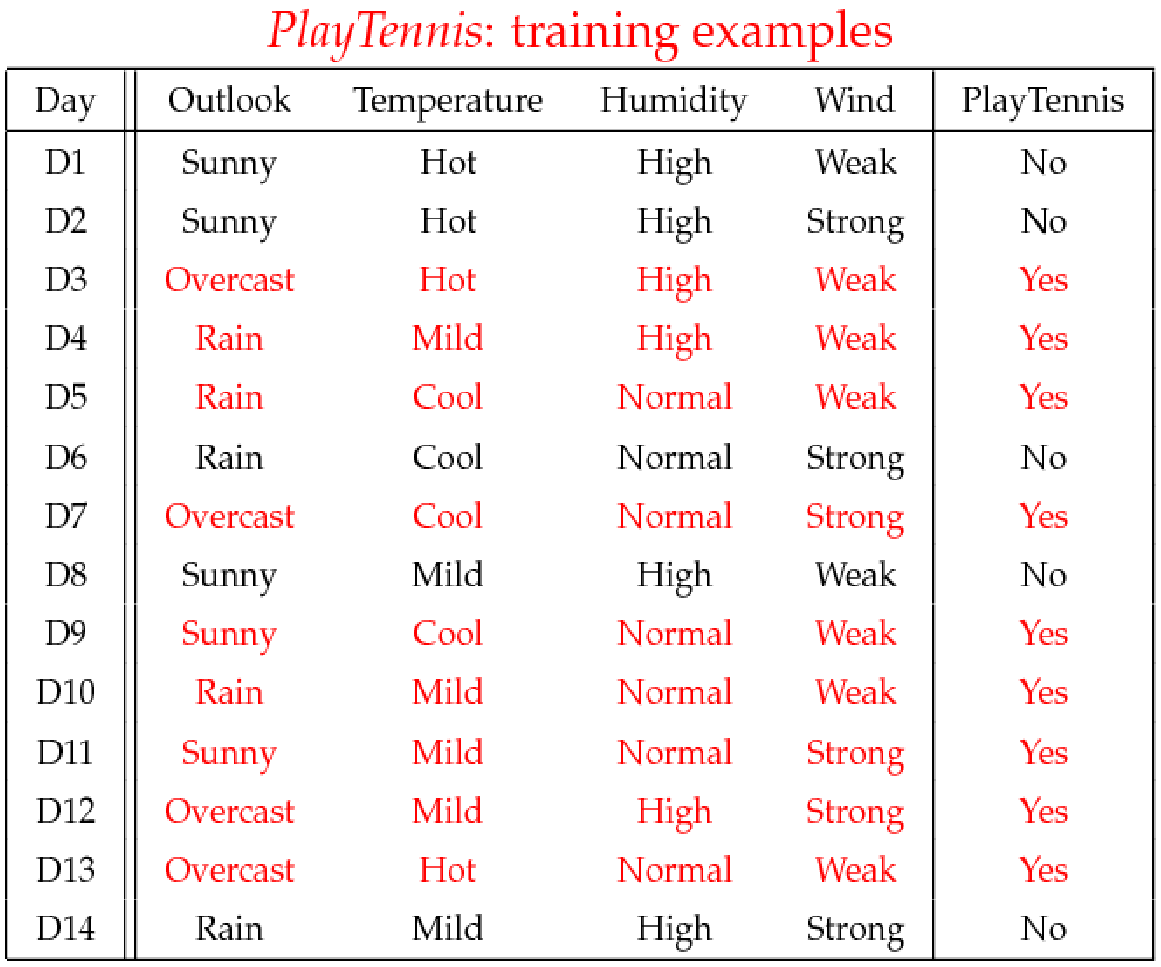
$$\hat{\theta}_{ijk} = \hat{P}\left(X_i = x_{ij} \middle| Y = y_k\right) = \frac{\#D\left\{X_i = x_{ij} \wedge Y = y_k\right\}}{\#D\{Y = y_k\}}$$

Where D = Number of items in data set D for which Y= y_k

Slide credit: Tom Mitchell

Example

- Example: Play Tennis



PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14 \quad P(\text{Play}=No) = 5/14$

Example

- Test Phase
  - Given a new instance,

    $\mathbf{x}'=(\text{Outlook}=\textit{Sunny}, \text{Temperature}=\textit{Cool}, \text{Humidity}=\textit{High}, \text{Wind}=\textit{Strong})$
  - Look up tables

