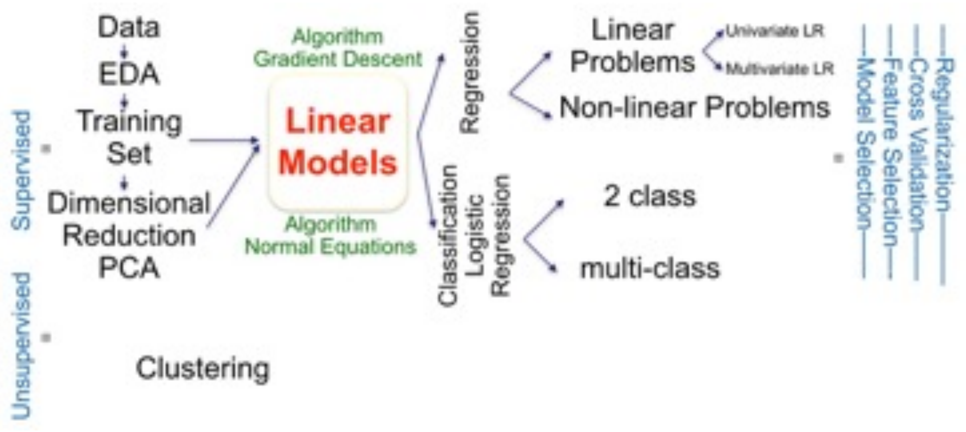
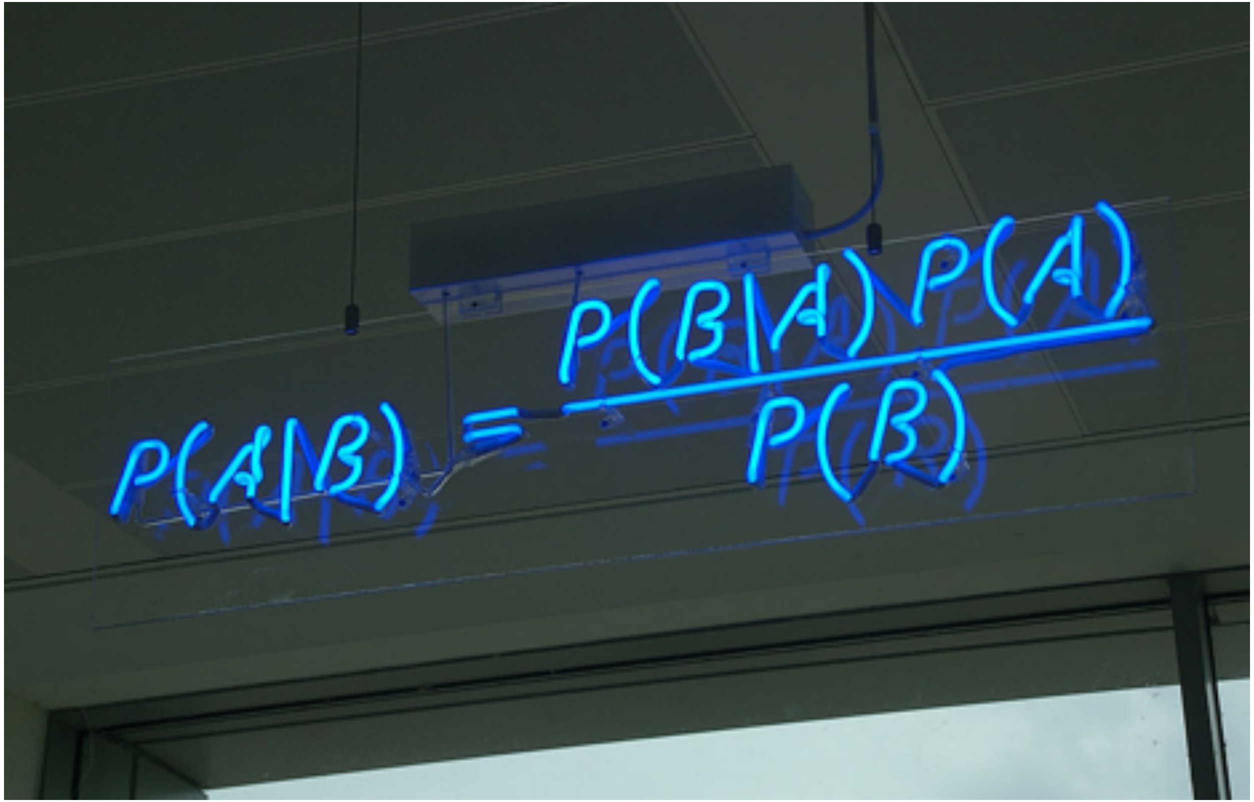
INTRO TO DATA SCIENCE
LECTURE 13: BAYES THEOREM & NAIVE BAYES CLASSIFIERS

Where are we on the data science road-map?



