# Bayes and Naïve Bayes Classifiers

CS434

# In this lecture

1. Review some basic probability concepts
2. Introduce a useful probabilistic rule - Bayes rule
3. Introduce the learning algorithm based on Bayes rule (thus the name Bayes classifier) and its extension, Naïve Beyes

# Commonly used discrete distributions

Binary random variable $x \sim Bernoulli(p)$

$$\left.\begin{aligned} P(x=1) &= p; \\ P(x=0) &= 1-p \end{aligned}\right\} \quad p(x) = p^x(1-p)^{(1-x)}$$

Categorical distribution: $x$ can take multiple values, $v_1, \ldots, v_k$

$$\left.\begin{aligned} P(x=v_1) &= p_1 \\ P(x=v_2) &= p_2 \\ P(x=v_k) &= p_k \end{aligned}\right\} \quad p(x) = \prod_{i=1}^{k} p_i^{I(x=v_i)}$$

$p_1 + p_2 + \cdots + p_k = 1$

# Learning the parameters of discrete distributions

- Let $x$ denote the event of getting a head when tossing a coin ($x \in \{0,1\}$)
- Given a sequence of $n$ coin tosses $x_1, \ldots, x_n$, we estimate

$$p(x=1) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Let $x$ denote the outcome of rolling a die ($x \in \{1, \ldots, 6\}$)
- Given a sequence of $n$ rolls, we estimate

$$p(x=j) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = j)$$

Estimation using such simple counting performs what we call "Maximum Likelihood Estimation" (MLE). There are other methods as well.

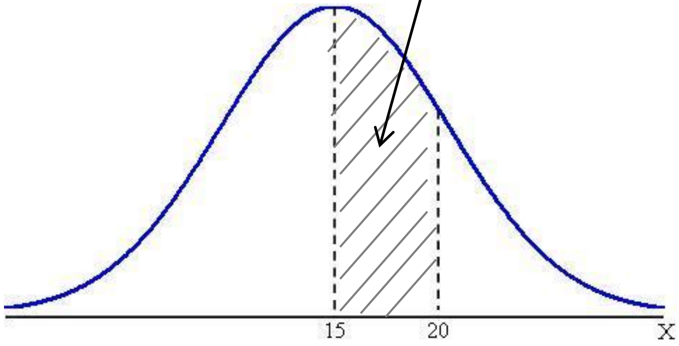
# Continuous Random Variables

- A continuous random variable $x$ can take any value in an interval on the real line
  - $x$ usually corresponds to some real-valued measurements, e.g., $x =$ *today's lowest temperature*
  - The probability of a continuous random variable taking an exact value is typically zero: P($x$=56.2)=0
  - It is more meaningful to measure the probability of a random variable taking a value within an interval $P(x \in [50, 60])$
  - This is captured in the ***Probability density function***

# PDF: probability density function

- We often use $f(x)$ to denote the PDF of $x$
- $f(x) \geq 0$
- $f(x)$ can be larger than 1
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $\int_{x_1}^{x_2} f(x)dx = P(x_1 < x < x_2)$

$P(15 < x < 20)$



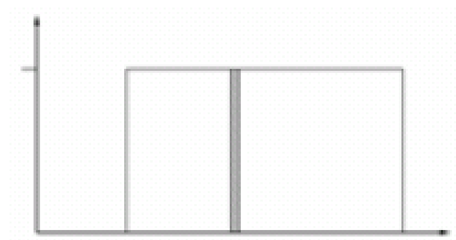

- If $f(x_1) = \alpha f(x_2)$:

  When x is sampled from $f(x)$, you are $\alpha$ times as likely to see a $x$ value “near” $x_1$ than that a $x$ value “near” $x_2$

# Commonly Used Continuous Distributions

## Uniform Probability Density Function

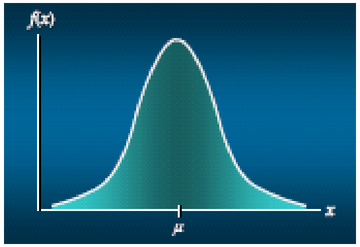
$$f(x) = 1/(b-a) \quad \text{for } a \le x \le b$$
$$= 0 \quad \text{elsewhere}$$



