11.02.2020

Statistical Methods in AI (CSE/ECE 471)

Lecture-10: Unsupervised Learning (k-means, GMM)

Ravi Kiran (ravi.kiran@iiit.ac.in)

Vineet Gandhi (v.gandhi@iiit.ac.in)

Center for Visual Information Technology (CVIT)

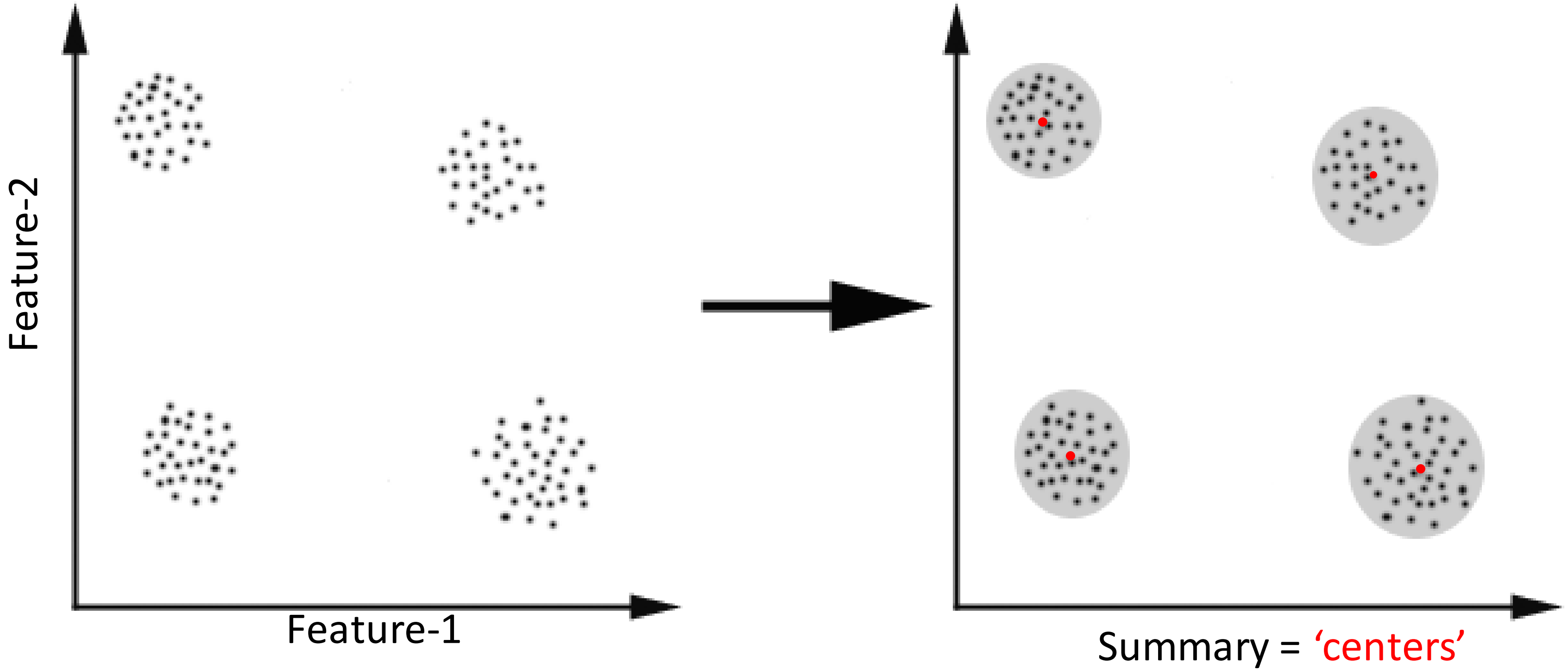
IIIT Hyderabad

Unsupervised Learning → Clustering

Group similar things e.g. images

[Goldberger et al.]

Perspective: Clustering as a 'summary' of input data version 2



$$\{x^{(1)}, \ldots, x^{(m)}\} \qquad x^{(i)} \in \mathbb{R}^n$$

The k-means clustering algorithm is as follows:

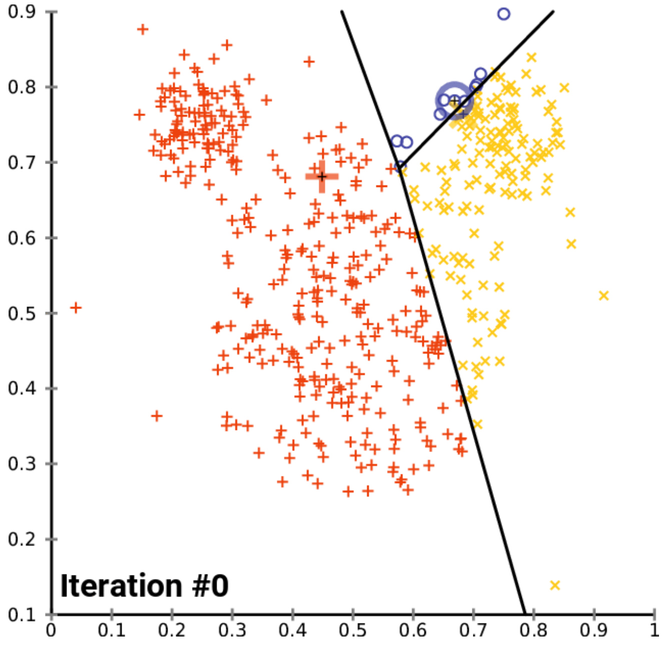
1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.

2. Repeat until convergence: {

   For every i, set
$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set
$$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

}



Repeat until convergence: {

For every i, set

$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

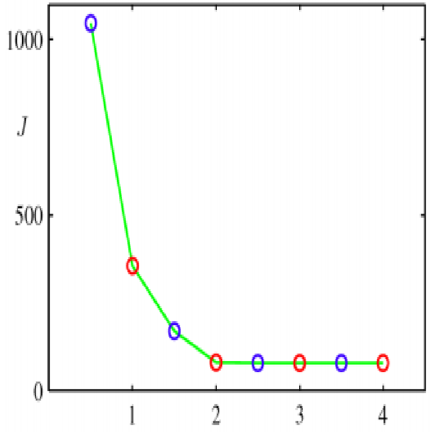
For each j, set

$$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

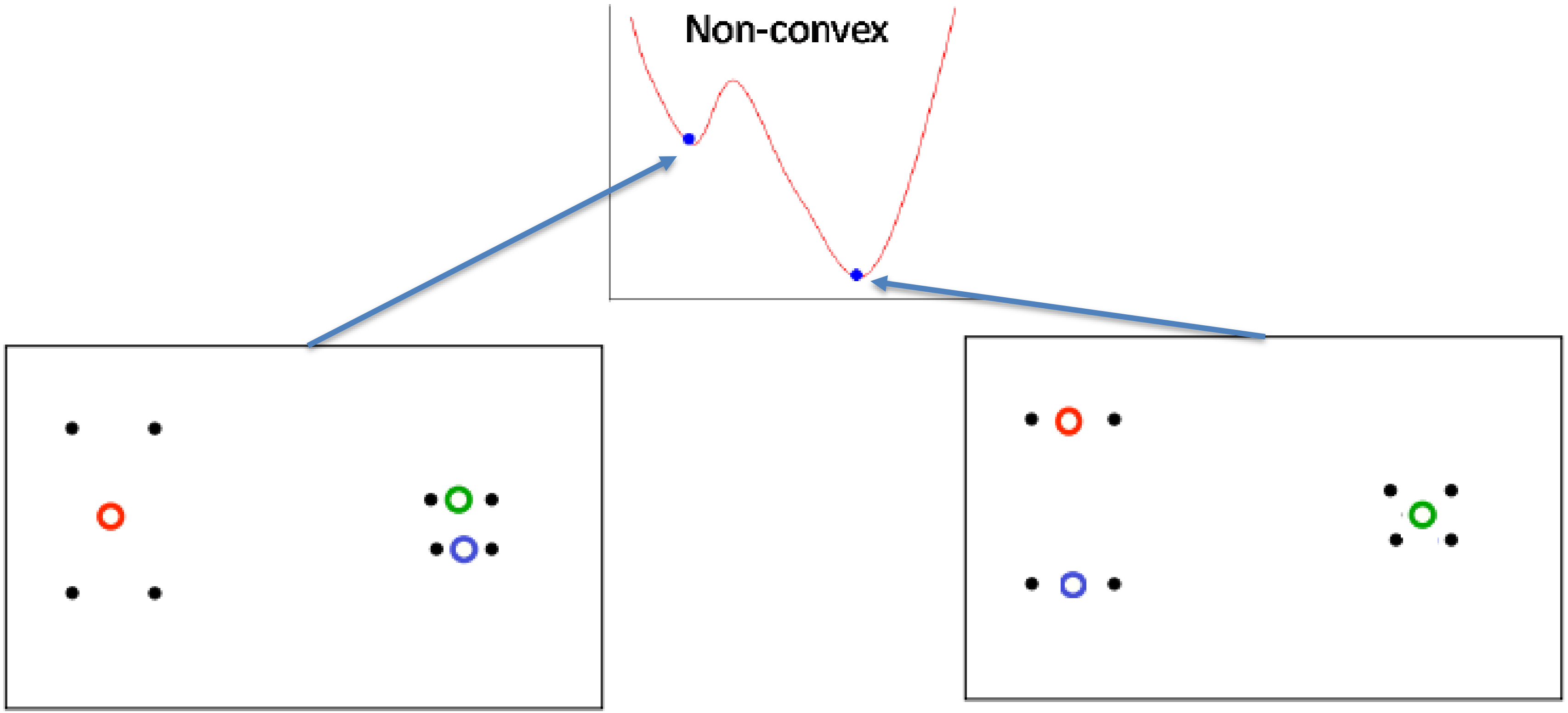
}

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\| x_{ki} - \mu_k \right\|^2$$

- Whenever an assignment is changed, the sum squared distances J of data points from their assigned cluster centers is reduced.
- Whenever a cluster center is moved, J is reduced.
- Test for convergence: If the assignments do not change in the assignment step, we have converged (to at least a local minimum).