$P(\text{Outlook}=\textit{Sunny}\mid\text{Play}=\textit{Yes}) = 2/9$	$P(\text{Outlook}=\textit{Sunny}\mid\text{Play}=\textit{No}) = 3/5$
$P(\text{Temperature}=\textit{Cool}\mid\text{Play}=\textit{Yes}) = 3/9$	$P(\text{Temperature}=\textit{Cool}\mid\text{Play}==\textit{No}) = 1/5$
$P(\text{Huminity}=\textit{High}\mid\text{Play}=\textit{Yes}) = 3/9$	$P(\text{Huminity}=\textit{High}\mid\text{Play}=\textit{No}) = 4/5$
$P(\text{Wind}=\textit{Strong}\mid\text{Play}=\textit{Yes}) = 3/9$	$P(\text{Wind}=\textit{Strong}\mid\text{Play}=\textit{No}) = 3/5$
$P(\text{Play}=\textit{Yes}) = 9/14$	$P(\text{Play}=\textit{No}) = 5/14$

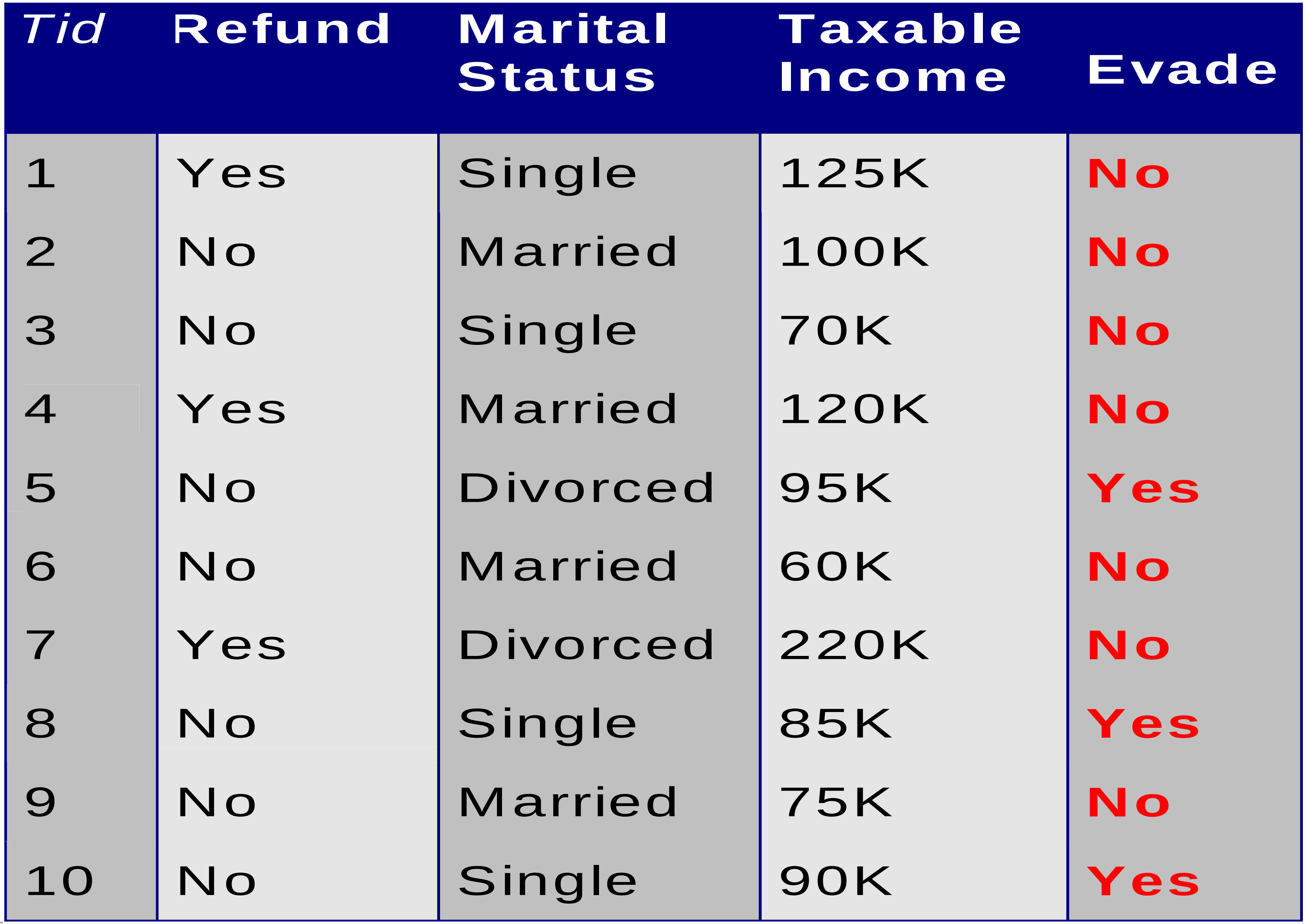
  - MAP rule

$P(\textit{Yes}\mid\mathbf{x}')$: $[P(\textit{Sunny}\mid\text{Yes})P(\textit{Cool}\mid\textit{Yes})P(\textit{High}\mid\text{Yes})P(\textit{Strong}\mid\textit{Yes})]P(\text{Play}=\textit{Yes}) =$ 0.0053