KEY CONCEPTS



Key Concepts - Frequentist or Bayesian?

Frequentist
- the classical view of statistics
- probability is the long-run frequency of an event
- returns a point estimate (a single number)

Example: what is the probability of a plane crash?

Bayesian
- probability measures your belief based on evidence (data)
- individuals may differ in their probability estimates (i.e. their beliefs) based upon their data
- as new information arrives this may change the belief
- returns a probability distribution or probabilities, from which a point estimate may be made

An imaginary frequentist function

IN: my coin has returned heads in the last 4 flips; is my coin bias?

OUT: ’Yes’

An imaginary bayesian function

IN: my coin has returned heads in the last 4 flips; I believe it is a fair coin; is my coin bias?

OUT: ‘Yes’ with p = 0.9, ‘No’ with p = 0.1

Key Concepts - Frequentist or Bayesian?

- The effect of prior information - I believe my coin is fair, i.e. has a 50/50 chance of coming up heads or tails
- As more evidence comes in this prior belief will get ‘washed out’
- As the number of examples increases and approaches infinity so the Bayesian estimate will align with the frequentist estimate

Key Concepts - Frequentist or Bayesian?

Andrew Gelman (2005):

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is "large enough," you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc.). N is never enough because if it were "enough" you'd already be on to the next problem for which you need more data.

PROBABILITY THEORY

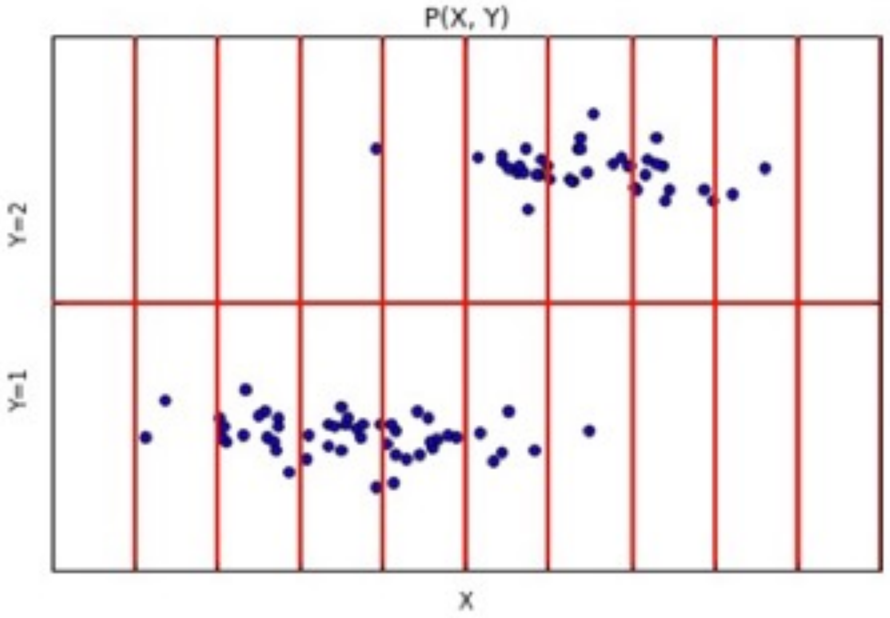
Key Concepts - Probability Theory

- A central foundation for pattern recognition
- Provides a consistent framework for handling uncertainty
  - Noise in data
  - Finite size of data sets
- Combined with decision theory allows for optimal predictions

Key Concepts - Probability Theory

- Probability (of an event) = the fraction of times that the event occurs out of the total number of events
- The set of all events is called the sample space
- By definition probabilities lie in the interval [0, 1]
- If events are mutually exclusive then the probabilities of all events in the sample space must sum to 1

