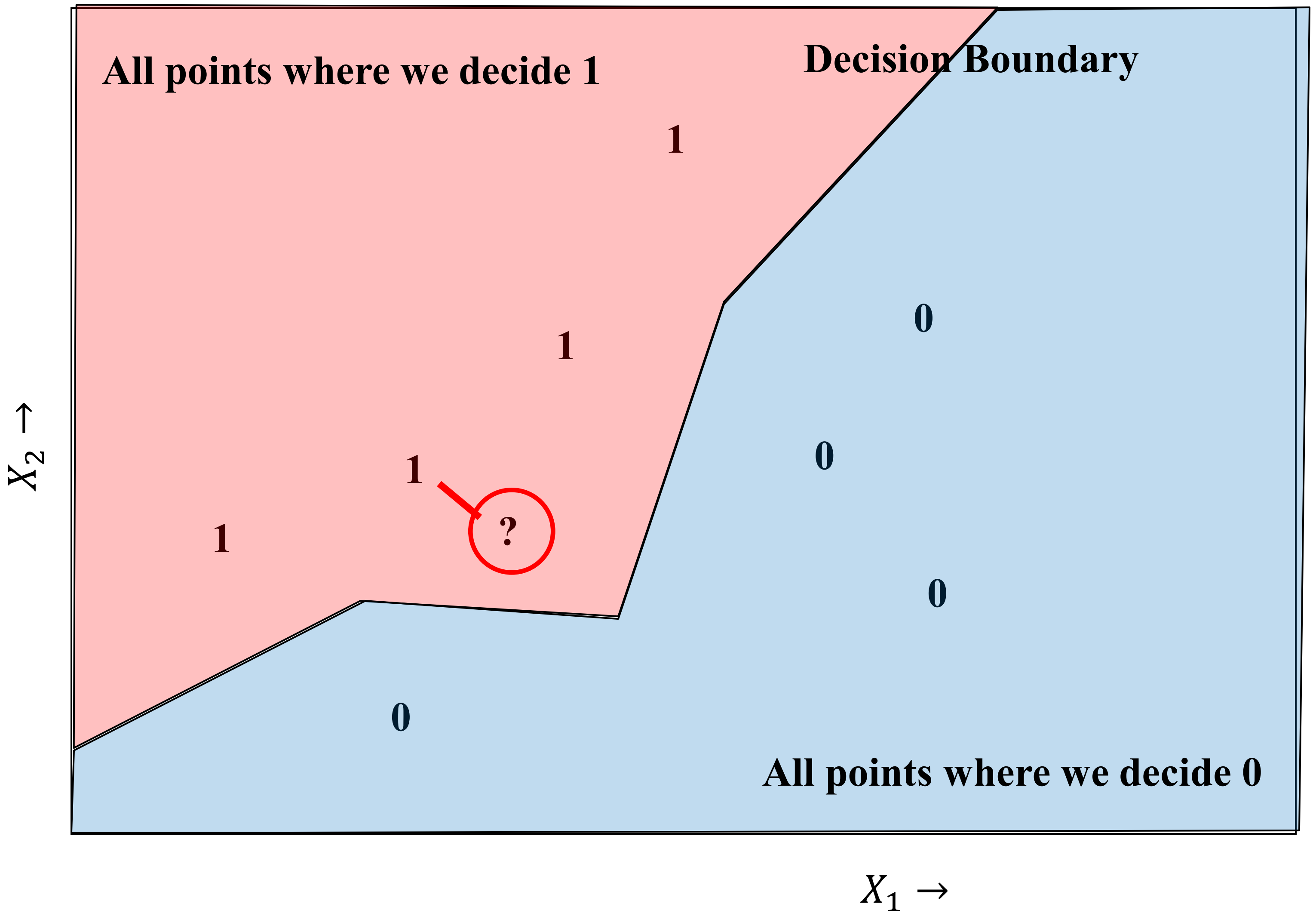
Bayes Classifiers

Adopted from slides by Alexander Ihler

Supervised Learning

- **Given** examples of a function $(X, Y = F(X))$
- **Find** function $\hat{Y} = h(X)$ to estimate $F(X)$
  - Continuous Y: Regression
  - Discrete Y: Classification

