




Inspire…Educate…Transform.

Probability

Dr. K. V. Dakshinamurthy

President, International School of Engineering

Stat versus prob

- Statistics measures tendencies from data
- Probability allows you to predict the data from tendencies
  - http://www.ft.com/intl/cms/s/0/2299fd16-9701-11dc-b2da-0000779fd2ac.html

FROM DATA

Single column

- Mean, Median, Mode
  - The arithmetic mean is the usual average
  - The median is the middle value
  - The mode is the number that is repeated more often than any other

Single column

- Standard Deviation, Variance
  - Spread of data around the mean

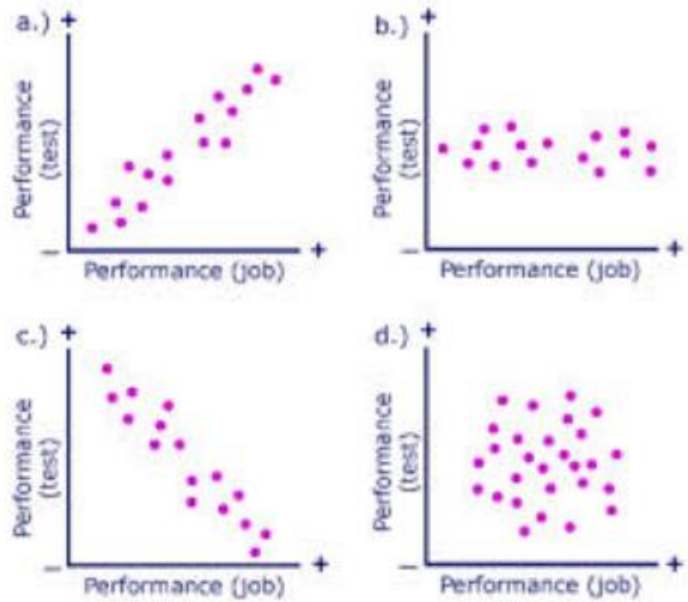
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(\bar{x} - x_i)^2}{n}}$$

Multiple columns

- Covariance, Correlation

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}$$

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$



Entire data: Variance-Covariance matrix

$$\begin{bmatrix} {\sigma_x}^2 & \rho_{xy} & \rho_{xz} \\ \rho_{yx} & {\sigma_y}^2 & \rho_{yz} \\ \rho_{zx} & \rho_{zy} & {\sigma_z}^2 \end{bmatrix}$$

A square matrix formed with the variance and covariance

FROM TENDENCIES

Empirical probability

- Who is more likely to hit 6 sixes in an over in a ODI: Kohli or Rohit?

Full name Rohit Gurunath Sharma

Born April 30, 1987, Bansod, Nagpur, Maharashtra

Current age 27 years 297 days

Major teams India, Deccan Chargers, India A, India Green, India Under-19s, Mumbai, Mumbai Cricket Association President's XI, Mumbai Indians, Mumbai Under-19s

Playing role Batsman

Batting style Right-hand bat

Bowling style Right-arm offbreak

Like 9.9k

Batting and fielding averages

	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	4s	6s	Ct	St
Tests	10	18	2	662	177	41.37	1285	51.51	2	2	71	11	10	0
ODIs	128	122	21	3905	264	38.66	4790	81.52	6	23	325	79	43	0
T20Is	42	35	11	739	79*	30.79	586	126.10	0	7	61	27	19	0
First-class	68	105	10	5464	309*	57.51			18	22			51	0
List A	195	185	28	6077	264	38.70			9	36			66	0
Twenty20	186	175	33	4613	109*	32.48	3547	130.05	2	31	381	193	81	0

Full name Virat Kohli

Born November 5, 1988, Delhi

Current age 26 years 108 days

Major teams India, Delhi, India Red, India Under-19s, Royal Challengers Bangalore

Playing role Middle-order batsman

Batting style Right-hand bat

Bowling style Right-arm medium

Like 66k

Batting and fielding averages

	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	4s	6s	Ct	St
Tests	33	59	4	2547	169	46.30	4798	53.08	10	10	301	8	30	0
ODIs	151	143	21	6339	183	51.95	7037	90.08	22	33	605	60	70	0
T20Is	28	26	5	972	78*	46.28	738	131.70	0	9	105	20	13	0
First-class	64	105	11	4735	197	50.37	8565	55.28	17	18	608	21	58	0
List A	185	176	24	7781	183	51.19	8573	90.76	26	41	769	84	88	0
Twenty20	157	146	22	4298	99	34.66	3344	128.52	0	29	416	127	69	0

Classical probability

- What is the probability of getting a head when I toss a coin?
- What is the probability of getting 1 when I roll a dice of 6 possibilities?
- You derive the answer using thought experiments

Probability is all about counting carefully!