For example, take N instances of 2 random variables, X and Y

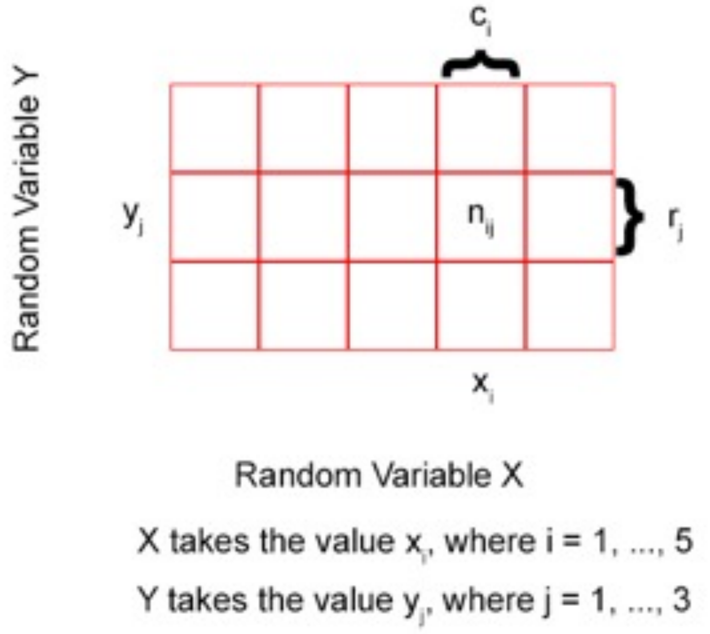


KEY CONCEPTS - JOINT DISTRIBUTION

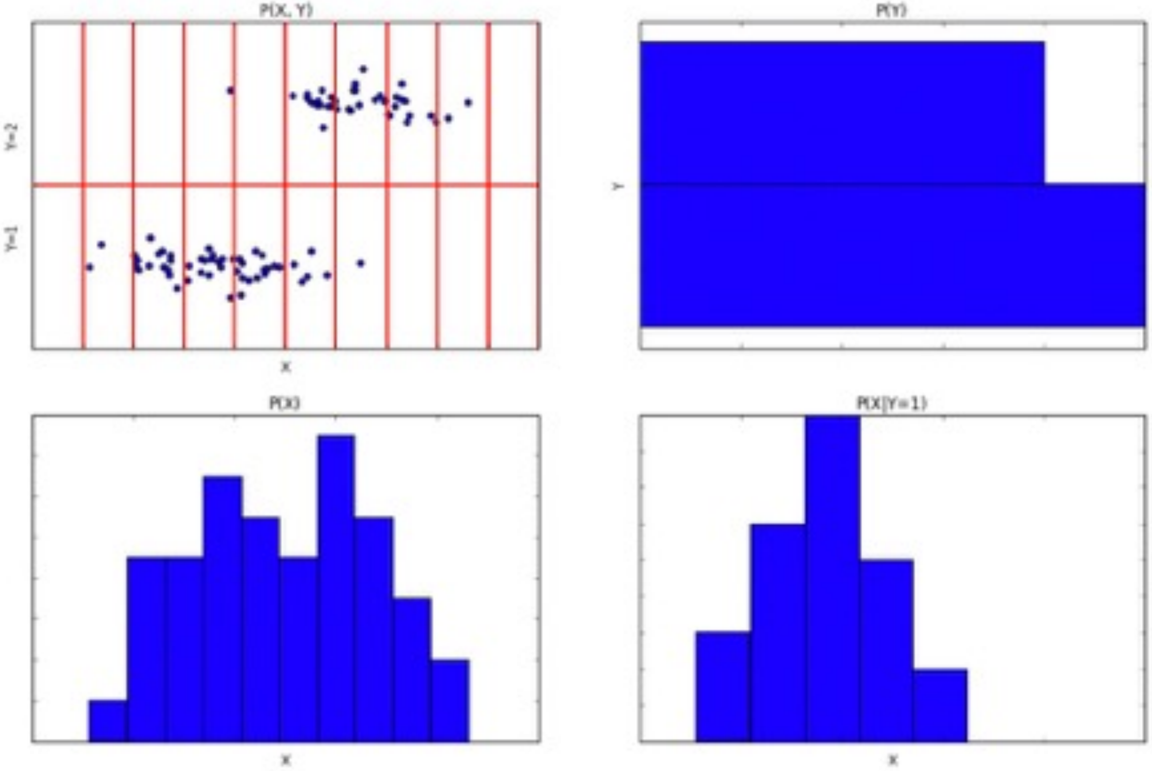
The number of instances of $X = x_i$ AND $Y = y_j$ is n_{ij}, which equals the number of points in the intersecting cell