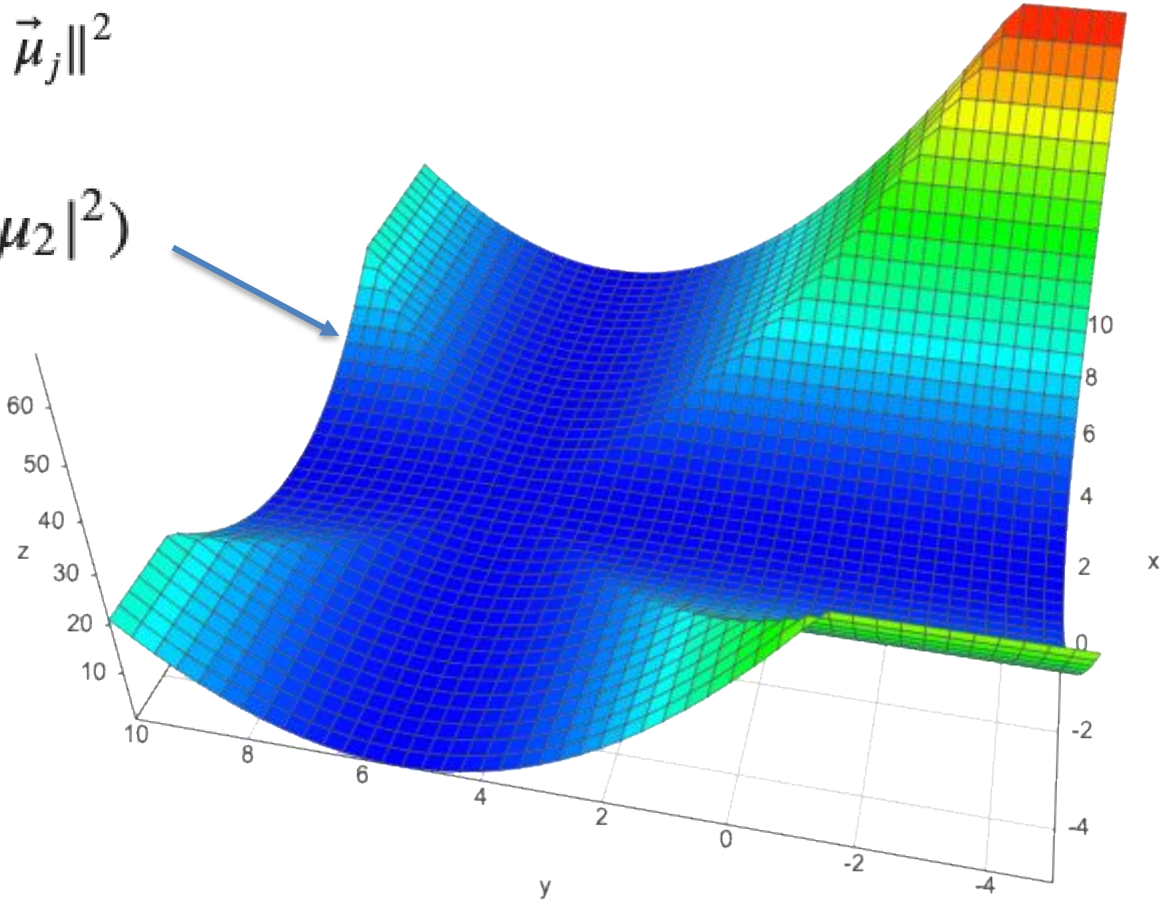
Objective function for k-means is non-convex



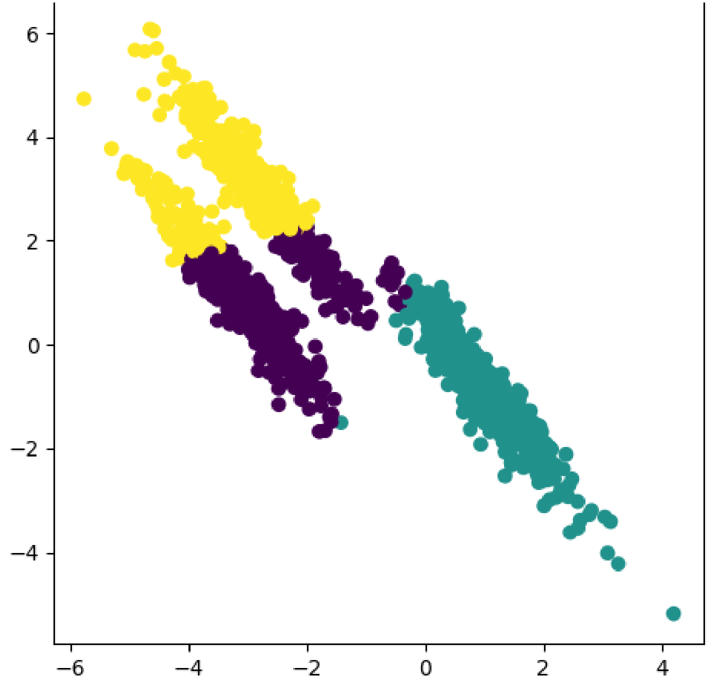
Let $\vec{x}_i, i = 1, 2, \ldots, n$ be the data points and $\vec{\mu}_j, j = 1, 2, \ldots, k$ be the k mean values.

$$\text{minimize} \quad \sum_{i=1}^{n} \min_{j=1..k} \|\vec{x}_i - \vec{\mu}_j\|^2$$

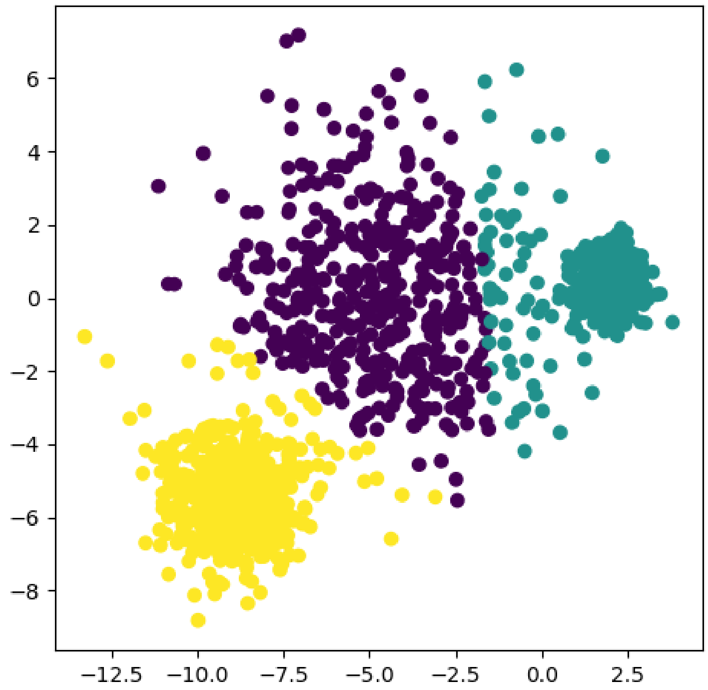
$$\min(|x_i - \mu_1|^2, |x_i - \mu_2|^2)$$



Limitations of k-means

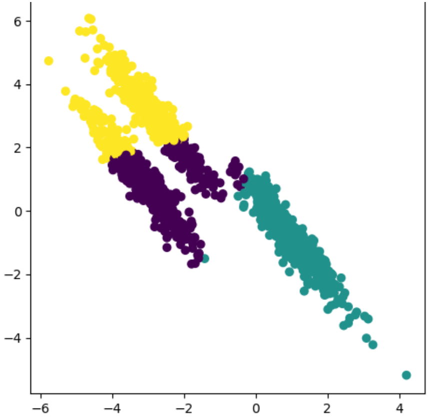


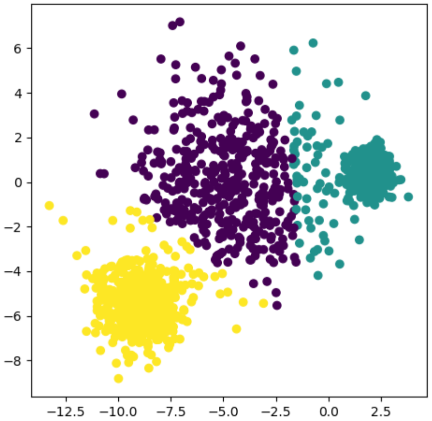
- Euclidean distance → spherical cluster boundaries



- Hard assignments → hard to characterize 'border cases'

- Can we have a distance-from-center based on ‘shape’ of the cluster ?
- Can we go beyond ‘hard’ assignments of points to clusters ?