E.g., Suppose we know that in the last 10 minutes, one customer arrived at OSU federal counter #1. The actual time of arrival X can be modeled by a uniform distribution over the interval of (0, 10)

## Normal (Gaussian) Probability Density Function

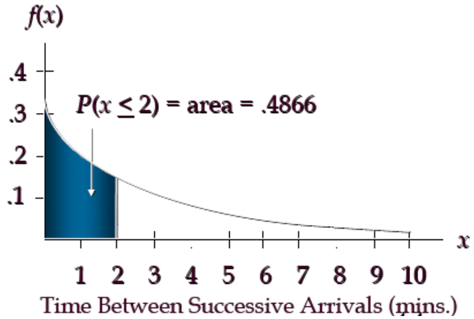
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$




E.g., the body temp. of a person, the average IQ of a class of 1st graders

## Exponential Probability Distribution

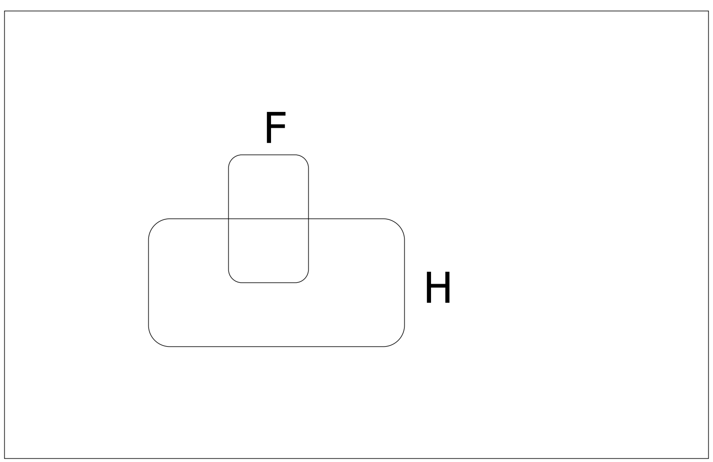
$$f(x) = \frac{1}{\mu} e^{-x/\mu}$$




E.g., the time between two successive forest fires

# Conditional Probability

- $P(A|B)$ = probability of A being true given that B is true



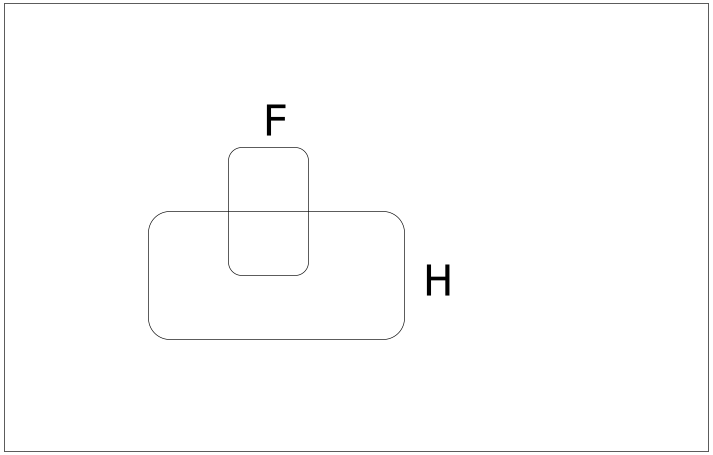

H = "Have a headache"
F = "Coming down with Flu"

$P(H) = 1/10$
$P(F) = 1/40$
$P(H|F) = 1/2$

"Headaches are rare and flu is rarer, but if you're coming down with a flu there's a 50-50 chance you'll have a headache."

Copyright © Andrew W. Moore

# Conditional Probability




H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

P(H|F)

= Area of "H and F" region
  ------------------------------
  Area of "F" region

= P(H ^ F)
  -----------
  P(F)

Copyright © Andrew W. Moore

# Definition of Conditional Probability

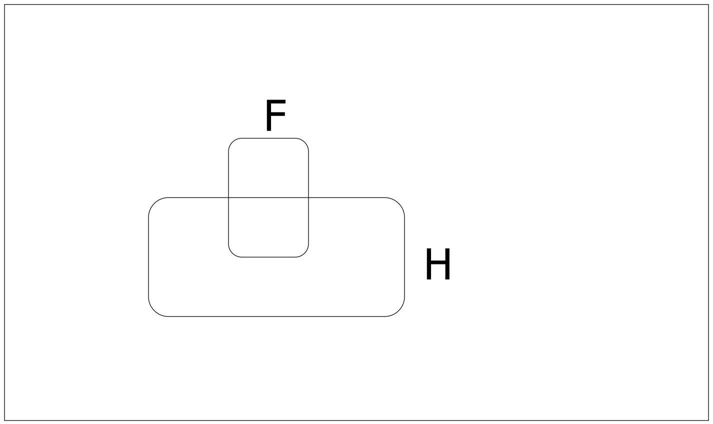
$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