The number of instances or points in column i, corresponding to $X = x_i$ is c_i

The number of points in row j, corresponding to $Y = y_j$ is r_j



KEY CONCEPTS - JOINT, MARGINAL AND CONDITIONAL DISTRIBUTIONS



Marginal Distribution

Marginal Distribution

Conditional Distribution

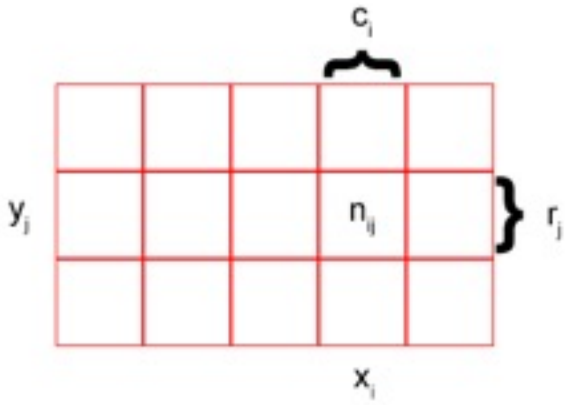
Joint Probability

The probability that X takes the value x_i AND Y takes the value y_j is denoted:

$p(X = x_i, Y = y_j)$ or $P(X, Y)$

$p(X, Y) = \frac{n_{ij}}{N}$

and called the joint probability of X and Y



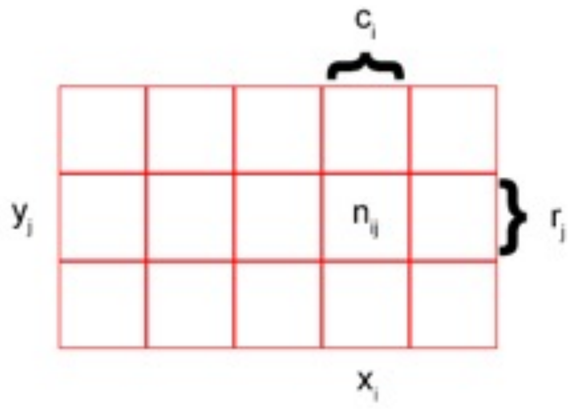
Key Concepts - Marginal Probability

The probability that X takes the value x_i irrespective of the value of Y is denoted:

$p(X = x_i)$ or $P(X)$

$p(X) = \frac{c_i}{N}$

Y has been 'marginalized' out…

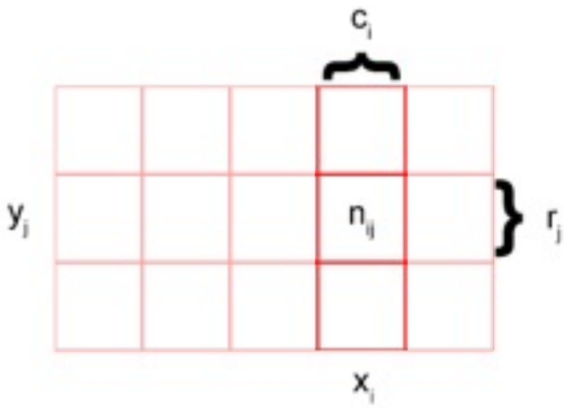


Key Concepts - Marginal Probability

c_i may be expressed as the sum of the values of n_{ij} summed over all the values of j

$$c_i = \sum_j nij$$

$$\therefore p(X) = \frac{c_i}{N} = \frac{\sum_j nij}{N} = \sum_j \frac{n_{ij}}{N} = \sum_Y p(X, Y)$$