$P(\textit{No}\mid\mathbf{x}')$: $[P(\textit{Sunny}\mid\text{No})\ P(\textit{Cool}\mid\text{No})P(\textit{High}\mid\textit{No})P(\textit{Strong}\mid\textit{No})]P(\text{Play}=\textit{No}) =$ 0.0206

Given the fact $P(\textit{Yes}\mid\mathbf{x}') < P(\textit{No}\mid\mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

How to classify the new record X = (Refund=‘Yes’, Status = ‘Single’, Taxable Income =80K)



Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Example of Naïve Bayes Classifier

Given a Test Record:

X = (Refund = Yes, Status = Single, Income =80K)

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1

P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:

If class=No:	sample mean=110 sample variance=2975
If class=Yes:	sample mean=90 sample variance=25

- P(X|Class=No) = P(Refund=Yes|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 3/7 * 2/7 * 0.0062 = 0.00075

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 0 * 2/3 * 0.01 = 0

- P(No) = 0.3, P(Yes) = 0.7

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

Naïve Bayes: Subtlety #1

- Often the X_i are not really conditionally independent

- Naïve Bayes often works pretty well anyway
  - Often the right classification, even when not the right probability [Domingos & Pazzani, 1996])

- What is the effect on estimated $\mathrm{P(Y|X)}$?
  - What if we have two copies: $X_i = X_k$

$$P(Y = y_k | X_1, \cdots, X_n) \propto P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Slide credit: Tom Mitchell

Naïve Bayes: Subtlety #2

MLE estimate for $P(X_i|Y = y_k)$ might be zero.

(for example, X_i = birthdate. X_i = Feb_4_1995)

- Why worry about just one parameter out of many?

$$P(Y = y_k|X_1, \cdots, X_n) \propto P(Y = y_k)\Pi_i P(X_i|Y = y_k)$$

- What can we do to address this?
  - MAP estimates (adding "imaginary" examples)

Slide credit: Tom Mitchell

Estimating parameters: discrete Y, X_i

- **Maximum likelihood estimates (MLE)**

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}\left(X_i = x_{ij} \middle| Y = y_k\right) = \frac{\#D\left\{X_i = x_{ij}, Y = y_k\right\}}{\#D\{Y = y_k\}}$$

- **MAP estimates (Dirichlet priors):**

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}$$

$$\hat{\theta}_{ijk} = \hat{P}\left(X_i = x_{ij} \middle| Y = y_k\right) = \frac{\#D\left\{X_i = x_{ij}, Y = y_k\right\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}$$

Slide credit: Tom Mitchell

What if we have continuous $\mathrm{X_i}$

- Gaussian Naïve Bayes (GNB): assume

$$P(X_i = x \,|Y = y_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} \mathrm{exp}(-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2})$$

- Additional assumption on σ_{ik}:
  - Is independent of Y (σ_i)
  - Is independent of X_i (σ_k)
  - Is independent of X_i and Y (σ_k)

Slide credit: Tom Mitchell

Naïve Bayes algorithm – continuous $\mathrm{X_i}$

- For each value $\mathrm{y_k}$

  Estimate $\pi_k = P(Y = y_k)$

  For each attribute $\mathrm{X_i}$ estimate

  Class conditional mean μ_{ik}, variance σ_{ik}

- Classify $\mathrm{X^{test}}$

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k) \Pi_i P\left(X_i^{\text{test}} \middle| Y = y_k\right)$$

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} \pi_k \, \Pi_i \quad Normal(X_i^{\text{test}}, \mu_{ik}, \sigma_{ik})$$

Slide credit: Tom Mitchell

Things to remember

- Probability basics
- Estimating parameters from data
  - Maximum likelihood (ML) maximize $P(\mathrm{Data}|\theta)$
  - Maximum a posteriori estimation (MAP) maximize $P(\theta|\mathrm{Data})$
- Naive Bayes

$$P(Y = y_k | X_1, \cdots, X_n) \propto P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$