## Corollary: The Chain Rule

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

Copyright © Andrew W. Moore

# Probabilistic Inference




H = "Have a headache"
F = "Coming down with Flu"

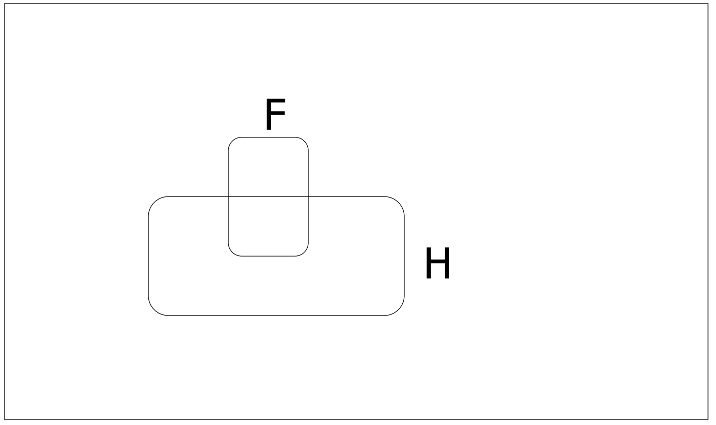
P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu"

Is this reasoning good?

Copyright © Andrew W. Moore

# Probabilistic Inference




H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

$P(F \wedge H) = \ldots$

$P(F|H) = \ldots$

Copyright © Andrew W. Moore

# What we just did…

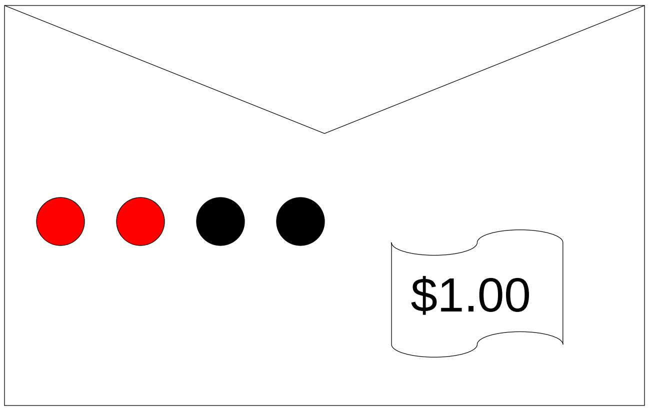
$$P(B|A) = \frac{P(A \wedge B)}{P(A)} = \frac{P(A|B)\,P(B)}{P(A)}$$

This is Bayes Rule

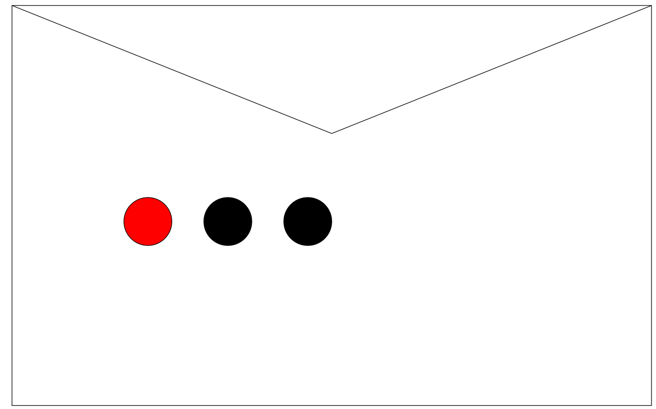
**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Copyright © Andrew W. Moore