KEY CONCEPTS - SUM RULE

The SUM RULE:

$$p(X) = \sum_Y p(X, Y)$$

$p(X)$ is called the MARGINAL PROBABILITY

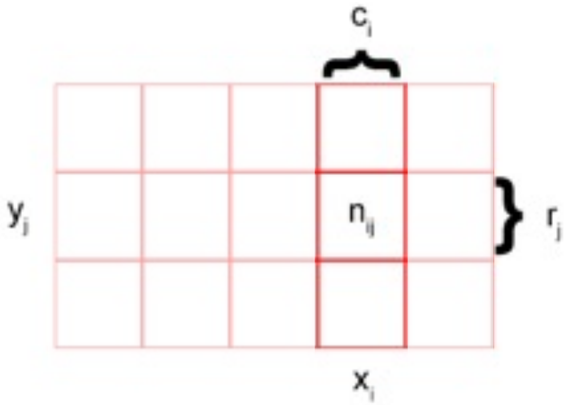
Other variables, Y here, are being summed out or marginalized

KEY CONCEPTS - CONDITIONAL PROBABILITY

Consider those instances for which $X = x_i$, then the fraction of such instances for which $Y = y_j$ is written $p(Y = y_j | X = x_i)$

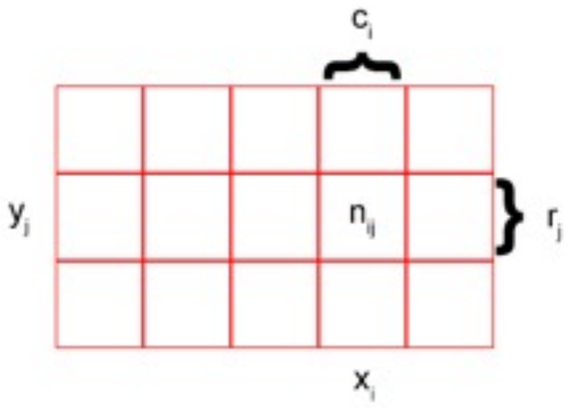
$p(Y|X)$ is called the CONDITIONAL PROBABILITY.

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$



Key Concepts - Rules of Probability

$$p(X, Y) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \times \frac{c_i}{N} = p(Y|X)p(X)$$



The PRODUCT RULE:

$$p(X, Y) = p(Y|X)p(X)$$

RULES OF PROBABILITY:

Sum Rule: $P(X) = \sum_Y P(X, Y)$

Product Rule: $P(X, Y) = P(Y|X)p(X)$

BAYES' THEOREM:

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

$$P(X, Y) = P(Y|X)P(X) \text{ and } P(Y, X) = P(X|Y)P(Y)$$

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

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

Bayes' Theorem

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

OR

$$\text{Posterior} = \frac{\text{Likelihood x Prior}}{\text{Normalization}}$$

INDEPENDENCE

If $p(X, Y) = p(X)p(Y)$ then X and Y are said to be independent

This means that

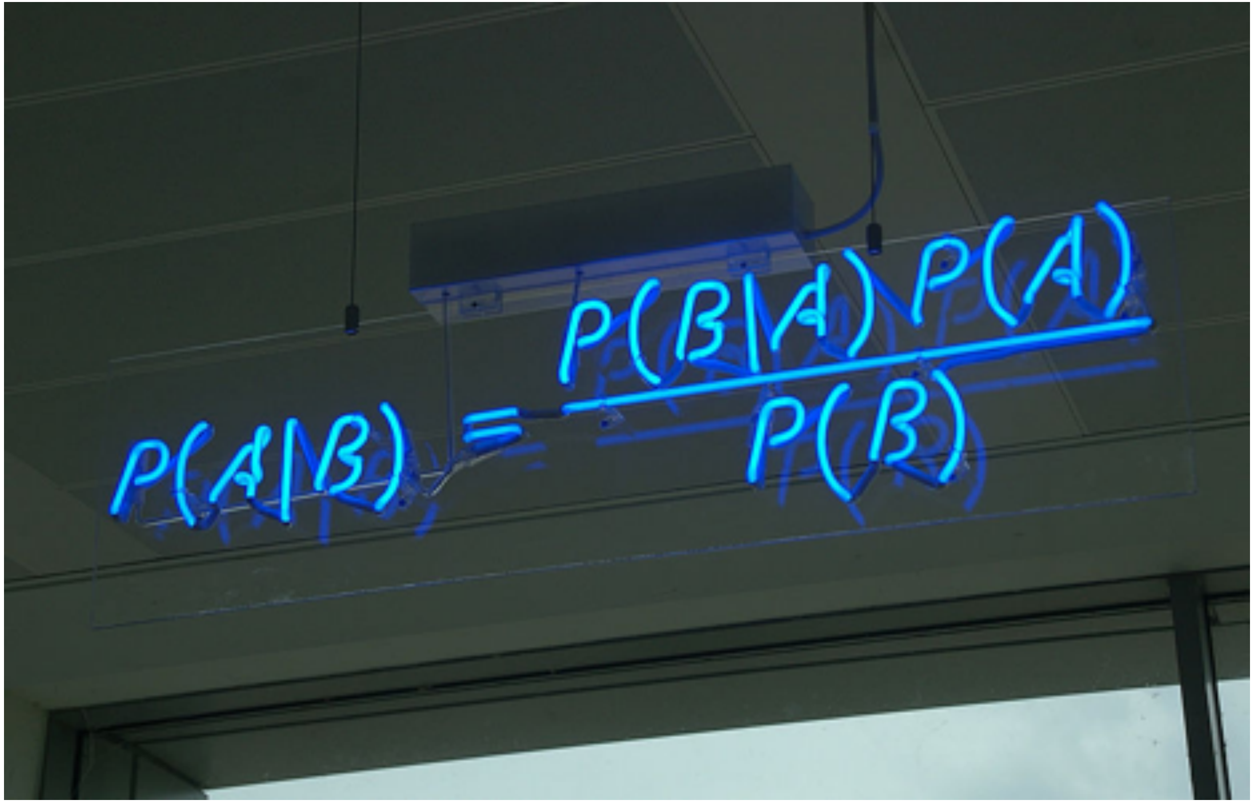
$p(Y|X) = P(Y)$ So the conditional distribution Y given X, is indeed independent of X