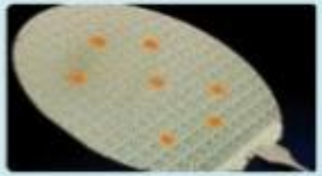


Uncertainty arises from many sources

Process Uncertainty

Processes contain "randomness"

Uncertain travel times



Semiconductor yield

Data Uncertainty

Data input is uncertain

Intended Spelling → Text Entry → Actual Spelling

GPS Uncertainty

Testimony

{Paris Airport}

Ambiguity

Contaminated?

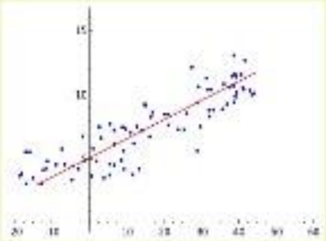
Rumors

{John Smith, Dallas}
{John Smith, Kansas}

Conflicting Data

Model Uncertainty

All modeling is approximate



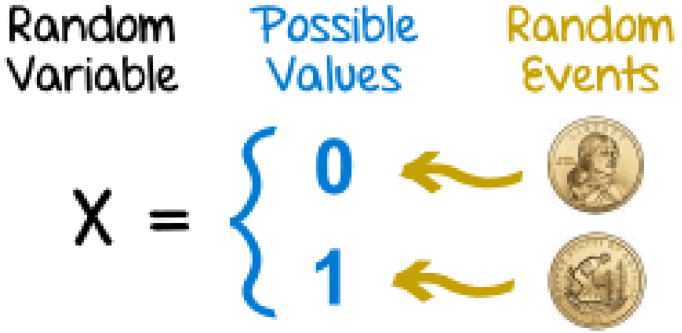
Fitting a curve to data

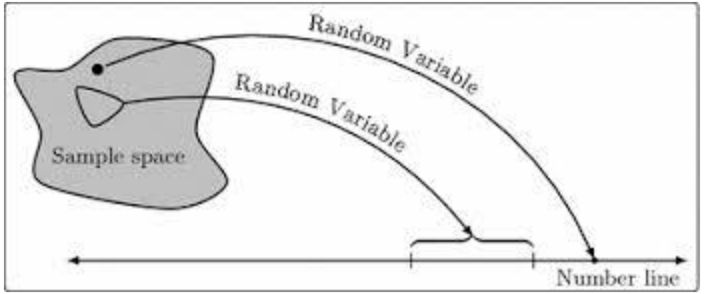
Forecasting a hurricane (www.noaa.gov)

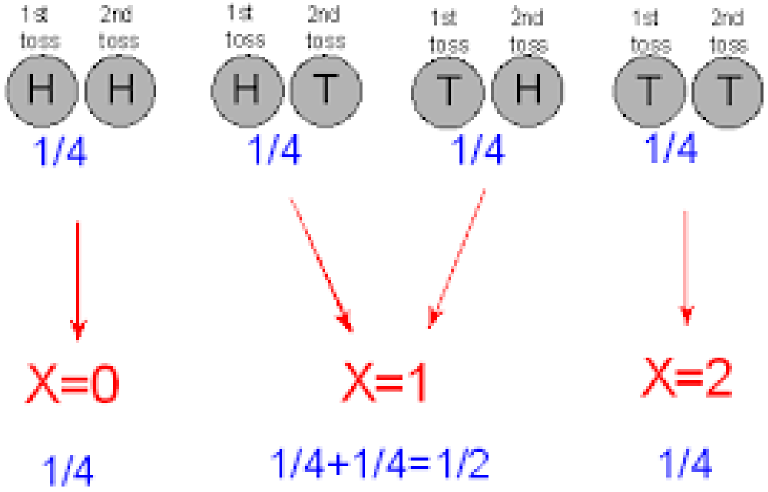

PROBABILITY = EVENT / OUTCOMES
wikiHow to Calculate Probability

Random Variables

R.V. = A numerical value assigned to a subset of events from a random experiment







P(X = a) = Probability of events associated with RV X taking the value 'a'

Random variables

- A **discrete random variable** can assume a countable number of values.
  - Number of steps to the top of the Eiffel Tower*

Random variables

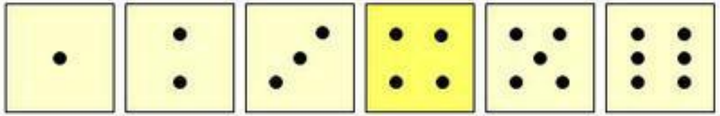
- A **discrete random variable** can assume a countable number of values.
  - Number of steps to the top of the Eiffel Tower*
- A **continuous random variable** can assume any value along a given interval of a number line.
  - The time a tourist stays at the top once s/he gets there

*Believe it or not, the answer ranges from 1,652 to 1,789. See Great Buildings

Discrete Random Variables

- Can only take on a countable number of values

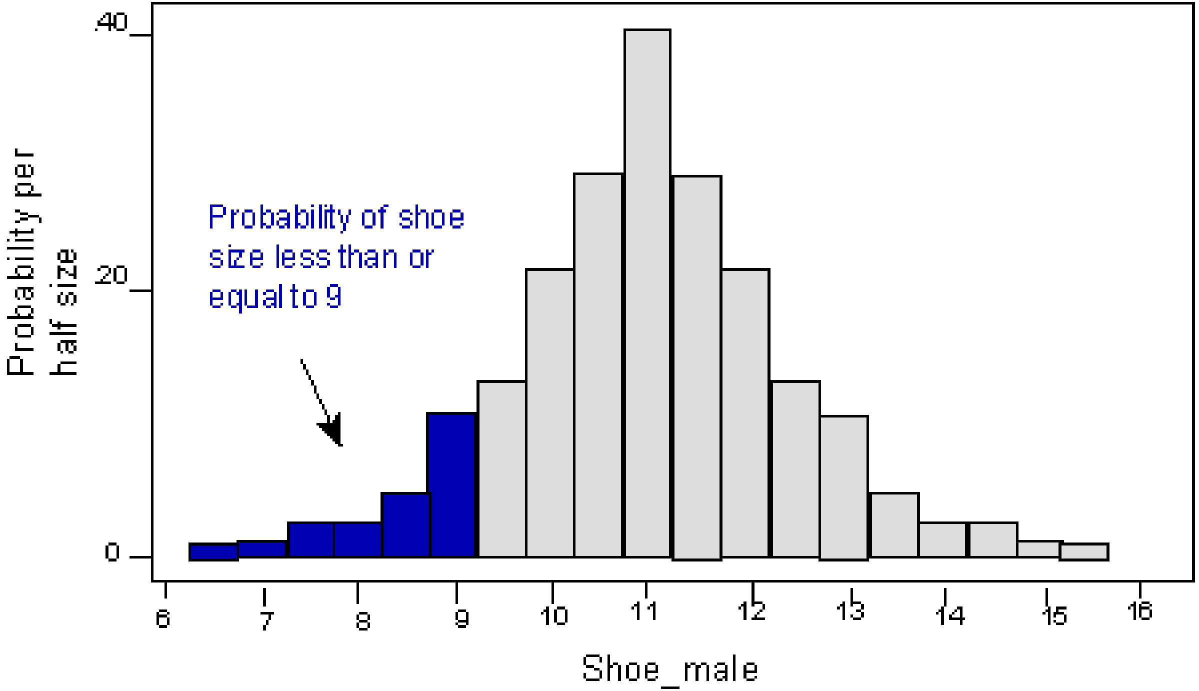
Examples:



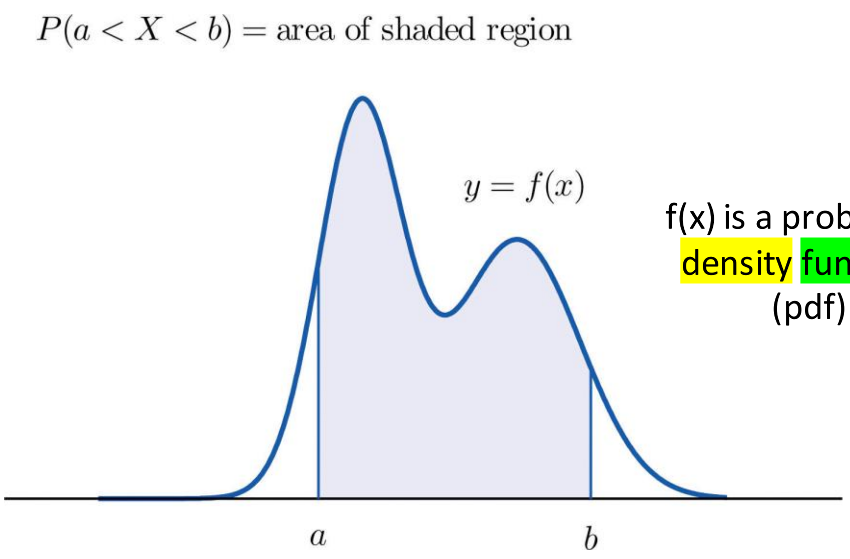
- **Roll a die twice**
  Let X be the number of times 4 comes up
  (then X could be 0, 1, or 2 times)