# Using Bayes Rule to Gamble




The "Win" envelope has a dollar and four beads in it

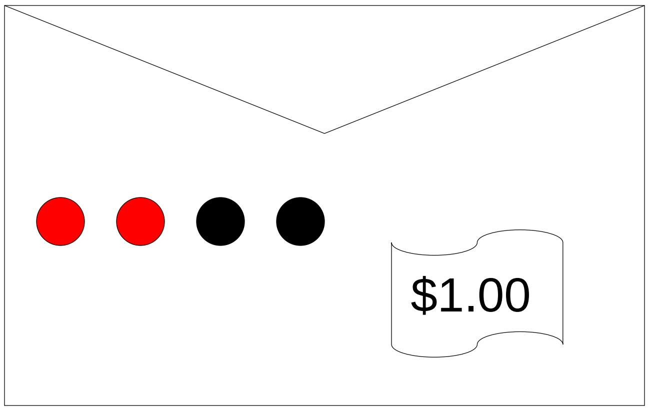


The "Lose" envelope has three beads and no money

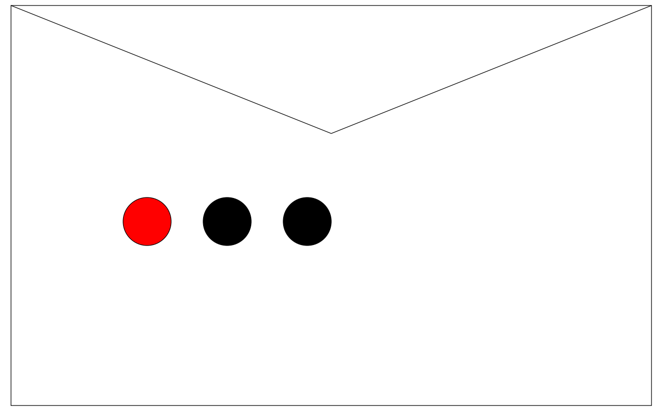
Trivial question: someone draws an envelope at random and offers to sell it to you. How much should you pay in order to not lose money on average?

Copyright © Andrew W. Moore

# Using Bayes Rule to Gamble




The “Win” envelope has a dollar and four beads in it



The “Lose” envelope has three beads and no money

Interesting question: before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it’s black: How much should you pay?

Suppose it’s red: How much should you pay?

Copyright © Andrew W. Moore

# Where are we?

- We have recalled the fundamentals of probability
- We have discussed conditional probability and Bayes rule
- Now we will move on to talk about the Bayes classifier

# Basic Idea

- Each example is described by $m$ input features, i.e., $\boldsymbol{x} = [x_1, x_2, \dots, x_m]^T$, for now we assume they are discrete variables
- Each example belongs to one of $c$ possible classes, denoted by $y \in \{1, \dots, c\}$
- If we don't know anything about the features, a randomly drawn example has a fixed probability $P(y = j)$ to belong to class $j$
- If an example belongs to class $j$, its features $\boldsymbol{x} = [u_1, \dots, u_m]^T$ will follow some particular distribution $p(u_1, \dots, u_m | y = j)$
- Given an example $\boldsymbol{x} = [u_1, \dots, u_m]^T$, a bayes classifier reasons about the value of $y$ using Bayes rule:

$$P(y = j | u_1, \dots, u_m) = \frac{P(u_1, \dots, u_m | y = j) P(y = j)}{P(u_1, \dots, u_m)}$$

# Learning a Bayes Classifier

$$P(y = j | u_1, \dots, u_m) = \frac{P(u_1, \dots, u_m | y = j) P(y = j)}{\boldsymbol{P(u_1, \dots, u_m)}}$$

Given a set of training data $S = \{(\boldsymbol{x_i}, y_i): i = 1, \dots, n\}$, we learn:

- Class Priors: $P(y = j)$ for $j = 1, \dots c$

$$P(y = j) = \frac{1}{n} \sum_{i=1}^{n} I(y_i = j)$$

- Class Conditional Distribution: $P(x = [u_1, \dots, u_m]^T | y = j)$ for $= 1, \dots, c$

$$P(u_1, \dots, u_m \mid y = j)$$

$$= \frac{\#\text{of class } j \text{ examples that have values } (u_1, u_2, \dots, u_m)}{\#\text{of total class } j \text{ examples}}$$