- There are three cards (A, B, C). A has white on both sides. B has black on both sides. C has white on one and black on 1. They are placed one on top of the other and the top is black.
  - If the reverse is white, you get Rs. 100 and reverse is black, you lose Rs. 60. Will you take the bet?

The possibilities

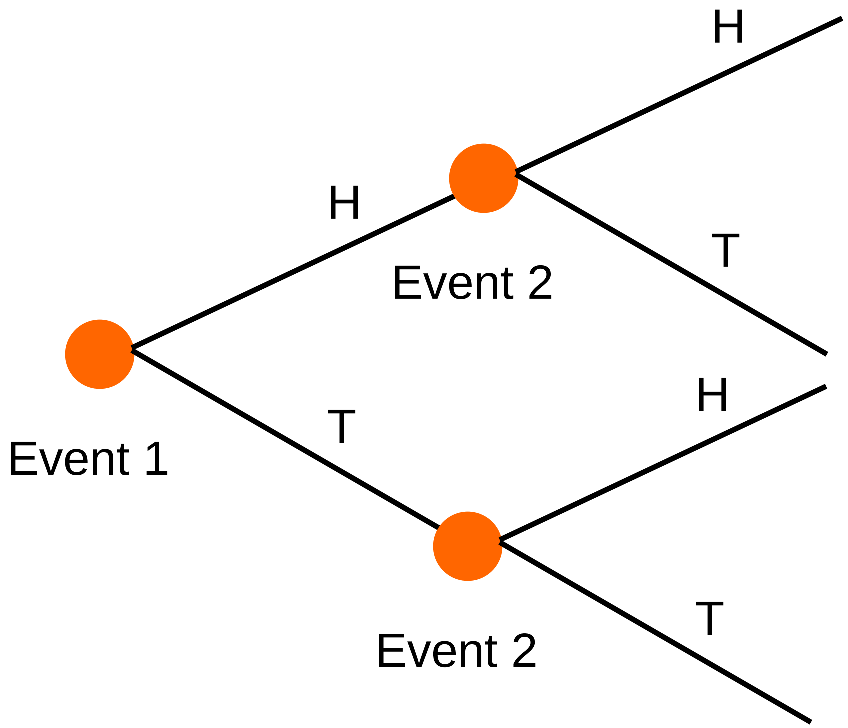
- A top black could be

Side 1 of B	Side 2 of B	Side 1 of C
You lose	You lose	You win

If you play 3 times, you lose 120 and gain 100. Don't take it.

Trees are a good way to count

- What is the probability of getting 2 heads in a row



Another problem

- You have 8 coins in a bag. 3 of them are unfair in that they have a 60% chance of coming up heads. Rest are fair. You randomly choose one coin from the bag and flip twice. What is the probability of getting 2 heads

Approach 2: the other end

- If I toss 5 times, what is the probability that I get at least one head
  - The tree gets complex
  - How many possibilities are there: 2 per each toss. So, for 5 tosses = 2X2X2X2X2=32
  - How many ways I can have all tails = 1
  - So, in 31 of 32, there will be at least one head.

Tools to count: Permutations and combinations

- Permutation is complex sounding. It is. Order is importation
  - 10 athletes are competing in Olympics. How many different ways the gold, silver and bronze can be won?
  - $10 * 9 * 8 = \frac{10!}{(10-3)!} = P_3^{10}$

Permutations and combinations

- Combination is simple. It is OK if order is relaxed
  - In the same problem as before, how many ways I can pick top three guys
  - $C_3^{10} = \frac{10!}{3!(10-3)!} = \frac{10*9*8}{3*2*1} = 120$

A tough one to count

- There are 365 different days in calendar. What is the probability that any two have the same birthday in a group of 25
  - How many groups are possible (P or C)
    - $C_2^{25} = 300$
  - What is the possibility that a group does not have same birthday = $\frac{364}{365}$
    - What is the probability that not even one have the same birth day = $(\frac{364}{365})^{300} = 0.44$