- **Toss a coin 5 times.**
  Let X be the number of heads
  (then X = 0, 1, 2, 3, 4, or 5)

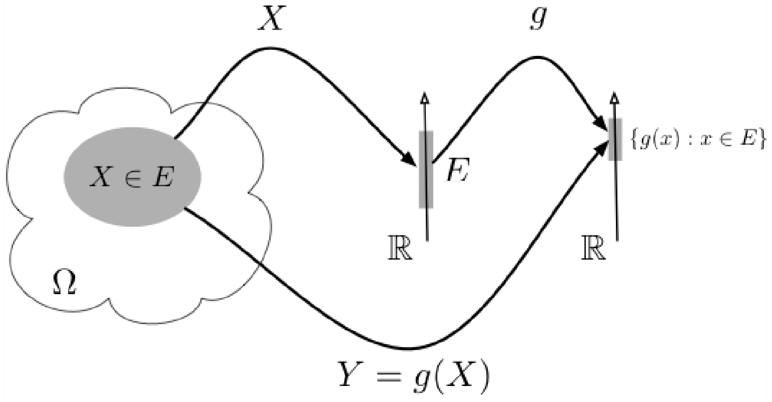
Discrete Random Variable



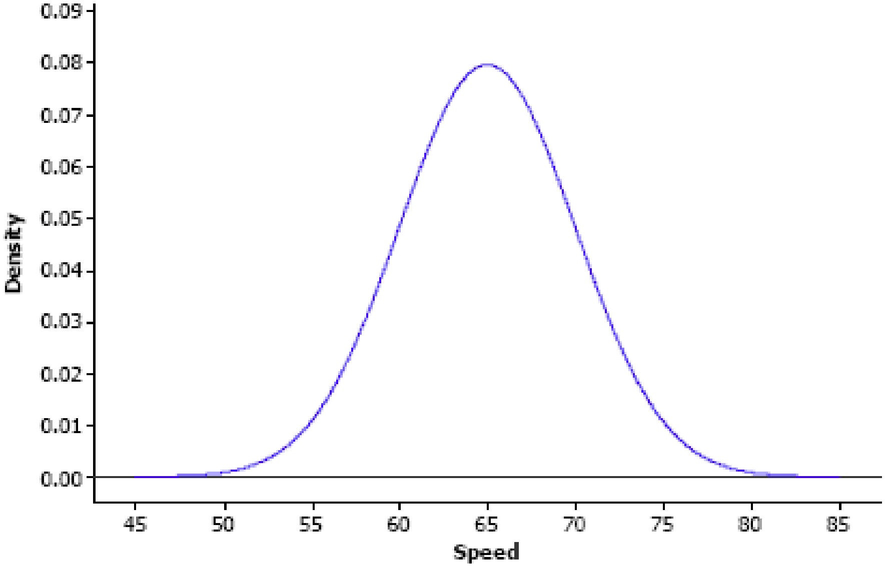
Continuous random variable

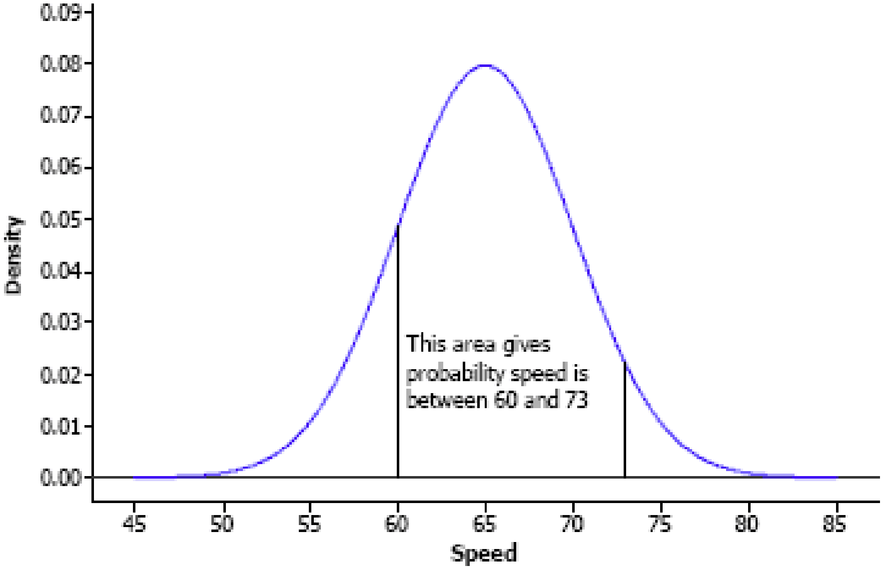


f(x) is a probability density function (pdf)