- Marginal Distribution of $\boldsymbol{x}$: $P(u_1, \dots, u_m)$

No need to learn, can be computed using

$$P(\boldsymbol{x}) = \sum_{j=1}^{c} P(\boldsymbol{x} \mid y = j) P(y = j)$$

# Learning a joint distribution

Build a JD table for your attributes in which the probabilities are unspecified

| A | B | C | Prob |
|---|---|---|---|
| 0 | 0 | 0 | ? |
| 0 | 0 | 1 | ? |
| 0 | 1 | 0 | ? |
| 0 | 1 | 1 | ? |
| 1 | 0 | 0 | ? |
| 1 | 0 | 1 | ? |
| 1 | 1 | 0 | ? |
| 1 | 1 | 1 | ? |

The fill in each row with

$$\hat{P}(\text{row}) = \frac{\text{records matching row}}{\text{total number of records}}$$

| 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 |

Fraction of all records in which A and B are True but C is False

# Example of Learning a Joint

- This Joint was obtained by learning from three attributes in the UCI “Adult” Census Database [Kohavi 1995]

*48842 examples in total*

*12363 examples with this value combination*

| 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 |

UCI machine learning repository: http://archive.ics.uci.edu/ml/datasets/Adult

# Bayes Classifiers in a nutshell

1. Estimate $P(y = j)$ as fraction of class j examples for j=1,…,c.
2. Learn conditional joint distribution $p(x_1, x_2, \dots x_m \mid y = j)$ for $j = 1, \dots, c$

*learning*

3. To make a prediction for a new example $\boldsymbol{x} = [u_1, \dots, u_m]^T$

$$y^{\text{predict}} = \underset{j}{\operatorname{argmax}} P(y = j \mid u_1, \cdots, u_m)$$

$$= \underset{j}{\operatorname{argmax}} \frac{P(u_1, \cdots, u_m \mid y = j) P(y = j)}{P(u_1, \cdots, u_m)}$$

$$= \underset{j}{\operatorname{argmax}} P(u_1, \cdots, u_m \mid y = j) P(y = j)$$

If you wish to have the probability value $P(y = 1 | [u_1, \dots, u_m]^T)$, you will need to compute the normalizing factor $P([u_1, \dots, u_m]^T)$.

# Example: Spam Filtering

- Use the Bag-of-words representation
- Assume a dictionary of $m$ words and tokens
- Create one binary attribute for each dictionary entry
  - i.e., $x_i = 1$ means the $i$-th word is used in the email
- Consider a reasonable dictionary size: $m = 5000$ --- we have 5000 binary attributes
- How many parameters that we need to learn?
  - We need to learn the class prior $P(y$=1),$P(y$=0): 1
  - For each of the two classes, we need to learn a joint distribution table of 5000 binary variables: $2^{5000} - 1$
  - Total: $2 * (2^{5000} - 1) + 1$
- Clearly we don't have nearly enough data to estimate that many parameters

# Joint Distribution Overfits

- It is common to encounter the following situation:
  - No training examples have the exact value combinations $\boldsymbol{x} = (u_1, u_2, \ldots . u_m)$,
  - $P(\boldsymbol{x}|y = j\ ) \ = 0$ for all values of $y$
  - How do we make predictions for such examples?
- To avoid overfitting
  - Make some bold assumptions to simplify the joint distribution

# The Naïve Bayes Assumption

- Assume that each feature is independent of any other features given the class label

$$\boxed{\begin{aligned}&P(x_1 = u_1, \cdots, x_m = u_m \mid y = j)\\ &= P(x_1 = u_1 \mid j) \cdots P(x_m = u_m \mid j)\end{aligned}}$$

# A note about independence

- Assume A and B are two random events. Then

  “A and B are independent”

  if and only if

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

# Independence Theorems

- Assume $P(A|B) = P(A)$
- Then

$$P(A \wedge B) = P(A)\, P(B)$$

- Assume $P(A|B) = P(A)$
- Then

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

# Examples of independent events

- Two separate coin tosses
- Consider the following four events:
  - T: Toothache ( I have a toothache)
  - C: Catch (dentist's steel probe catches in my tooth)
  - A: Cavity (I have a cavity)
  - W: Weather (weather is good)
  - $P(T, C, A, W) = P(T,C,A)P(W)$

# Conditional Independence

- $P(A|B,C) = P(A|C)$
  - A and B are conditionally independent given C
  - $P(B|A,C)=P(B|C)$
- If A and B are conditionally independent given C, then we have
  - $P(A,B|C) = P(A|C)\,P(B|C)$