Nearest neighbor classifier



A basic classifier

- Training data $D = \{x^{(i)}, y^{(i)}\}$, Classifier $f(x)$
  - Discrete feature vector X
  - $f(x)$ is a contingency table
- Ex: credit rating prediction (bad/good)
  - X = income (low/med/high)
  - How can we make the most # of correct predictions?

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

A basic classifier

- Training data $D = \{x^{(i)}, y^{(i)}\}$, Classifier $f(x)$
  - Discrete feature vector X
  - $f(x)$ is a contingency table
- Ex: credit rating prediction (bad/good)
  - X = income (low/med/high)
  - How can we make the most # of correct predictions?

  - Predict more likely outcome
    for each possible observation

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

A basic classifier

- Training data $D = \{x^{(i)}, y^{(i)}\}$, Classifier $f(x)$
  - Discrete feature vector X
  - $f(x)$ is a contingency table
- Ex: credit rating prediction (bad/good)
  - X = income (low/med/high)
  - How can we make the most # of correct predictions?

  - Predict more likely outcome
    for each possible observation

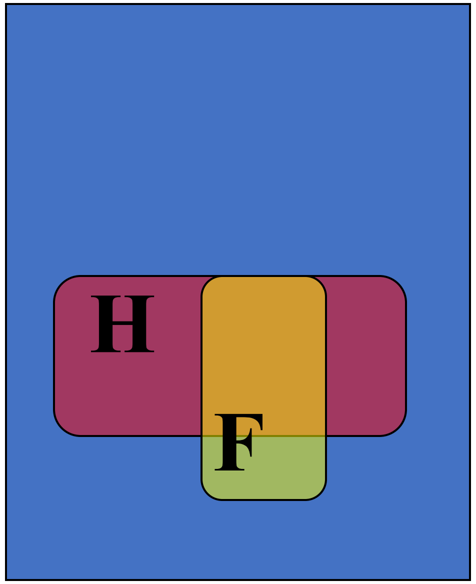
  - Can normalize into probability:
    $p(\, y = good \mid X = x \,)$

  - How to generalize?

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

Bayes rule

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



- Two events: headache, flu
- p(H) = 1/10
- p(F) = 1/40
- p(H|F) = 1/2

- You wake up with a headache – what is the chance that you have the flu?

Example from Andrew Moore's slides

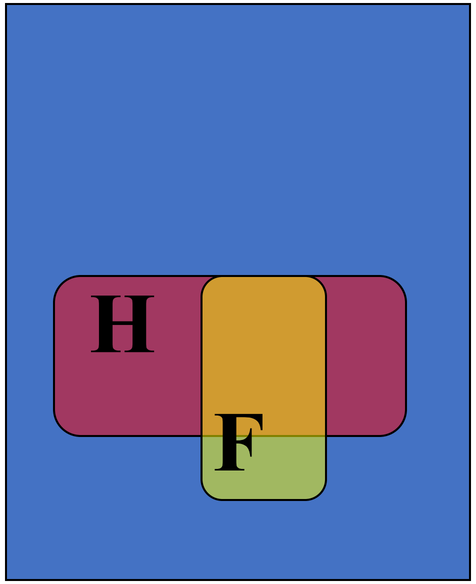
Bayes rule

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

- Two events: headache, flu
- p(H) = 1/10
- p(F) = 1/40
- p(H|F) = 1/2

- P(H & F) = ?

- P(F|H) = ?