PROBABILITY PROVIDES THE FRAMEWORK FOR STATISTICS

Random variables

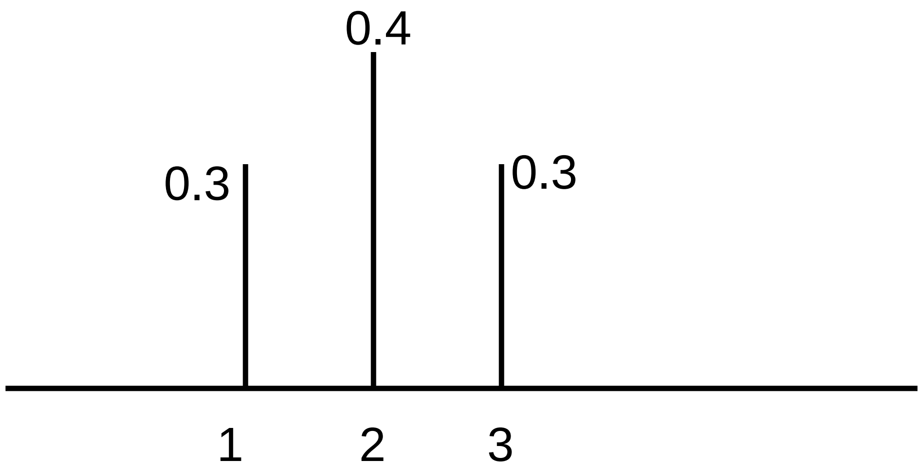
- A function that takes different values with different probabilities

Y =
1 (0.3 probability)
2 (0.4 probability)
3 (0.3 probability)

This is called discrete random variable as it can only take discrete values.

Aha! All categorical variables are discrete random variables.

Plotting the discrete random variables



This is called probability mass function

Discrete Random variable

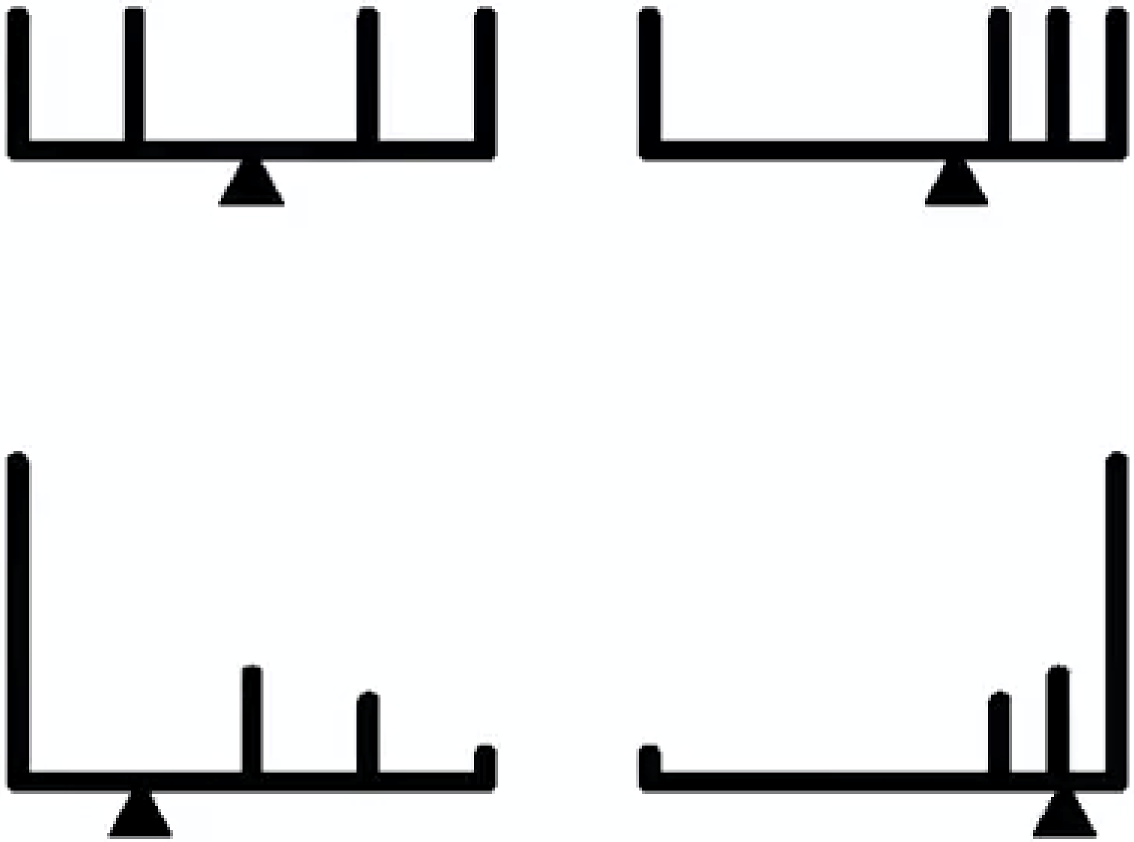
- Sum of all probabilities is 1

$$\sum p_i = 1$$

- Probability weighted average is called the expectation (arithmetic mean)

