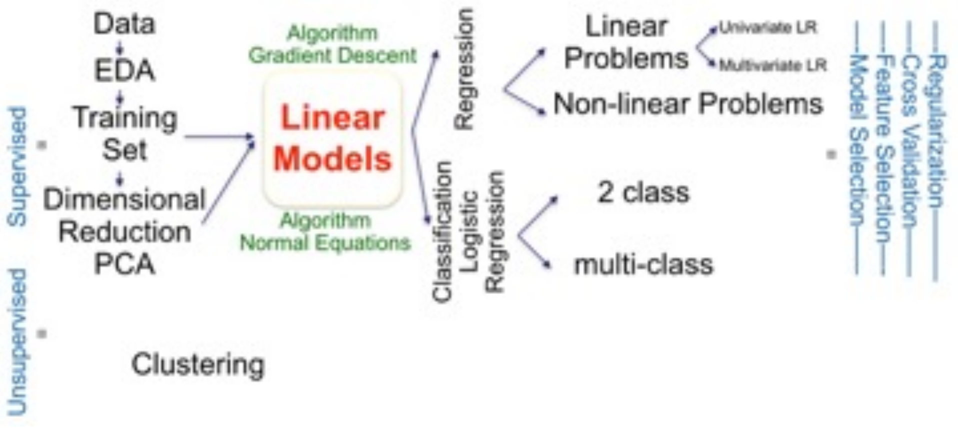
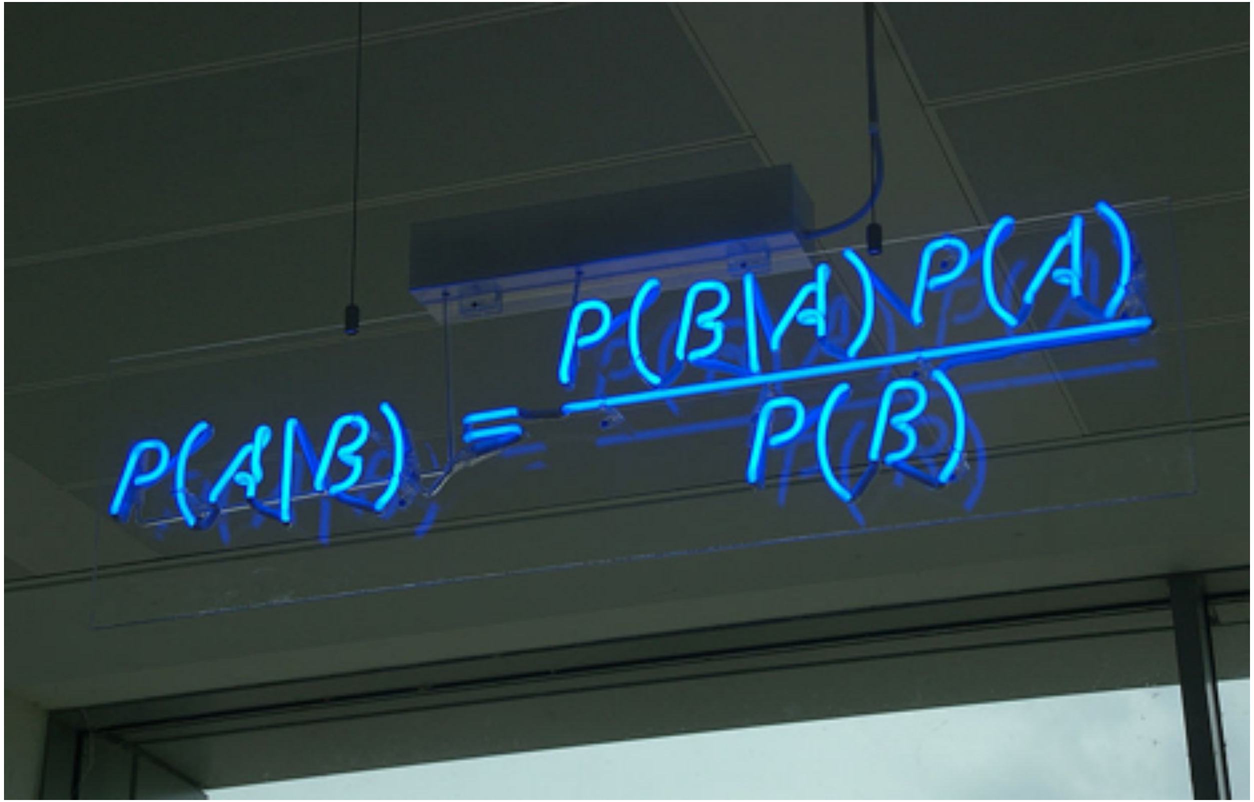
GENERAL ASSEMBLY

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

WHERE ARE WE ON THE DATA SCIENCE ROAD-MAP?