KEY CONCEPTS - THOMAS BAYES 1701 - 1761

English clergyman, amateur scientist and mathematician. One problem of his time concerned ‘inverse probability’, to which he proposed a solution in a paper called ‘Essay towards solving a problem in the doctrine of chances’. This was published 3 years after his death.

Bayes only formulated his theory for the case of the uniform prior. Pierre-Simon Laplace who independently rediscovered the theory in general form and demonstrated its broad applicability.

Key Concepts - Bayes



KEY CONCEPTS - BAYES

- The overall goal of Bayesian computation is to determine the posterior distribution of a particular variable given some data

- From this distribution you can derive point estimates

- It is important to realize the importance of the prior distribution. Often, in real world problems, we do not know the prior. A ‘default’ assumption is equal priors

- The Likelihood is derived from the (labeled) data

Comments about Bayes:

- It is a relatively simple algebraic relationship
- It is extremely powerful as a computational tool
- It is unbelievably confusing

- If it all sounds crazy don't worry!

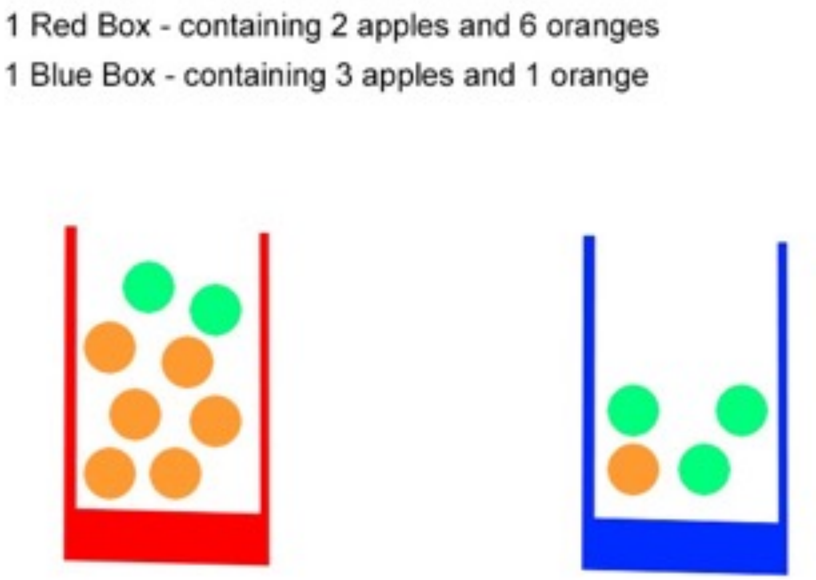
KEY CONCEPTS - PROBABILITY - A SIMPLE EXAMPLE

Suppose you pick a box, blindfolded, and over the course of many trials determine that you pick the blue box 60% of the time

Having chosen a box to pick fruit from you randomly select an item of fruit

Each piece of fruit in a box is equally likely to be chosen

You replace the fruit between trials



- Probability box chosen is blue,
  - P(Box=blue) = 0.6, and
  - P(Box=red) = 1.0 - 0.6 = 0.4

- Choosing a box is mutually exclusive, you cannot choose both boxes at the same time, it is either one or the other

1. Conditional Probability:

What is P(F=apple | B=blue)?