$$p_1 x_1 + p_2 x_2 + \cdots = \sum p_i x_i$$

Expectation is like the center of mass



What is the expectation of

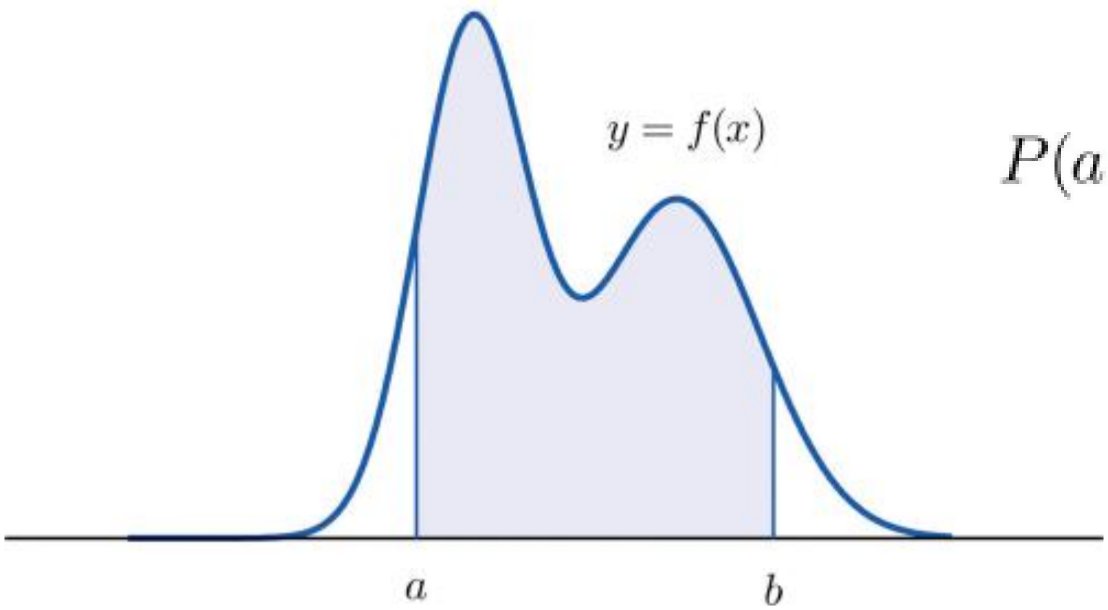
- A dice with six outcomes
- A coin toss
- Expectation need not be one of the possible values

Continuous Random variable

- Function can take any value

- All numeric attributes are continuous random variables

Probability distribution function



$$P(a < X < b) = \text{area of shaded region}$$

Some interesting facts

- Probability that the function takes a value a is 0
- All the formulae for discrete are valid for continuous
  - $\sum$ *must be replaced by* $\int$ *sign*.
- Any function whose area is 1 in a range, never takes a negative value nor is discontinuous can be a probability distribution function.

Properties of expectation

- $E[aX + b] = aE[X] + b$
  - If you multiply a random variable by a constant, you multiply the expectation
  - If you add a constant to a random variable, you add the constant to the expectation
  - Expectation is a linear operator
  - Let us verify with R!