Key Concepts



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

- 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]

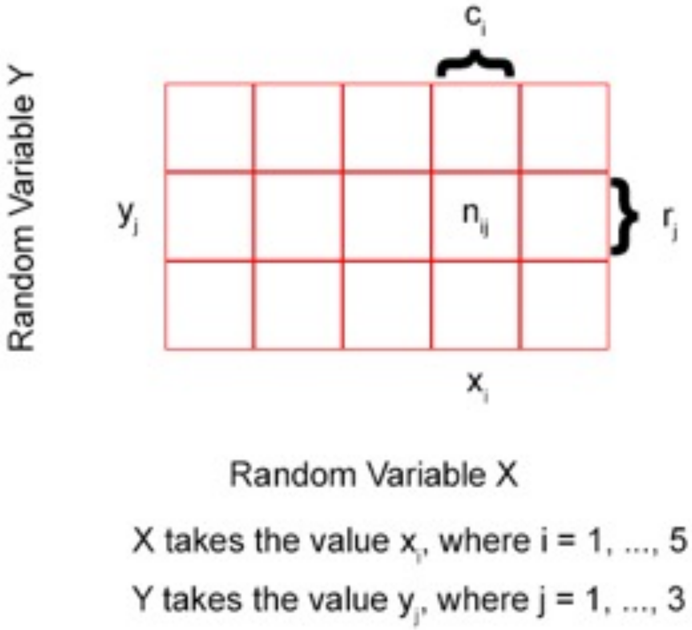
If events are mutually exclusive then they must sum to 1

KEY CONCEPTS - JOINT DISTRIBUTION

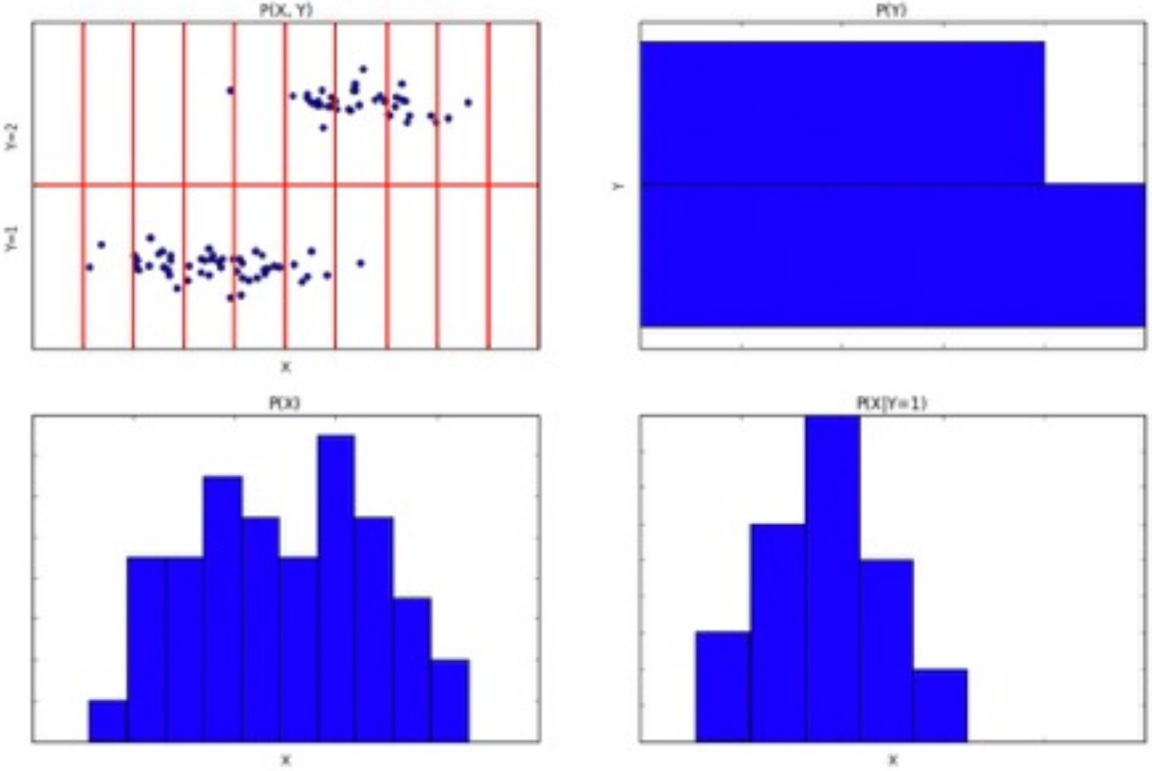
For example, take N instances of 2 random variables, X and Y
The number of instances of $X = x_i$ AND $Y = y_j$ is n_{ij}
Which equals the number of points in the intersecting cell