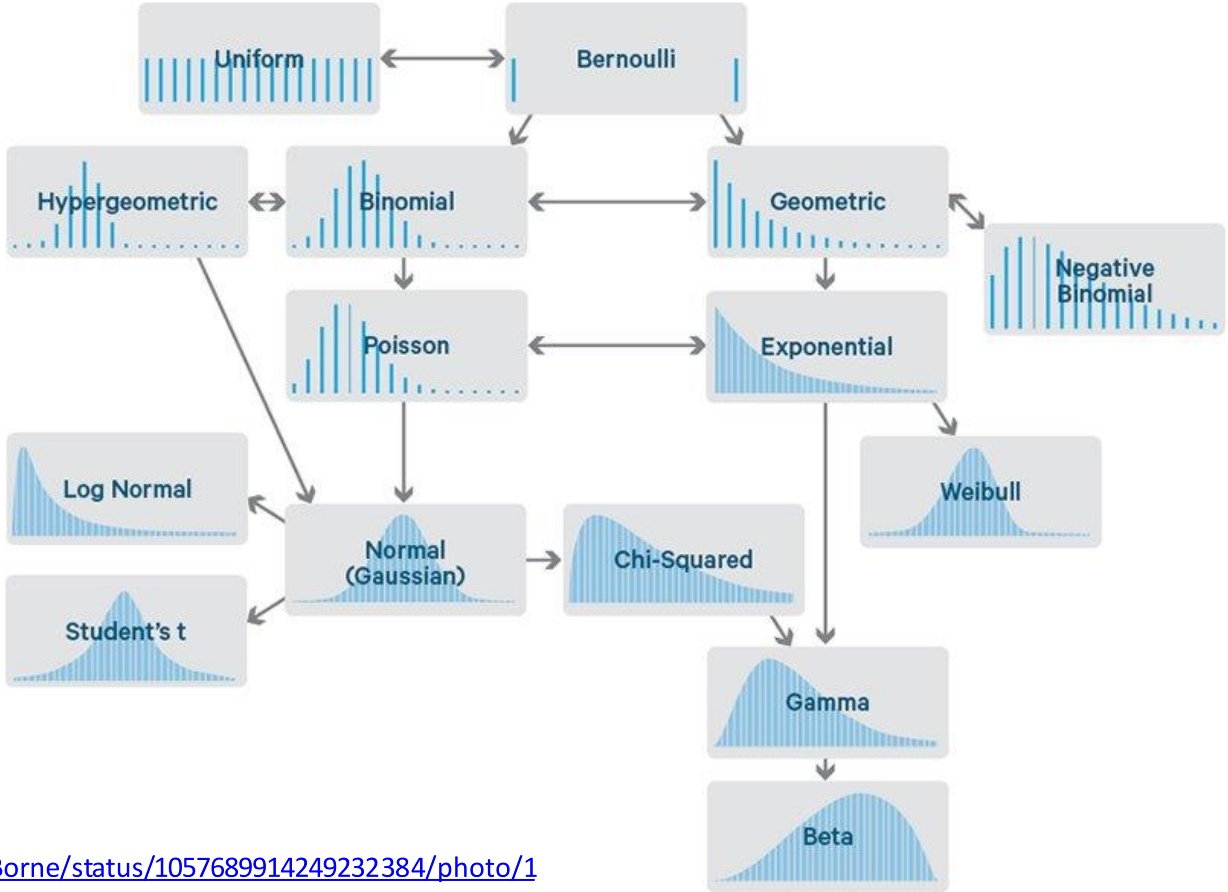


Continuous random variable - example



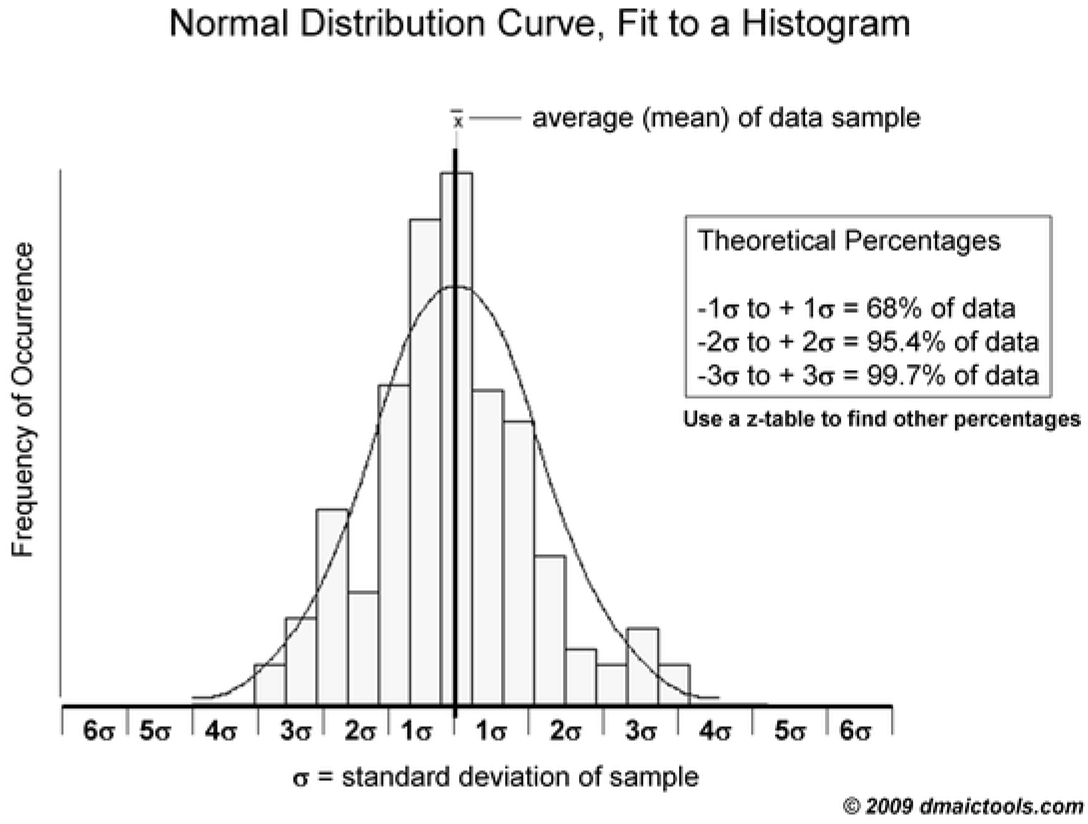


Some common probability distributions

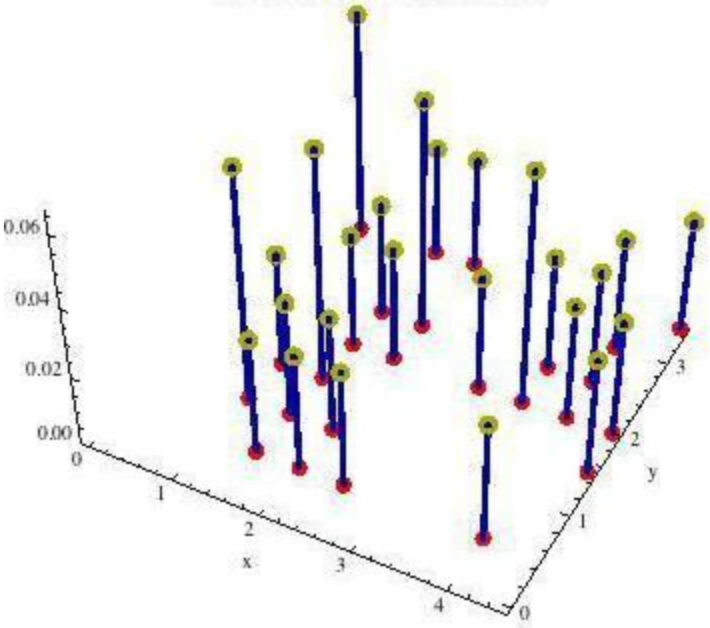


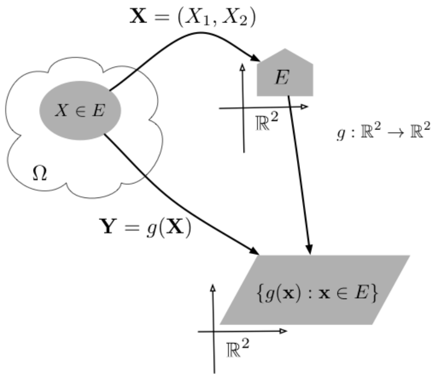
https://twitter.com/KirkDBorne/status/1057689914249232384/photo/1

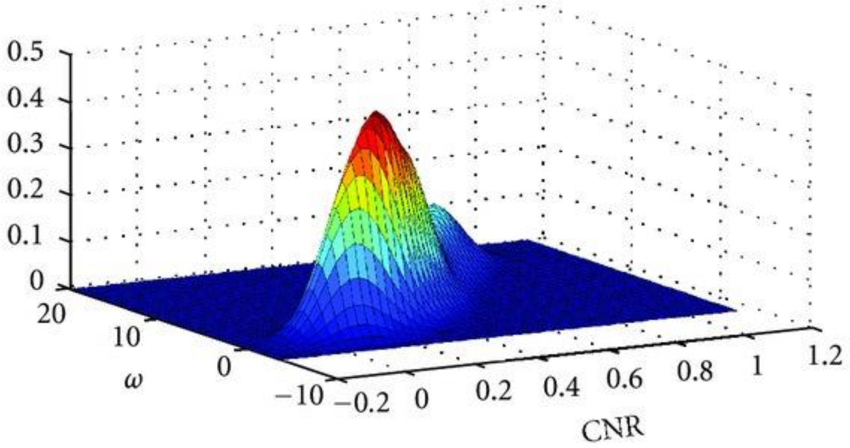
Data → r.v.


Normal Distribution Curve, Fit to a Histogram
x̄ average (mean) of data sample
Frequency of Occurrence
Theoretical Percentages
-1σ to + 1σ = 68% of data
-2σ to + 2σ = 95.4% of data
-3σ to + 3σ = 99.7% of data
Use a z-table to find other percentages
6σ 5σ 4σ 3σ 2σ 1σ 1σ 2σ 3σ 4σ 5σ 6σ
σ = standard deviation of sample
© 2009 dmaictools.com

Random vectors



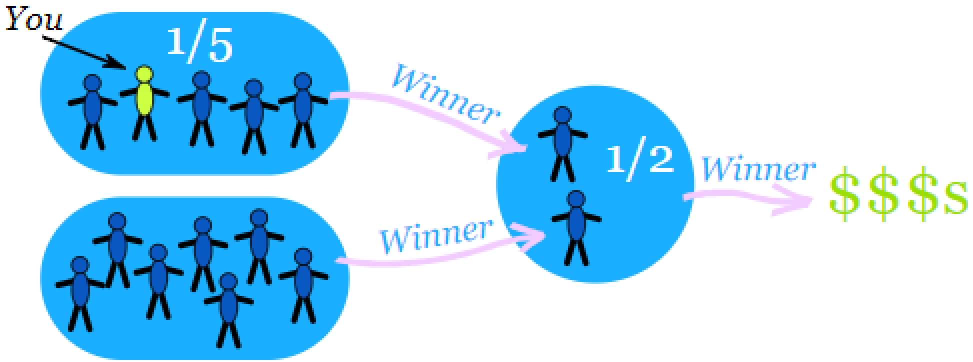




Independent Events

Imagine there are two groups:

- A member of each group gets randomly chosen for the winners circle,
- **then** one of those gets randomly chosen to get the big money prize:



What is your chance of winnning the big prize?

Independent vs. Dependent Events

Using the bag of marbles on the left, what is the probability of pulling a black marble two times in a row? $P(black, black)$

When you put 1st marble back in (*Independent Events*)

$$\frac{2}{10} * \frac{2}{10}$$

$$\frac{1}{5} * \frac{1}{5} \boxed{= \frac{1}{25}}$$

When you KEEP 1st marble (*Dependent Events*)

Independent and Dependent Random Variables

Independent Events

The outcome of one event does not affect the outcome of the other.

If A and B are independent events then the probability of both occurring is

$$P(A \text{ and } B) = P(A) \times P(B)$$

Dependent Events

The outcome of one event affects the outcome of the other.