Variance

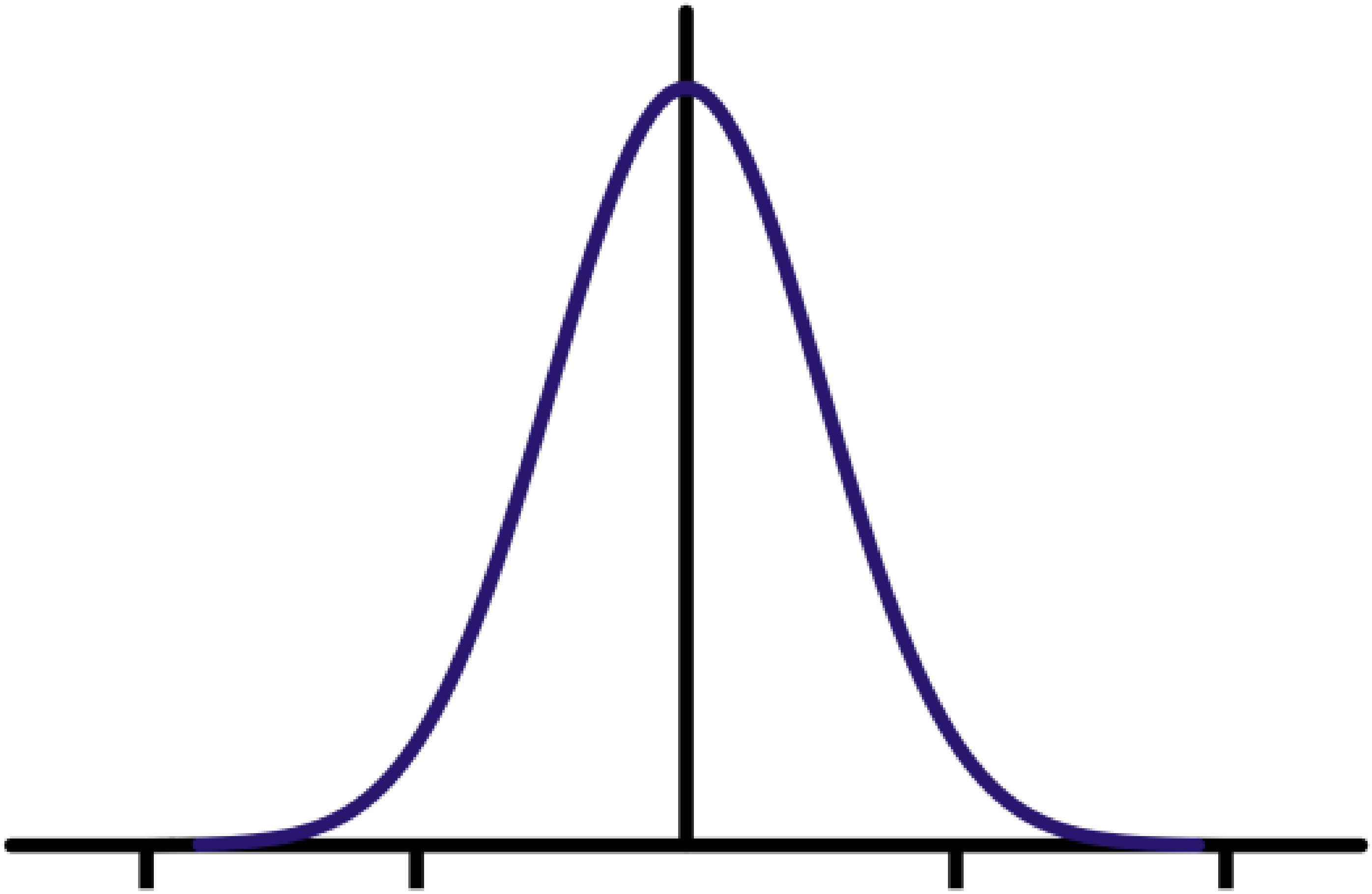
$$Var(x) = E((x - E(x))^2)$$

$$= E(x^2 - 2xE(x) + E(x)^2)$$
$$= E(x^2) - 2E(x)E(x) + E(x)^2$$
$$= E(x^2) - E(x)^2$$

Variance relations

- $V(ax) = a^2 V(x)$
- $V(x+b) = V(x)$
- $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$

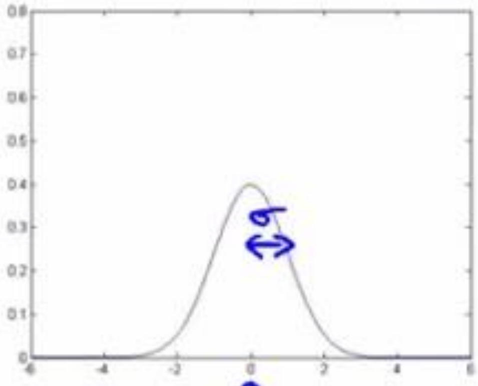
GAUSSIAN EXPLAINED



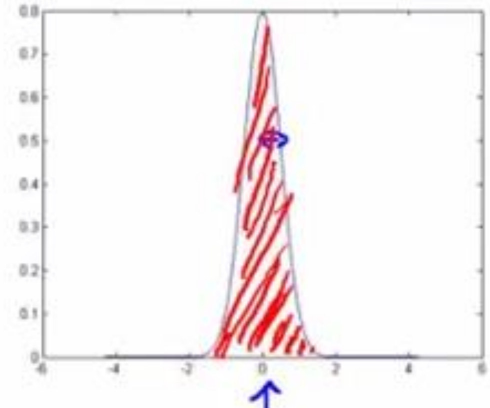
$$PDF\ of\ Bell\ curve = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)}$$