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 DISTRIBUTION



Marginal Distribution

Marginal Distribution

Conditional Distribution

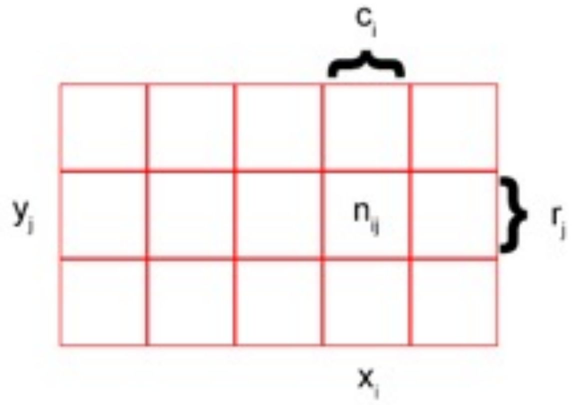
Joint Probability

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

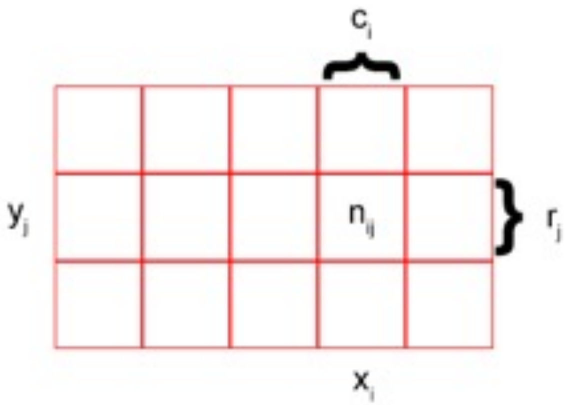


Key Concepts - Marginal Probability

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

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

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

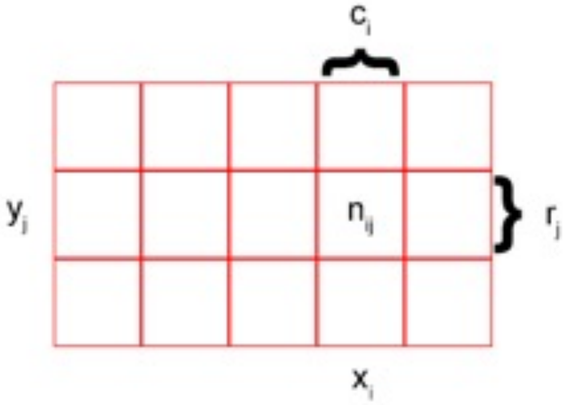


KEY CONCEPTS - MARGINAL PROBABILITY

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

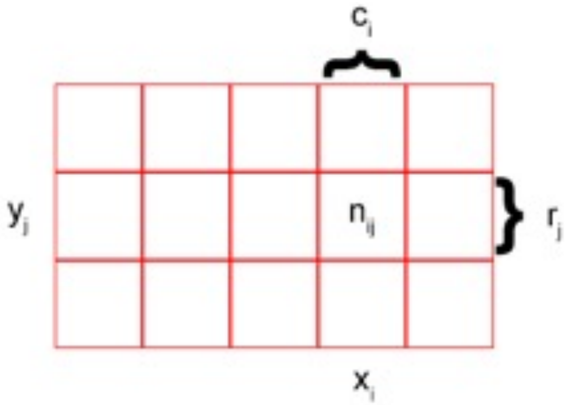
Other variables, Y here, are being summed out or marginalized

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 $p(Y = y_j | X = x_i)$

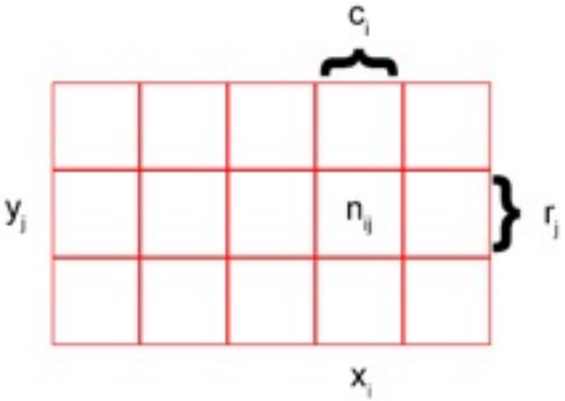
$p(Y|X)$ is called the CONDITIONAL PROBABILITY.