Example from Andrew Moore's slides

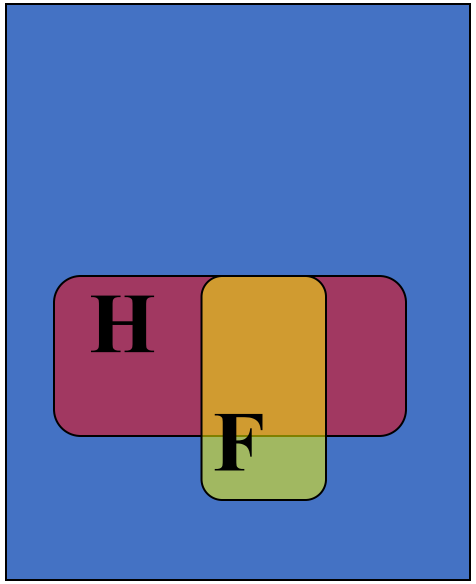
Bayes rule

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

- Two events: headache, flu
- p(H) = 1/10
- p(F) = 1/40
- p(H|F) = 1/2

- P(H & F) = p(F) p(H|F)

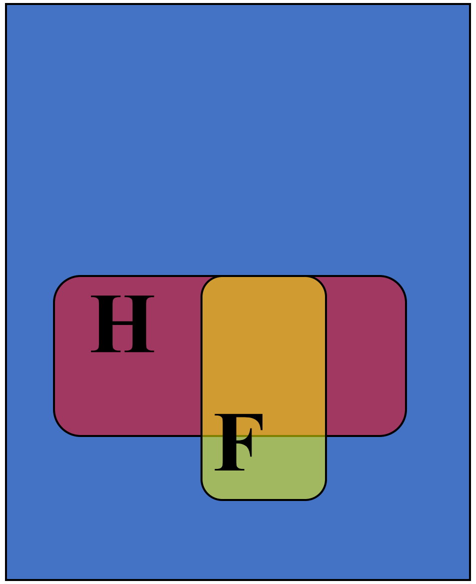
  = (1/2) * (1/40) = 1/80
- P(F|H) = ?



Example from Andrew Moore's slides

Bayes rule

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



- Two events: headache, flu
- p(H) = 1/10
- p(F) = 1/40
- p(H|F) = 1/2

- P(H & F) = p(F) p(H|F)
  = (1/2) * (1/40) = 1/80
- P(F|H) = p(H & F) / p(H)
  = (1/80) / (1/10) = 1/8

Example from Andrew Moore's slides

Classification and probability

- Suppose we want to model the data

- Prior probability of each class, $p(y)$
  - E.g., fraction of applicants that have good credit
- Distribution of features given the class, $p(x \mid y = c)$
  - How likely are we to see "x" in users with good credit?

- Joint distribution $$p(y|x)p(x) = p(x,y) = p(x|y)p(y)$$

- Bayes Rule: $$\Rightarrow \quad p(y|x) = p(x|y)p(y)/p(x)$$

$$= \frac{p(x|y)p(y)}{\sum_c p(x|y=c)p(y=c)}$$

(Use the rule of total probability to calculate the denominator!)

Bayes classifiers

- Training data
  - Estimate $p(y = c)$
  - Split by class
  - $D_c = \{ x^{(i)} : y^{(i)} = c \}$
- Estimate $p(x \mid y = c)$ using D_c
- Estimate $p(y \mid x)$ using Bayes rule
- For a discrete X, this recalculates the same table…