Gaussian distribution example

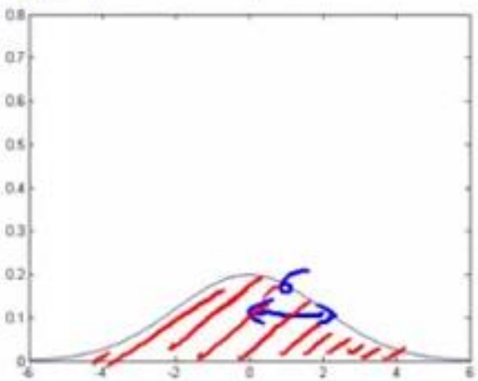
$\mu = 0, \sigma = 1$



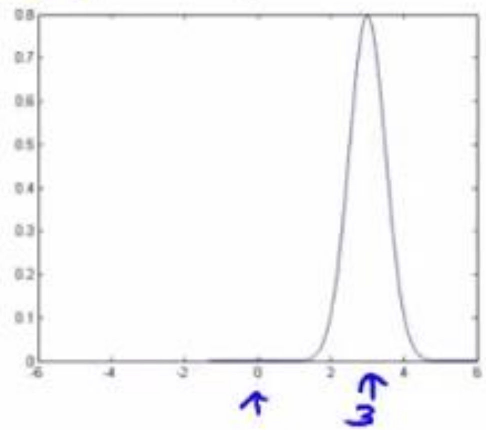
$\mu = 0, \sigma = \underline{0.5}$ $\qquad \sigma^2 = 0.25$



$\mu = 0, \sigma = 2$



$\mu = 3, \sigma = 0.5$

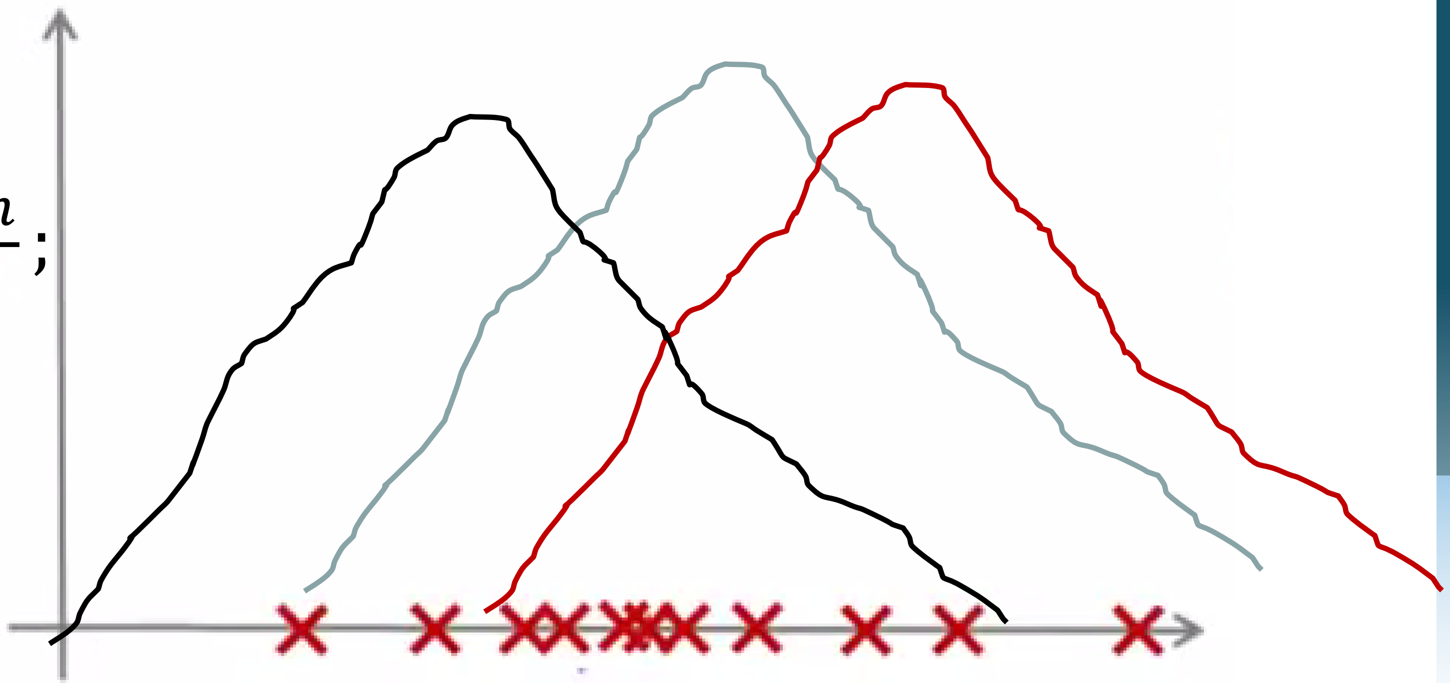


Parameter estimation

Dataset: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ $\quad x^{(i)} \in \mathbb{R}$