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



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:

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

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

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

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

KEY CONCEPTS - BAYES' THEOREM

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

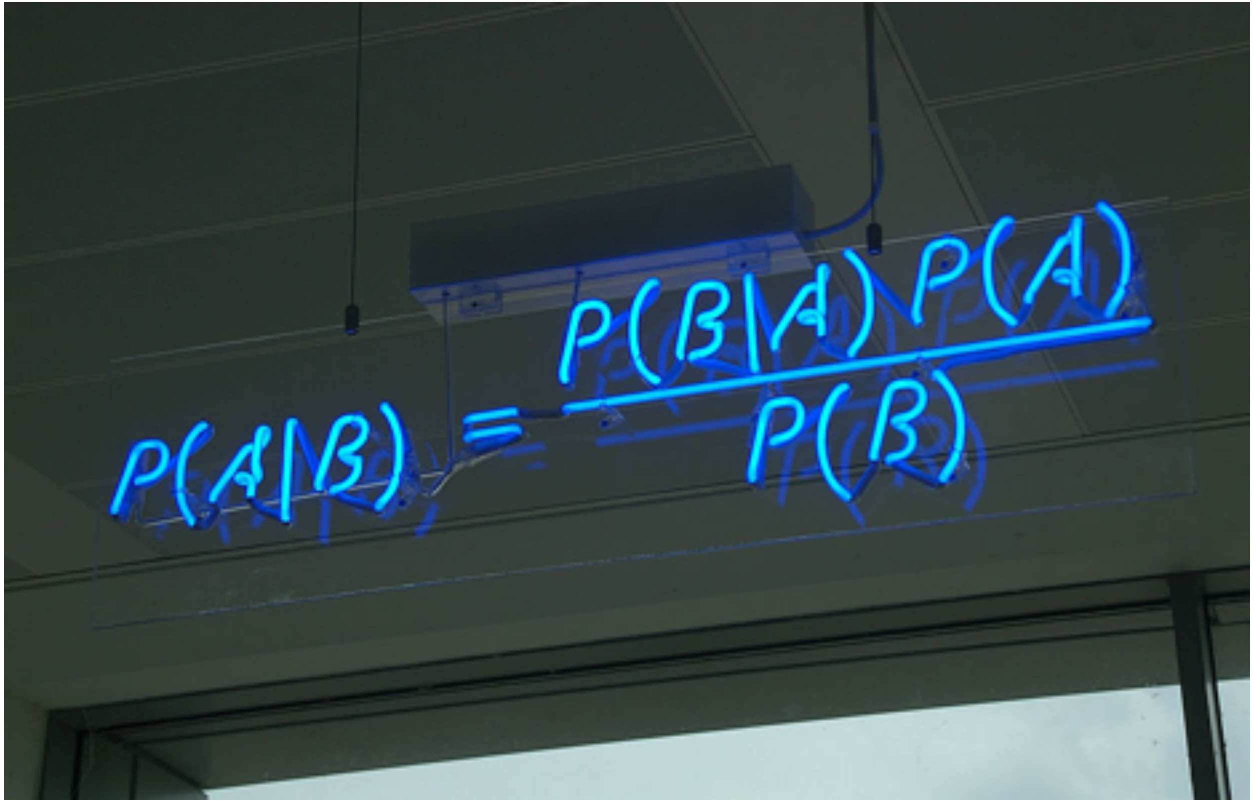
$p(Y|X) = P(Y)$ So the conditional distribution Y given X, is indeed independent of X

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



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!