2. Marginal Probability:

What is P(apple)

3. Bayes' Theorem:

Given you have selected an apple what is the probability it came from the blue box?

Key Concepts - Classification Using Bayes

- Family of algorithms based upon a common algorithmic structure
- Highly applicable where frequency counts are the features
- Highly scalable
- Very competitive, in terms of performance
- Only requires a relatively small amount of date in order to make estimates

KEY CONCEPTS - NAIVE BAYES CLASSIFIERS

$P(C_k|\vec{x})$ where $x = (x_1, x_2, \ldots, x_N)$

using Bayes:

$$P(C_k|\vec{x}) = \frac{P(\vec{x}|C_k) \times P(C_k)}{P(\vec{x})}$$

Because we will be comparing 2 probabilities (for at least a 2 class problem) we can ignore the denominator. The relatively magnitude of $P(C_k|\vec{x})$ will be the same whether or not we divide by $P(\vec{x})$

Key Concepts - Naive Bayes Classifiers

Let $N = 3$

$$P(C_k|\vec{x}) = P(C_k|x_1, x_2, x_3)$$

Expand using Bayes and forget the denominator

$$= P(x_1, x_2, x_3|C_k) \times p(C_k)$$

Now use the product rule to expand out the first term

$$= P(x_1|C_k) \times P(x_2, x_3|C_k, x_1) \times p(C_k)$$

$$= P(x_2|C_k, x_1) \times P(x_3|C_k, x_1, x_2) \times P(x_1|C_k) \times p(C_k)$$

RULES OF PROBABILITY:

Sum Rule: $P(X) = \sum_Y P(X, Y)$

Product Rule: $P(X, Y) = P(Y|X)p(X)$

KEY CONCEPTS - ESTIMATING THE LIKELIHOOD FUNCTION

$$
\begin{aligned}
P(x_1, x_2, x_3|C_k) &= P(x_1|C_k) \times P(x_2, x_3|C_k, x_1) \\
&= P(x_2|C_k, x_1) \times P(x_3|C_k, x_1, x_2)
\end{aligned}
$$

To estimate these conditional probabilities would require data on all possible combinations!

It is simply impractical

Key Concepts - Naive Bayes Classifiers

If $x_1, x_2,$ and x_3 **are independent then**

$P(x_2|C_k, x_1) = P(x_2|C_k)$ **and**

$P(x_3|C_k, x_1, x_2) = P(x_3|C_k)$

INDEPENDENCE

If $p(X, Y) = p(X)p(Y)$ then X and Y are said to be independent

This means that

$p(Y|X) = P(Y)$ So the conditional distribution Y given X, is indeed independent of X

$$P(C_k|\vec{x}) = P(C_k, x_1, x_2, x_3)$$

$$= P(C_k) \times P(x_1|C_k) \times P(x_2|C_k) \times P(x_3|C_k)$$

Key Concepts - Naive Bayes Classifiers

$$\hat{y} = \text{argmax}_{k \in 1,\ldots,K}\, P(C_k) \prod_{i=1}^{N} P(x_i|C_k)$$

KEY CONCEPTS - THE 'NAIVE' IN NAIVE BAYES

- It is called naive, because we choose to ignore an important assumption
- We assume that the features in x are independent of each other
- Naive Bayes classifiers do not output accurate probabilities!, they are meant to classify purely based upon which class output is greatest.
- The independence assumption avoids the curse of dimensionality problem, hence the reason why these models scale so well

KEY CONCEPTS - SKLEARN - NAIVE BAYES CLASSIFIERS

- Gaussian. For each class the features of x are the mean and std deviation

- Multinomial. For each class the features of x are a histogram representing the frequency of occurrence

- Bernoulli. For each class the features of x are the occurrence or not