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n};$$

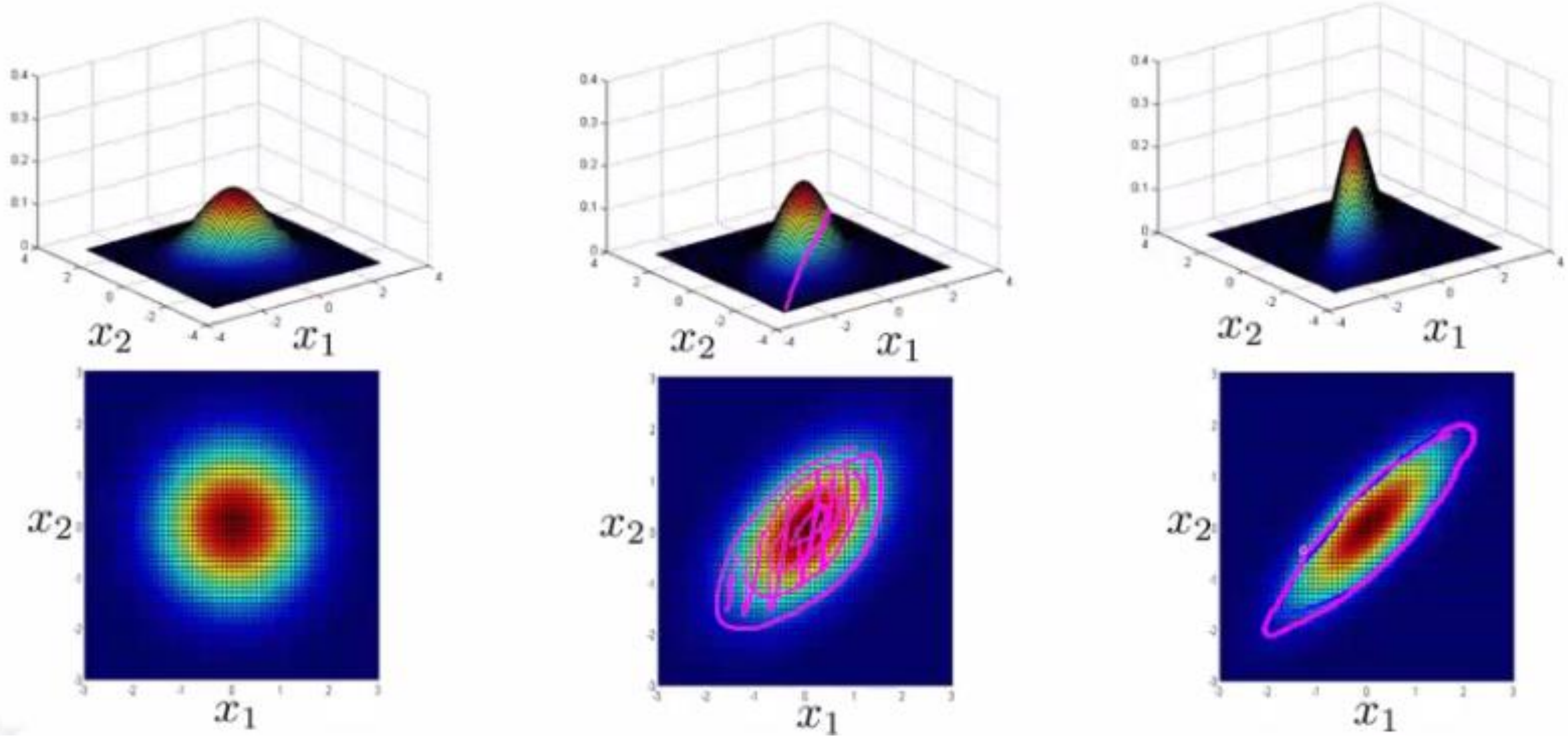
$$\sigma^2 = \frac{\sum_1^n (x-\mu)^2}{n}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



Multi variate Gaussian

$$Bell\ curve\ in\ multivariate\ case =$$

$$2\pi^{\frac{-N}{2}}|\Sigma|^{\frac{-1}{2}}e^{(\frac{-1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu))}$$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