If A and B are dependent events then the probability of both occurring is

$$P(A \text{ and } B) = P(A) \times P(B | A)$$

Probability of B given A

Independent vs. Dependent Events

Using the bag of marbles on the left, what is the probability of pulling a black marble two times in a row? $P(black, black)$

When you put 1st marble back in (*Independent Events*)

$$\frac{2}{10} * \frac{2}{10}$$

$$\frac{1}{5} * \frac{1}{5} = \frac{1}{25}$$

When you KEEP 1st marble (*Dependent Events*)

$$\frac{2}{10} * \frac{1}{9}$$

$$\frac{1}{5} * \frac{1}{9}$$

$$P(A \text{ and } B) = P(A) \times P(B)$$

$$P(A \text{ and } B) = P(A) \times P(B \mid A)$$

Probability of B given A

Expected Value of a Discrete RV

Suppose the random variable x can take on the n values $x_1, x_2, \ldots, x_n$. Also, suppose the probabilities that these values occur are respectively p_1, $p_2, \ldots, p_n$. Then the expected value of the random variable is:

$$E(x) = x_1 p_1 + x_2 p_2 + \ldots. + x_n p_n$$

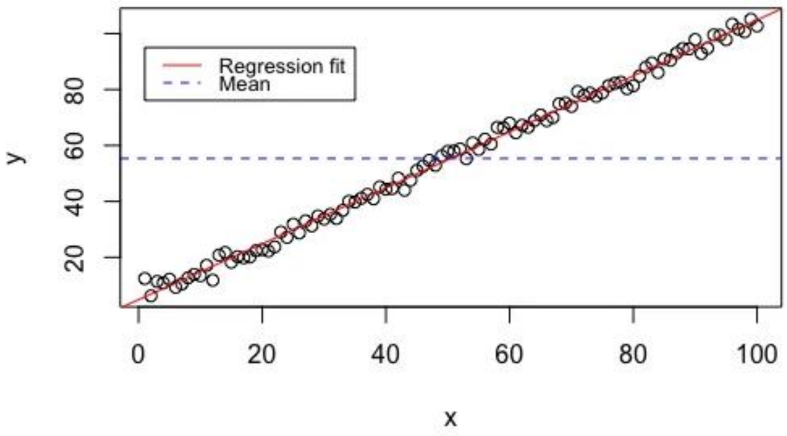
Recall: Unsupervised Learning

Task: Given $X \in \mathcal{X}$, learn $f(X)$.

- f can be
  - Deterministic
  - Probabilistic

Deterministic Models

- $f(x) = a_0 + a_1 x$
- Hypothesize exact relationships
- Suitable when error of prediction is negligible
- Repeated parameter estimation runs give same estimates for each run.



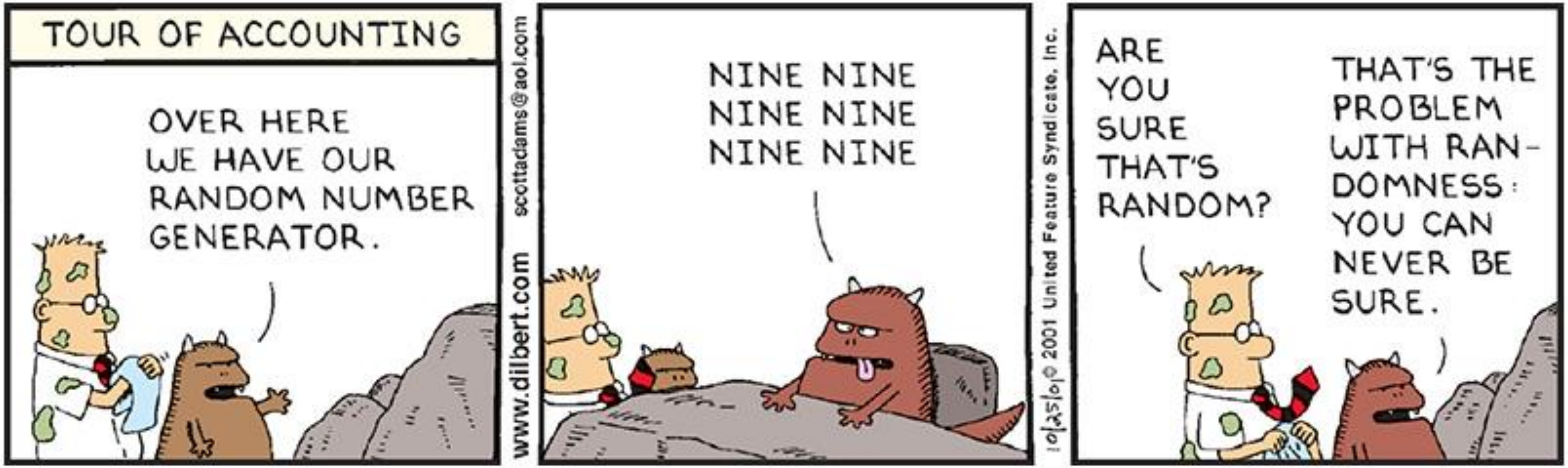
Probabilistic Models

- Help capture uncertainty
- Sales volume (y) is ‘about’ 10 times advertising spending (x)
  - $y = 10\ x + \ \varepsilon$

Sales volume is also due to ‘random’ unseen factors

Probabilistic Generative Model

- `Uniform Random Number Generator - rand()`



Probabilistic Generative Model

Observed data is the ‘realization’ of a probabilistic model

An example

- Consider the estimation of heads probability of a coin tossed *n* times
- Heads probability p
- Data = HHTTHTHHTTT

An example

- Consider the estimation of heads probability of a coin tossed n times
- Heads probability p
- Data = HHTTHTHHTTT
- $L(p) = \Pr(D|p) = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = p^5(1-p)^6$

Maximum Likelihood

$L(p) = p^5(1-p)^6$

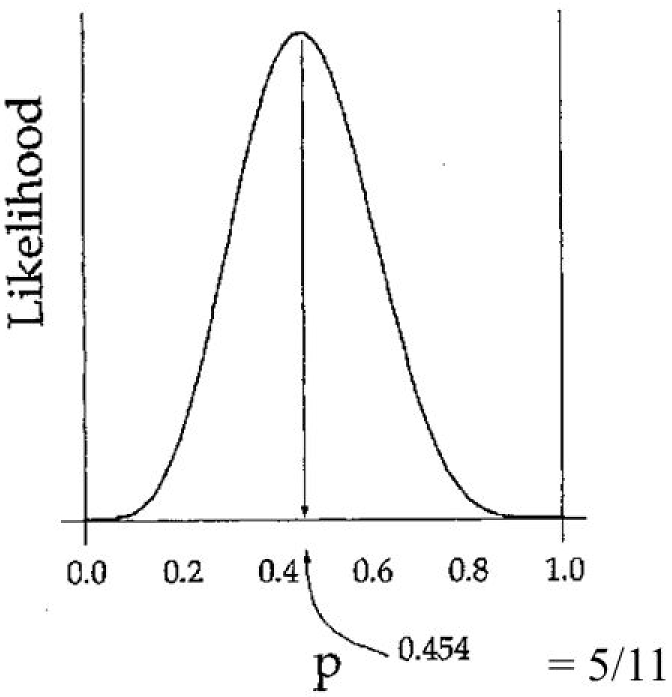
Take the derivative of L with respect to p:

$$\frac{dL}{dp} = 5p^4(1-p)^6 - 6p^5(1-p)^5$$

Equate it to zero and solve:

$$\hat{p} = 5/11$$

$$L(p) = p^5(1-p)^6$$



Log Likelihood

$L(p) = p^5(1-p)^6$

- For computational reasons, we maximise the logarithm

  $\ln L = 5 \ln p + 6 \ln(1-p)$

  with derivative

$$\frac{d(\ln L)}{dp} = \frac{5}{p} - \frac{6}{(1-p)} = 0$$

$\hat{p} = 5/11$

Maximum Likelihood

- The likelihood function is the simultaneous density of the observation, as a function of the model parameters.

$$L(\Theta) = \Pr(Data|\Theta)$$

Maximum Likelihood

- The likelihood function is the simultaneous density of the observation, as a function of the model parameters.

$$L(\Theta) = \Pr(Data|\Theta)$$

- If the observations are independent, we can decompose the term into

$$\Pr(Data \mid \Theta) = \prod_{i=1}^{n} \Pr(X_i \mid \Theta)$$

Estimating Parameters of a Probabilistic Model [Maximum Likelihood Approach]

- Consider the estimation of heads probability of a coin tossed *n* times
- Heads probability p
- Data = HHTTHTHHTTT
- $L(p) = \Pr(D|p) = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = p^5(1-p)^6$

Maximum Likelihood

Take the derivative of L with respect to p:

$$\frac{dL}{dp} = 5p^4(1-p)^6 - 6p^5(1-p)^5$$

Equate it to zero and solve:

$$\hat{p} = 5/11$$

$\ln L = 5 \ln p + 6 \ln(1-p)$
with derivative

$$\frac{d(\ln L)}{dp} = \frac{5}{p} - \frac{6}{(1-p)} = 0$$

$$\hat{p} = 5/11$$

$$Data = \{X_1, X_2, \ldots X_n\}$$

- The likelihood function is the simultaneous density of the observation, as a function of the model parameters.