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

$\Rightarrow$

p(x | y=0)	p(x | y=1)
42 / 383	15 / 307
338 / 383	287 / 307
3 / 383	5 / 307

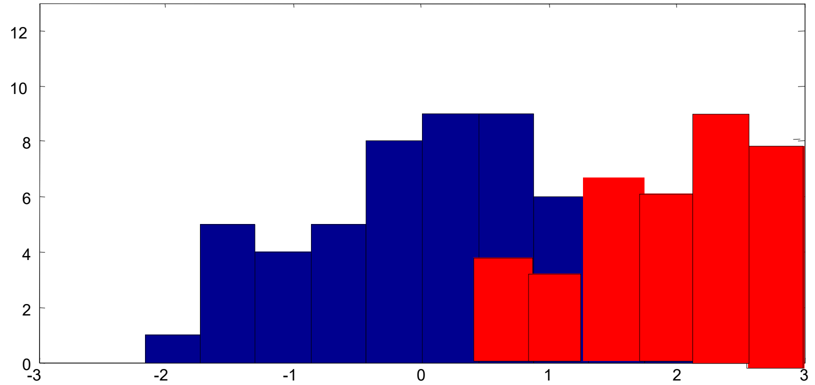
$\Rightarrow$

p(y=0|x)	p(y=1|x)
.7368	.2632
.5408	.4592
.3750	.6250

p(y)	383/690	307/690

Bayes classifiers

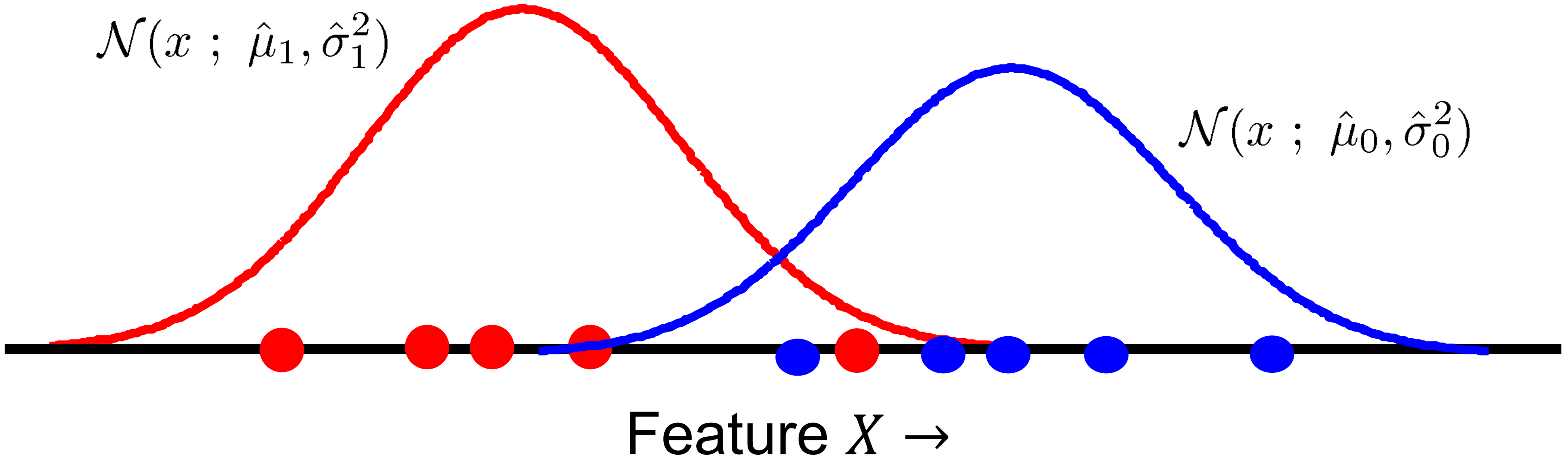
- Training data
  - Estimate $p(y = c)$
  - Split by class
  - $D_c = \{ x^{(i)} : y^{(i)} = c \}$
- Estimate $p(x \mid y = c)$ using D_c
- Estimate $p(y \mid x)$ using Bayes rule
- For continuous X, can use any density estimate like
  - Histogram
  - Gaussian
  - …