# Example of conditional independence

- T: Toothache ( I have a toothache)
- C: Catch (dentist's steel probe catches in my tooth)
- A: Cavity

T and C are conditionally independent given A: $P(T, C|A) = P(T|A)*P(C|A)$

So , **events that are not independent from each other might be conditionally independent given some fact**

It can also happen the other way around. **Events that are independent might become conditionally dependent given some fact.**

B=Burglar in your house; A = Alarm **(**Burglar**)** rang in your house
E = Earthquake happened
B is independent of E (ignoring some minor possible connections between them)
However, if we know A is true, then B and E are no longer independent. Why?
$P(B|A) >> P(B|A, E)$ Knowing E is true makes it much less likely for B to be true

# Naïve Bayes Classifier

- By assuming that each attribute is independent of any other attributes given the class label, we now have a *Naïve* Bayes Classifier
- Instead of learning a joint distribution of all features, we learn $p(x_i|y=j)$ separately for each feature $x_i$
- And we compute the joint by taking the product:

$$P(\boldsymbol{x} = [u_1, \ldots, u_m]^T | y = j) = \prod_{i=1}^{m} P(x_i = u_i | y = j)$$

# Naïve Bayes Classifiers in a nutshell

1. Learn the $p(x_i | y = j)$ for each feature $x_i$, and y value $j$
2. Estimate $P(y = j)$ as fraction of records with $y = j$ .

*learning*

3. For a new example $\boldsymbol{x} = [u_1, \dots, u_m]^T$:

$$y^{\text{predict}} = \underset{j}{\operatorname{argmax}} P(y = j \mid x_1 = u_1, \cdots, x_m = u_m)$$

$$= \underset{j}{\operatorname{argmax}} P(x_1 = u_1, \cdots, x_m = u_m \mid y = j) P(y = j)$$

$$= \underset{j}{\operatorname{argmax}} P(y = j) \prod_{i=1}^{m} P(x_i = u_i \mid y = j)$$

# Example

| $X_1$ | $X_2$ | $X_3$ | Y |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 1 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 1 |
| 0 | 0 | 1 | 1 |
| 0 | 1 | 1 | 1 |

Apply Naïve Bayes, and make prediction for (1,0,1)?

# When Bag-of-words uses counts (Multinomial Naïve Bayes)

- Often text is represented by the number of times each word appeared in it
  - E.g., "There are some shinning apples on the apple tree"
  - Word "apple" will be counted as appearing twice (ignoring the difference between plural and singular form)
  - How do we use naïve bayes in this case?
- We still learn $p(w_i|y=1)$ for each word $i$
- It will be estimated as

$$\frac{\text{\# of times word } i \text{ appeared in spam emails}}{\text{total \# words in spam emails}}$$

- For the sentence S above, we will have:

$$P(S|y=1)$$
$$= P(\text{"there"}|y=1)P(\text{"are"}|y=1)\ldots P(\text{"}apple\text{"}|y=1)^2\ldots P(\text{"}tree\text{"}|y=1)$$

Similarly for $P(S|y=0)$

- Bayes rule can then be used to compute $P(y=1|S)$

# Laplace Smoothing

- With Naïve Bayes Assumption, we still have zero probabilities
- E.g., if we receive an email that contains a word $w$ that has never appeared in the training emails
  - $P(w|spam) = 0$ and $P(w|nonspam) = 0$
- As such we ignore all the other words in the email because of this single rare word
- Laplace smoothing can help

Binary:

$$P(w|spam) = \frac{\text{\# of spam emails with word } w + 1}{\text{\# of spam emails} \quad + 2}$$

Multinomial:

$$P(w|spam) = \frac{\text{\# of times word } w \text{ appeared in spam emails} + 1}{\text{total \# words in spam emails} + m}$$

# Final Notes about (Naïve) Bayes Classifier

- Any density estimator can be plugged in to estimate $P(x_1,x_2, \ldots, x_m | y)$ for Bayes, or $P(x_i|y)$ for Naïve Bayes

- Real valued attributes can be discretized or directly modeled using simple continuous distributions such as Gaussian (Normal) distribution

- Naïve Bayes is wonderfully cheap and survives tens of thousands of attributes easily

- Laplace smoothing is important to avoid extreme probabilities