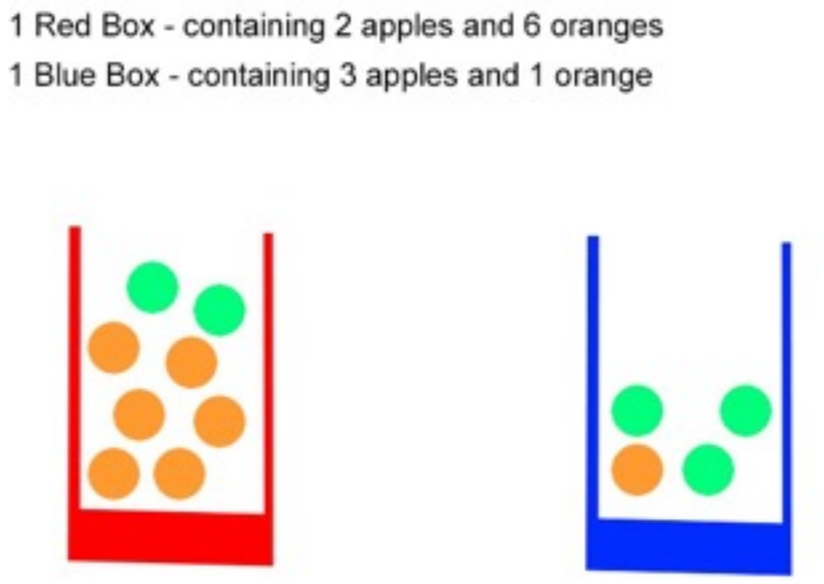
Key Concepts - Probability - A Simple Example

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=red) = 0.4, and P(Box=blue) = 1.0 - 0.4 = 0.6
- Choosing a box is mutually exclusive, you cannot choose both boxes at the same time, it is either one or the other

What is the probability of choosing an apple?

Having chosen an orange what is the probability of having chosen the blue box?

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

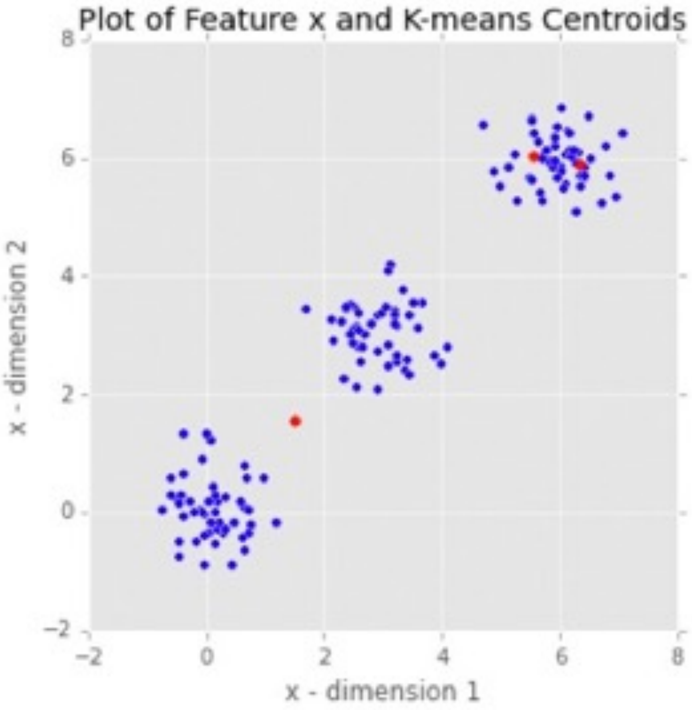
KEY CONCEPTS - K-MEANS ADVANTAGES

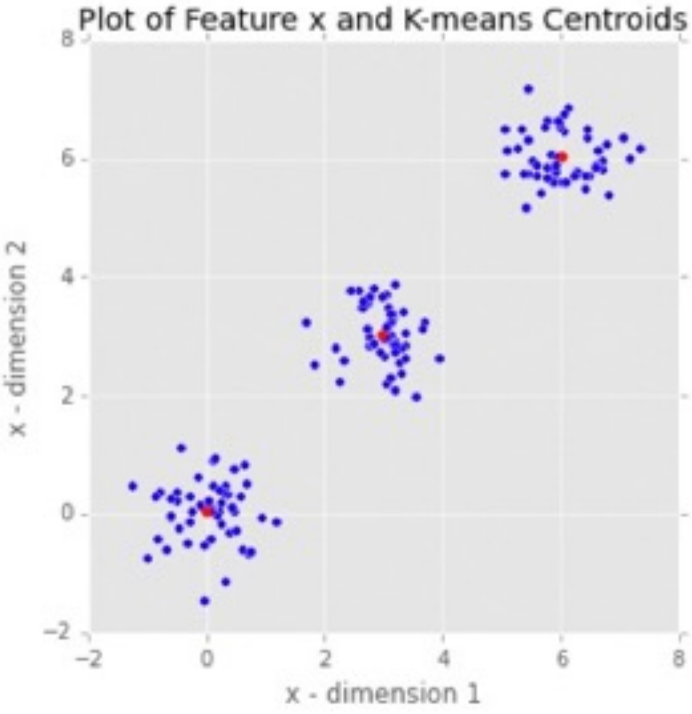
- The use of euclidean distance makes the algorithm susceptible to outliers