Gaussian models

- Estimate parameters of the Gaussians from the data

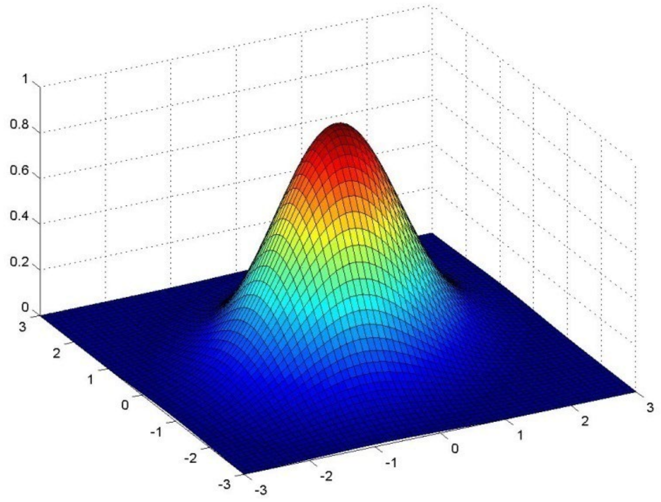
$$\alpha = \frac{m_1}{m} = \hat{p}(y = c_1) \qquad \hat{\mu} = \frac{1}{m}\sum_j x^{(j)} \qquad \hat{\sigma}^2 = \frac{1}{m}\sum_j (x^{(j)} - \mu)^2$$

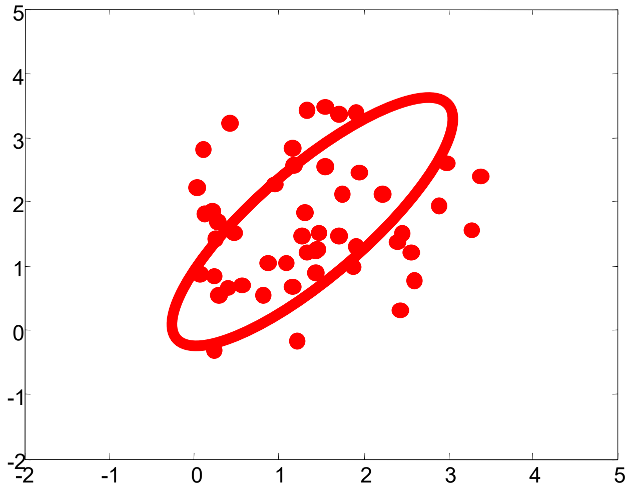


Multivariate Gaussian models

- Similar to univariate case

$$\mathcal{N}(\underline{x}\ ;\ \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right\}$$