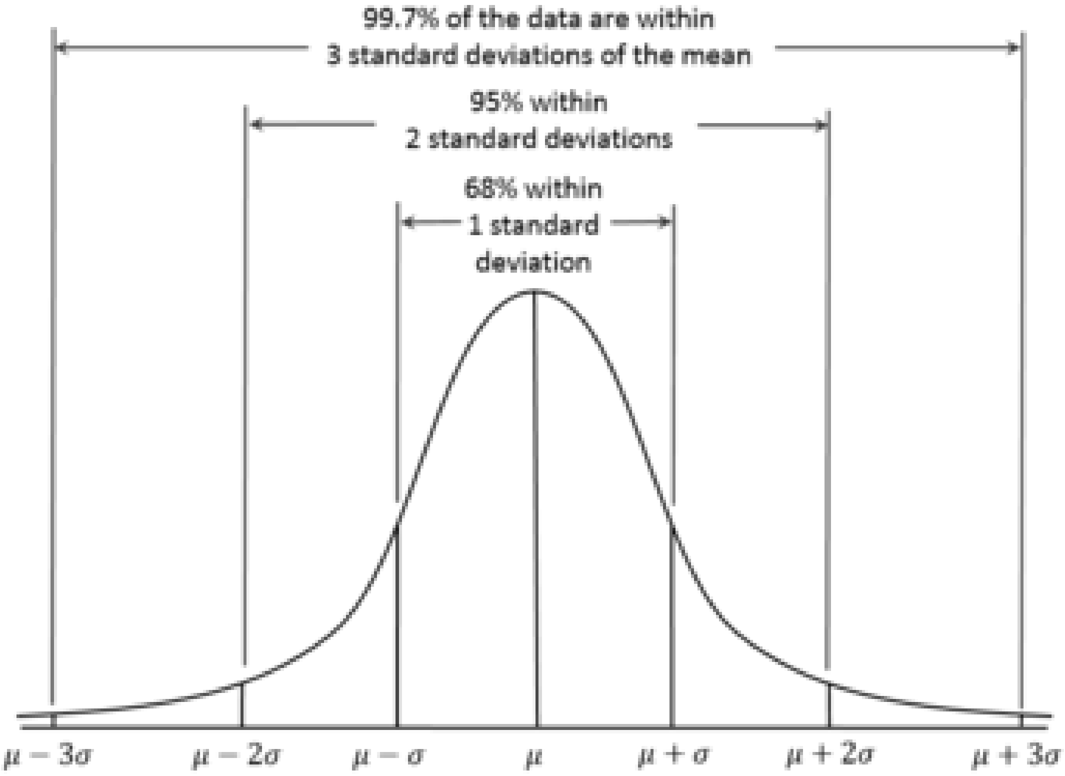
$$L(\Theta) = \Pr(Data|\Theta)$$

- If the observations are independent, we can decompose the term into

$$\Pr(Data \mid \Theta) = \prod_{i=1}^{n} \Pr(X_i \mid \Theta)$$

$$\boxed{\Theta^* = \arg\max_{\Theta} Pr(Data|\Theta)}$$

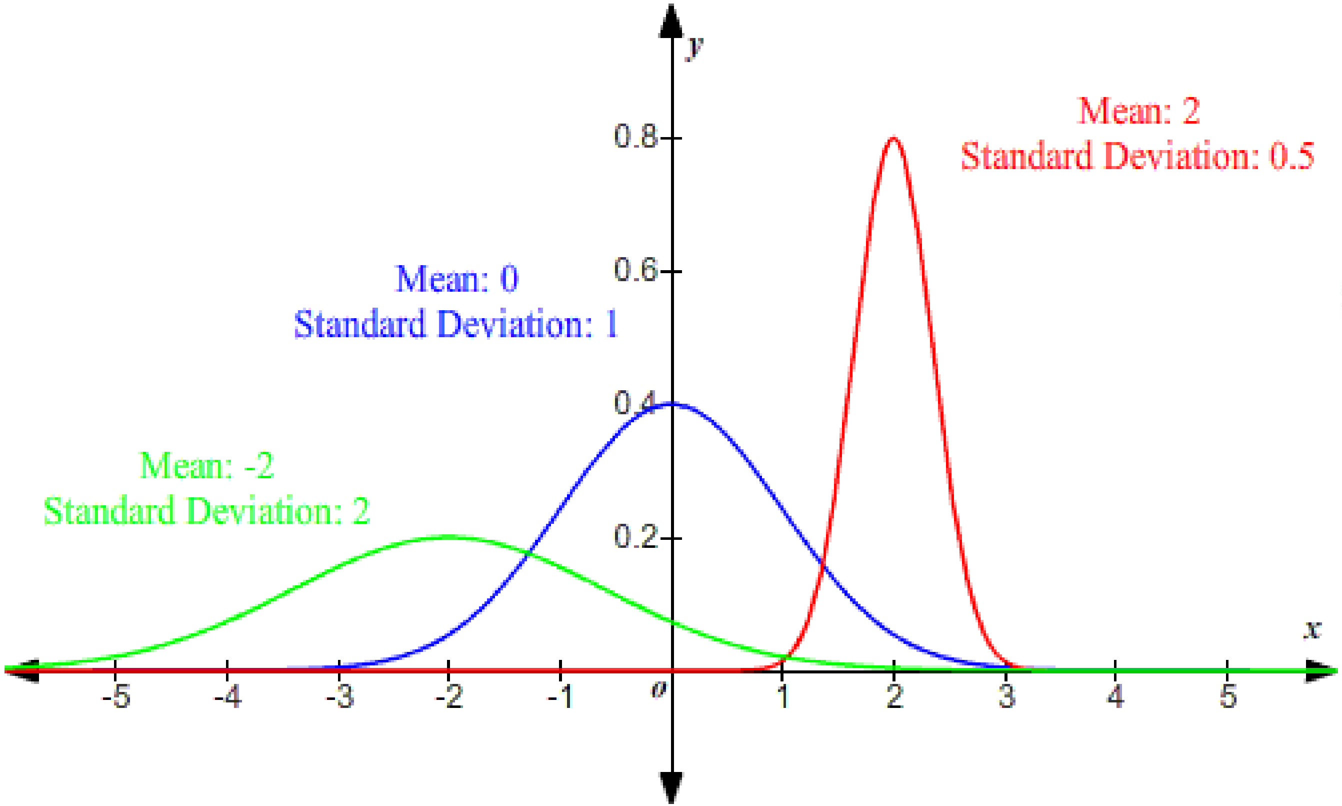
Gaussian Distribution



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

$\mu =$ Mean
$\sigma =$ Standard Deviation
$\pi \approx 3.14159\cdots$
$e \approx 2.71828\cdots$

Gaussian Distribution



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

$\mu =$ Mean

$\sigma =$ Standard Deviation

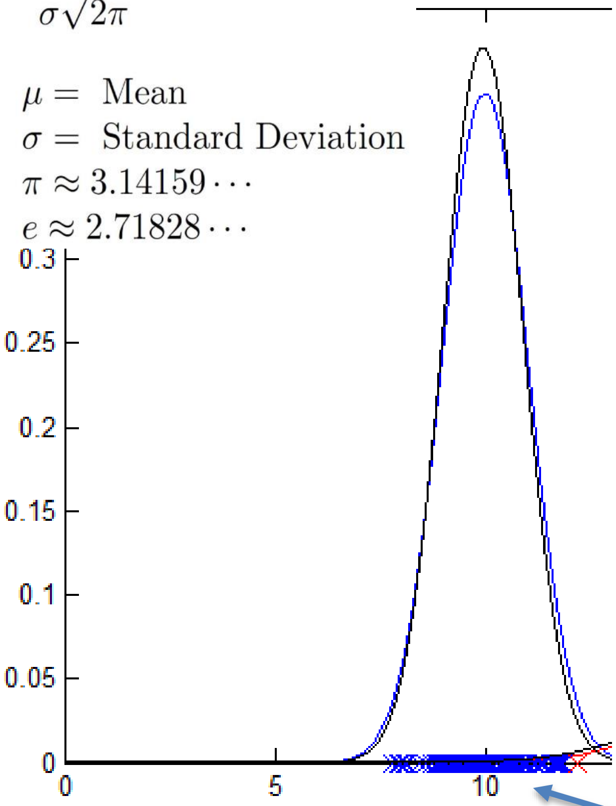
$\pi \approx 3.14159 \cdots$

$e \approx 2.71828 \cdots$

ML Estimation of Gaussian Parameters

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

$\mu =$ Mean
$\sigma =$ Standard Deviation
$\pi \approx 3.14159\cdots$
$e \approx 2.71828\cdots$



$$Data = \{X_1, X_2, \ldots X_n\}$$

$$\Theta =$$

$$Data = \{X_1, X_2, \ldots X_n\}$$

- The likelihood function is the simultaneous density of the observation, as a function of the model parameters.

$$L(\Theta) = \Pr(Data|\Theta)$$

- If the observations are independent, we can decompose the term into

$$\Pr(Data \mid \Theta) = \prod_{i=1}^{n} \Pr(X_i \mid \Theta)$$

$$\Theta^* = \arg\max_{\Theta} Pr(Data|\Theta)$$

Maximum Likelihood Solution

- Maximizing w.r.t. the mean gives the *sample mean*

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$$

- Maximizing w.r.t covariance gives the *sample covariance*

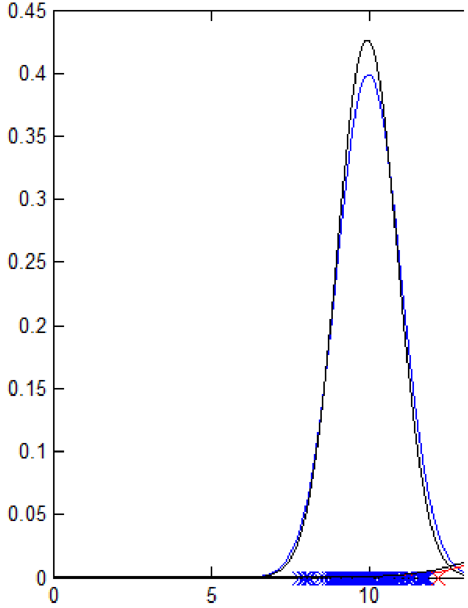
$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

Note: if N is small you want to divide by N-1 when computing sample covariance to get an unbiased estimate.