- You have to find a good value for k

- K-means is subject to the local minima problem.
  - An un-lucky initial randomization of centroids may yield a poor clustering result

KEY CONCEPTS - K-MEANS ADVANTAGES LOCAL OPTIMA

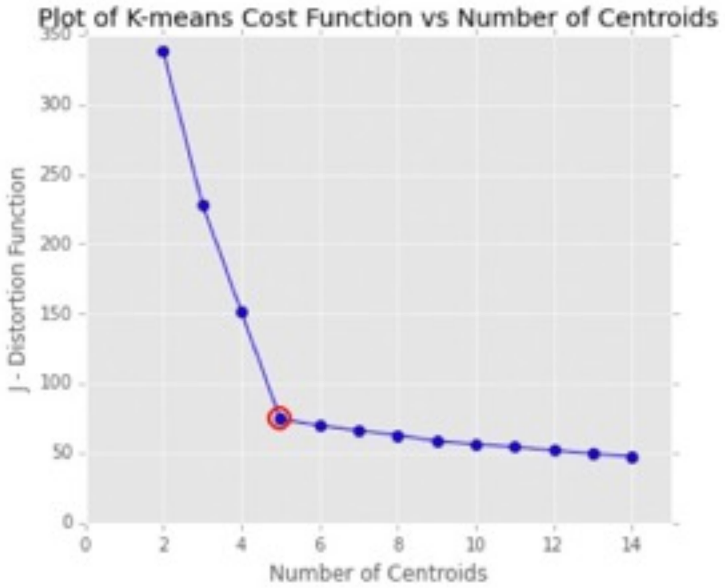


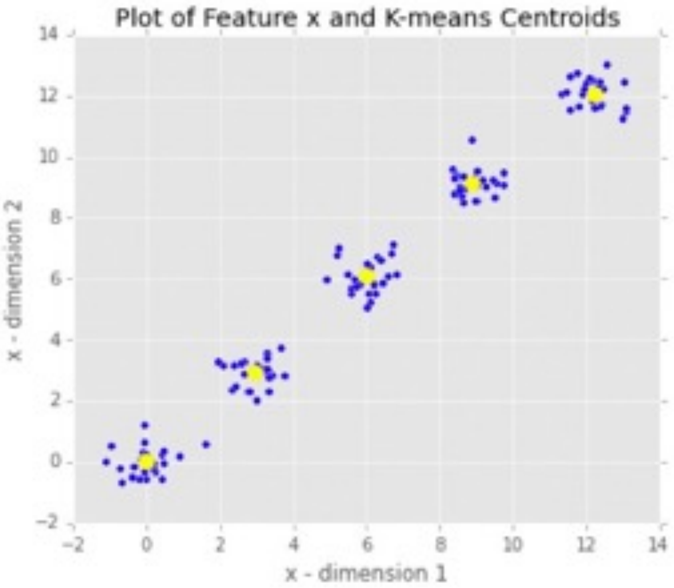


KEY CONCEPTS - K-MEANS ALGORITHM OPTIMIZING - FINDING A GOOD VALUE FOR K

- The elbow (or knee-of-the-curve) method plots the value of the cost function produced by different values of k
- As k increases, the average distances (and hence J) will decrease; each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids
- However, the improvement to J will decline as k increases. The value of k at which the improvement to J declines the most is called the elbow

KEY CONCEPTS - K-MEANS ALGORITHM OPTIMIZING - FINDING A GOOD VALUE FOR K





Key Concepts - K-means Algorithm Optimizing - Avoiding Local Minima

- Because K-means is a fast algorithm you can usually run 50, 100, or even 200 runs, each with a different random initialization, so as to avoid poor solutions

- For each choice of K, you would run the algorithm 50, 100 or even 200 times with a different random starting configuration

- This, in general, ‘solves’ the local minima problem

OTHER CLUSTERING ALGORITHMS

- Affinity Propagation
- MeanShift
- Spectral
- Ward
- Agglomerative
- DBSCAN