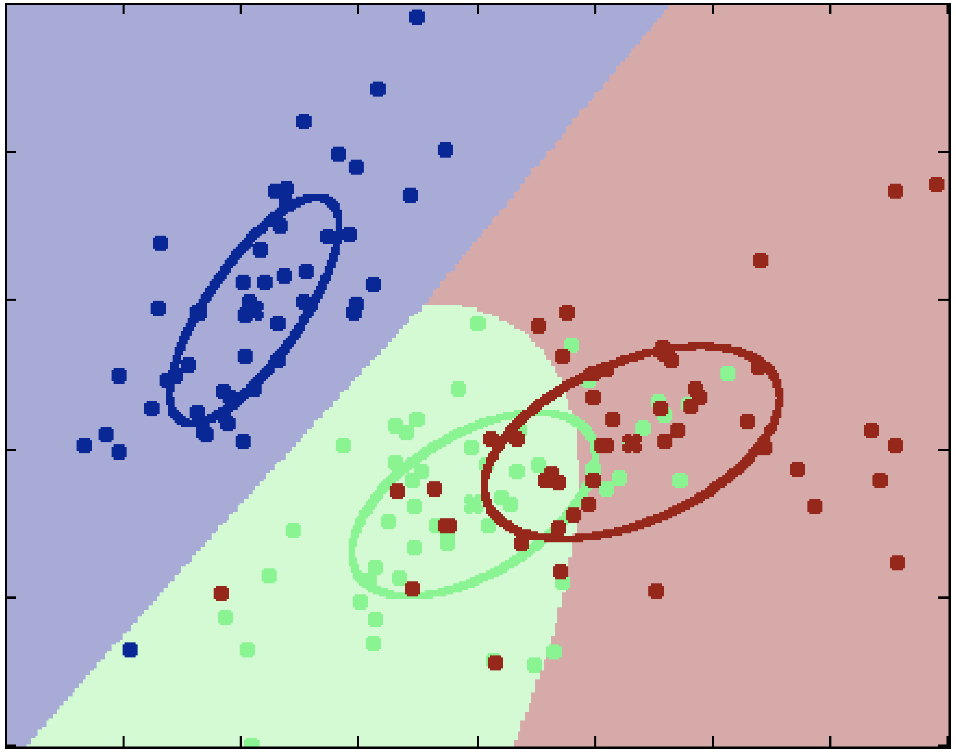


Example: Gaussian Bayes for Iris Data

- Fit Gaussian distribution to each class {0,1,2}

$$p(y) = \text{Discrete}(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3})$$

$$p(x_1, x_2 | y = 0) = \mathcal{N}(x\,;\, \mu_0, \Sigma_0)$$
$$p(x_1, x_2 | y = 1) = \mathcal{N}(x\,;\, \mu_1, \Sigma_1)$$
$$p(x_1, x_2 | y = 2) = \mathcal{N}(x\,;\, \mu_2, \Sigma_2)$$



Bayes Classifiers: Naïve Bayes

Bayes classifiers

- Estimate $p(y) = [\, p(y=0)\, , p(y=1)\ \ldots]$
- Estimate $p(x \mid y = c)$ for each class c
- Calculate $p(y = c \mid x)$ using Bayes rule
- Choose the most likely class c

- For a discrete X, can represent as a contingency table…
  - What about if we have more discrete features?

Features	# bad	# good
X=0	42	15
X=1	338	287
X=2	3	5

$\Rightarrow$

p(x | y=0)	p(x | y=1)
42 / 383	15 / 307
338 / 383	287 / 307
3 / 383	5 / 307

$\Rightarrow$

p(y=0|x)	p(y=1|x)
.7368	.2632
.5408	.4592
.3750	.6250

p(y)	383/690	307/690

Joint distributions

- Make a truth table of all combinations of values

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Joint distributions

- Make a truth table of all combinations of values
- For each combination of values, determine how probable it is
- Total probability must sum to one
- How many values did we specify?

A	B	C	p(A,B,C | y=1)
0	0	0	0.4
0	0	1	0.1
0	1	0	0.0
0	1	1	0.0
1	0	0	0.1
1	0	1	0.2
1	1	0	0.1
1	1	1	0.1

Overfitting & density estimation

- Estimate probabilities from the data
  - E.g., how many times (what fraction) did each outcome occur?

- M data $<<$ 2^N parameters?

- What about the zeros?
  - We learn that certain combinations are impossible?
  - What if we see these later in test data?

- Overfitting!

A	B	C	p(A,B,C | y=1)
0	0	0	4/10
0	0	1	1/10
0	1	0	0/10
0	1	1	0/10
1	0	0	1/10
1	0	1	2/10
1	1	0	1/10
1	1	1	1/10

Overfitting & density estimation

- Reduce the model complexity
  - E.g., assume that features are conditionally independent of one another given the class label
- Conditional Independence:
- $p(a,b,c|y) = p(a|y)\, p(b|y)\, p(c|y)$
- $p(x_1, x_2, \dots x_N \mid y=1) = p(x_1 \mid y=1)\, p(x_2 \mid y=1) \dots p(x_N \mid y=1)$
- Only need to estimate each individually

A	p(A |y=1)
0	.4
1	.6

B	p(B |y=1)
0	.7
1	.3

C	p(C |y=1)
0	.1
1	.9

$\Rightarrow$

A	B	C	p(A,B,C | y=1)
0	0	0	.4 * .7 * .1
0	0	1	.4 * .7 * .9
0	1	0	.4 * .3 * .1
0	1	1	…
1	0	0	
1	0	1	
1	1	0	
1	1	1	