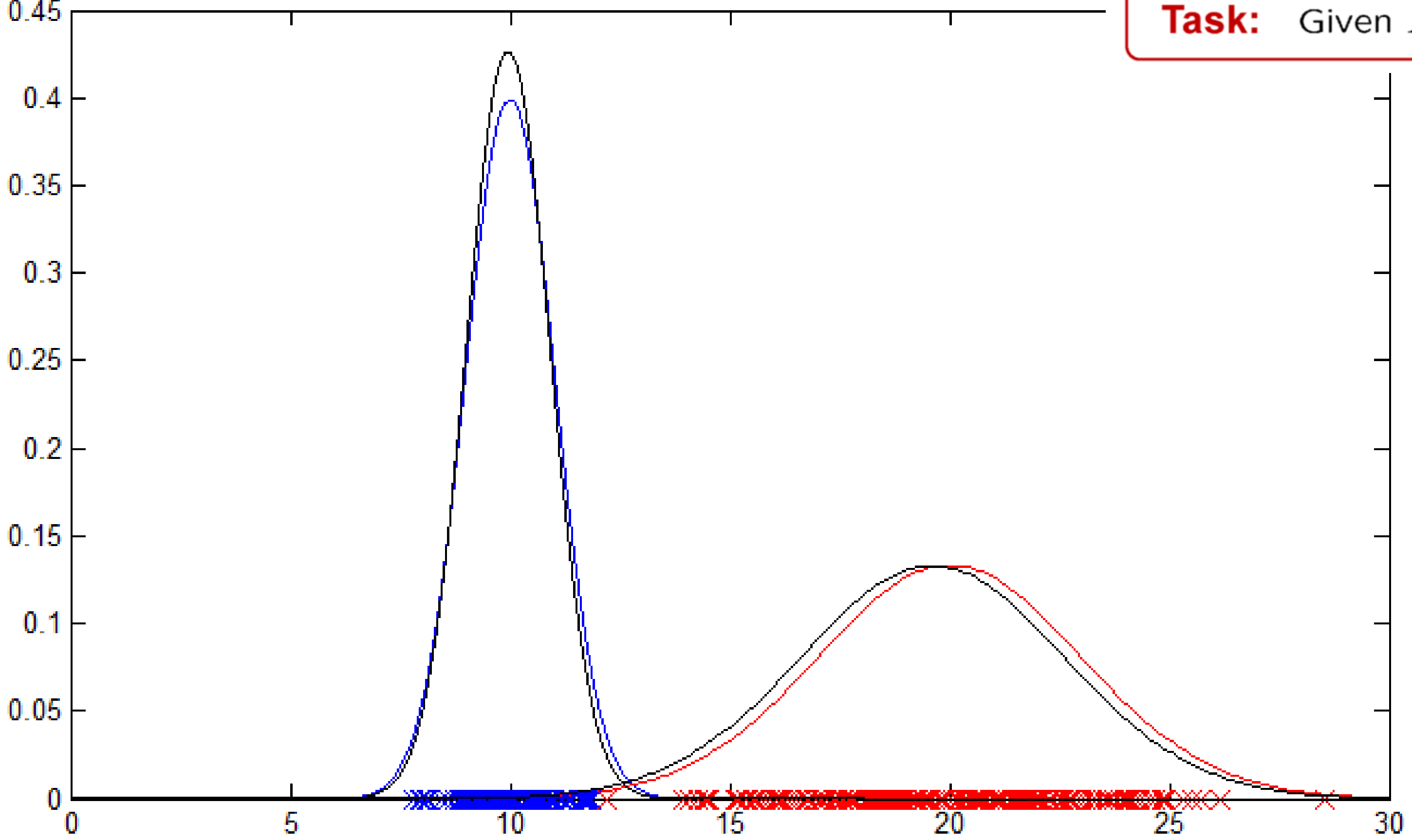
Bishop, 2003

- Note: Knowing the parameters allows us to compute probability (density) of data
- Previously (k-means): Obtain cluster centers from cluster memberships
- Alternative: Obtain from probabilistic modelling of 'cluster data density'

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



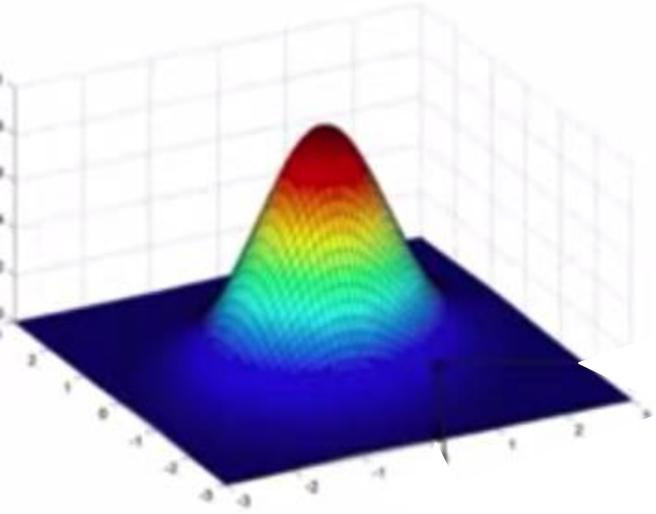
Data Probability Density is often Multi-modal



Task: Given $X \in \mathcal{X}$, learn $f(X)$.

Multivariate Gaussian

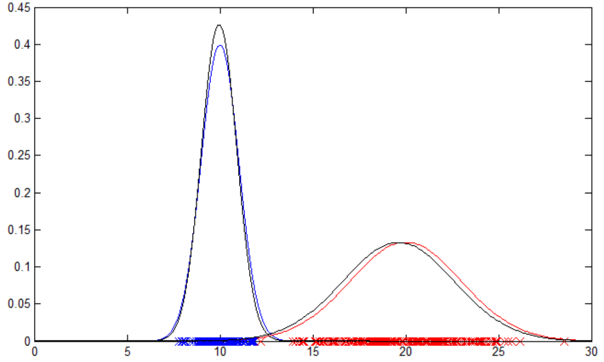
$$\mathcal{N}(\underline{x}\ ;\ \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(\underline{x} - \underline{\mu})\Sigma^{-1}(\underline{x} - \underline{\mu})^T \right\}$$



μ = length-d row vector
Σ = d x d matrix

$|\Sigma|$ = matrix determinant

Mixture of Gaussians



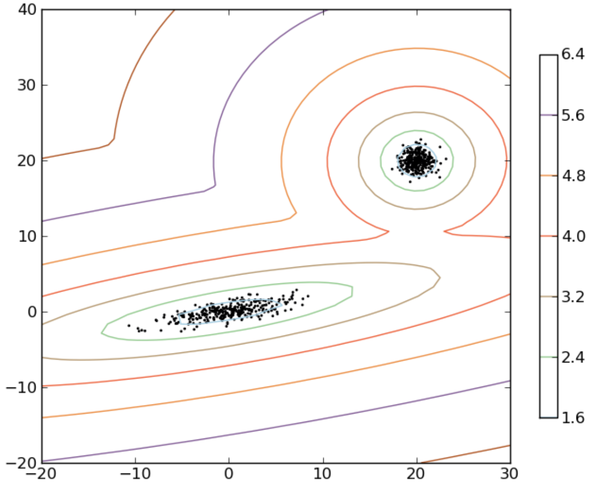
- Convex Combination of Distributions

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Normalization and positivity require

$$\sum_{k=1}^{K} \pi_k = 1 \qquad 0 \leqslant \pi_k \leqslant 1$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k)$$

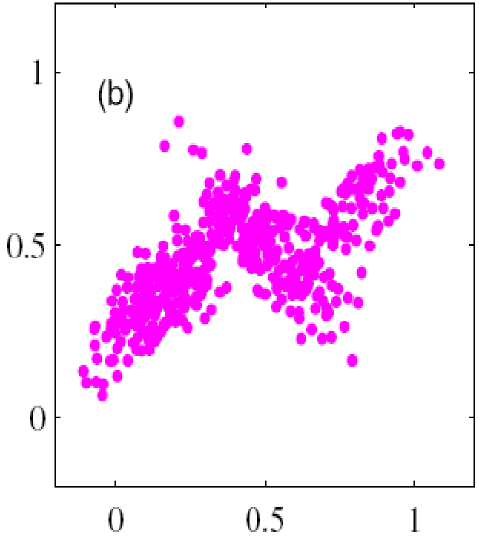


Bishop, 2003

MLE of Mixture Parameters

- However, MLE of mixture parameters is HARD!
- Joint distribution:

$$p(\mathbf{X}|\pi, \mu, \Sigma) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \Sigma_k)$$



MLE of Mixture Parameters

- However, MLE of mixture parameters is HARD!
- Joint distribution:

$$p(\mathbf{X}|\pi, \mu, \Sigma) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Log likelihood

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \underbrace{\left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right\}}_{\text{Uh-oh, log of a sum}}$$