Naïve Bayes Models

- Naïve Bayes:
  - $p(y \mid \boldsymbol{x}) = p(\boldsymbol{x} \mid y)\, p(y) \,/\, p(\boldsymbol{x})$
  - Estimate $p(y)$ for each class y
  - $p(\boldsymbol{x} \mid y) = \prod_i p(x_i \mid y)$
  - Estimate $p(x_i \mid y)$ for each feature x_i and class y

$$\boxed{\text{Predict } y = c_1 \text{ if } p(\boldsymbol{x} \mid y = c_1)\, p(y = c_1) > p(\boldsymbol{x} \mid y = c_2)\, p(y = c_2)}$$

- Note: may not be a good model of the data
  - Doesn’t capture correlations in x’s
  - Can’t capture some dependencies
- But in practice it often does quite well!

Example: Naïve Bayes

Observed Data:

x_1	x_2	y
1	1	0
1	0	0
1	0	1
0	0	0
0	1	1
1	1	0
0	0	1
1	0	1

$$\hat{p}(y=1) = \frac{4}{8} \qquad = (1 - \hat{p}(y=0))$$

$$\hat{p}(x_1 = 1 | y = 0) = \frac{3}{4} \qquad \hat{p}(x_1 = 1 | y = 1) = \frac{2}{4}$$

$$\hat{p}(x_2 = 1 | y = 0) = \frac{2}{4} \qquad \hat{p}(x_2 = 1 | y = 1) = \frac{1}{4}$$

$$\hat{p}(x_1 = 1, x_2 = 1 | y = 1) = \hat{p}(x_1 = 1 | y = 1)\, \hat{p}(x_2 = 1 | y = 1)$$
$$= \frac{2}{4} \times \frac{1}{4}$$

Example: Naïve Bayes

Observed Data:

x_1	x_2	y
1	1	0
1	0	0
1	0	1
0	0	0
0	1	1
1	1	0
0	0	1
1	0	1

$$\hat{p}(y=1) = \frac{4}{8} \qquad = (1 - \hat{p}(y=0))$$

$$\hat{p}(x_1 = 1 | y = 0) = \frac{3}{4} \qquad \hat{p}(x_1 = 1 | y = 1) = \frac{2}{4}$$

$$\hat{p}(x_2 = 1 | y = 0) = \frac{2}{4} \qquad \hat{p}(x_2 = 1 | y = 1) = \frac{1}{4}$$

Prediction given some observation x?

$$\underset{\frac{4}{8} \times \frac{2}{4} \times \frac{1}{4}}{\hat{p}(y=1)\hat{p}(x=11|y=1)} \quad \begin{matrix} < \\ > \end{matrix} \quad \underset{\frac{4}{8} \times \frac{3}{4} \times \frac{2}{4}}{\hat{p}(y=0)\hat{p}(x=11|y=0)}$$

Decide class 0

Naïve Bayes Models for Spam

- $y \in \{spam,\ not\ spam\}$
- X = observed words in email
  - Ex: [“the” … “probabilistic” … “lottery”…]
  - “1” if word appears; “0” if not

- 1000’s of possible words: 2^{1000s} parameters?
- # of atoms in the universE: » 2^{270}…

- Model words ***given*** email type as independent
- Some words more likely for spam (“lottery”)
- Some more likely for real (“probabilistic”)
- Only 1000’s of parameters now…

Summary

- Bayes rule; $p(\, y \mid x \,)$
- Bayes classifiers
  - Learn $p(x \mid y = C)$, $p(y = C) \Longrightarrow p(y = C \mid x)$
- Naïve Bayes classifiers
  - Assume features are independent given class:

    $$p(\, \boldsymbol{x} \mid y = C \,) \quad = \quad p(\, x_1 \mid y = C \,)\, p(\, x_2 \mid y = C \,) \ldots$$

- Maximum likelihood (empirical) estimators for
  - Discrete variables
  - Gaussian variables
  - Overfitting