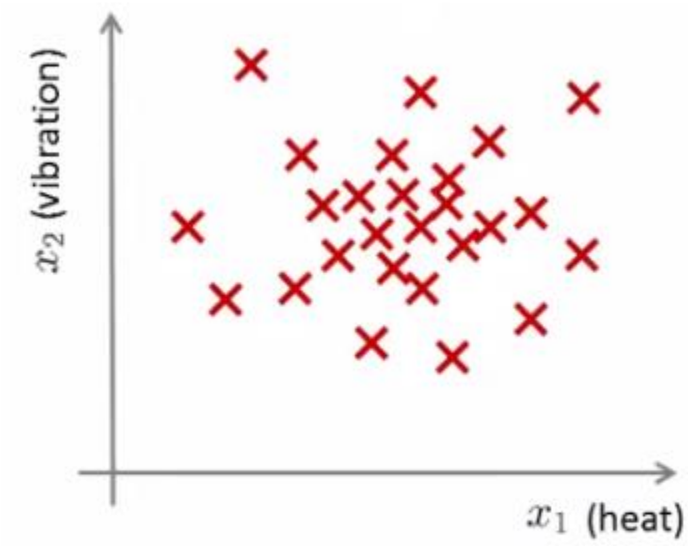
APPLICATIONS

Six sigma

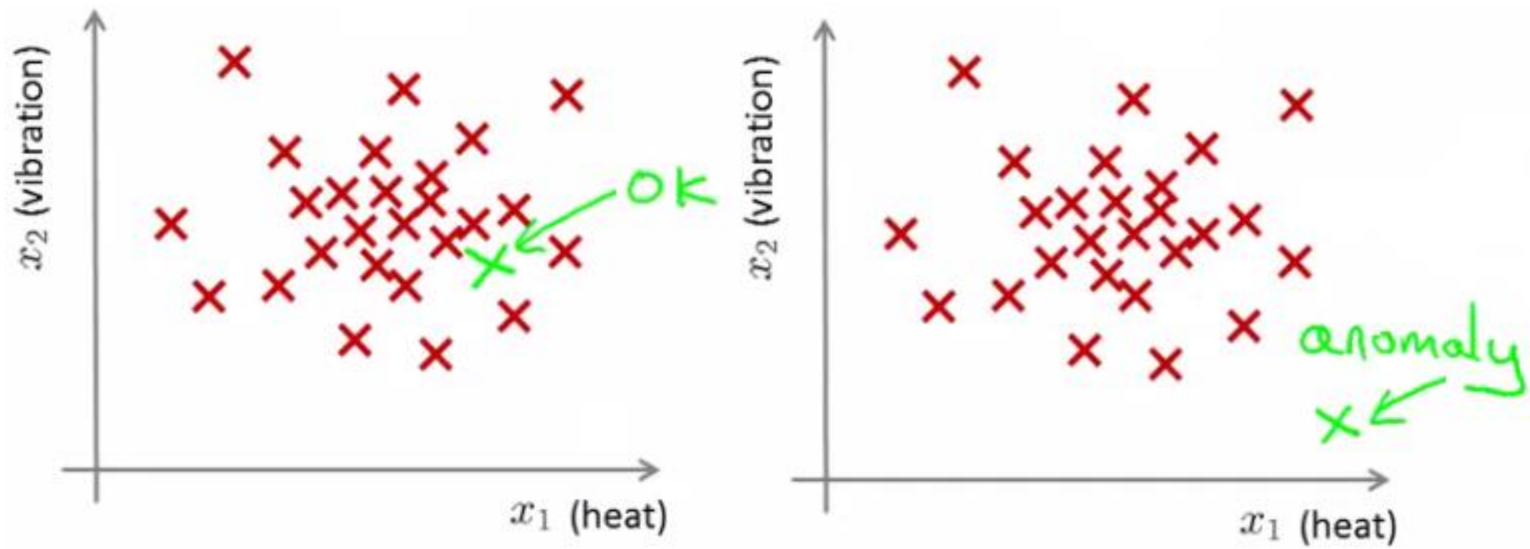
- Normal distribution has a property 69-95-99
- Productivity
  - For less complex components, more defects
  - For more complex components, high defects
- Quality stringency in God's factory

Anomaly detection

- Imagine you're an aircraft engine manufacturer. As engines roll off your assembly line you're doing QA
- Measure some features from engines (e.g. heat generated and vibration)
- You now have a dataset of x^1 to x^m (i.e. *m* engines were tested) and say we plot that dataset



Expectation of a model



Anomaly detection algorithm

1. Choose features x_i that you think might be indicative of anomalous examples.
2. Fit parameters $\mu_1, \ldots, \mu_n, \sigma_1^2, \ldots, \sigma_n^2$

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$$

$$\sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2$$

3. Given new example x, compute $p(x)$:

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

Anomaly if $p(x) < \varepsilon$

Aircraft engines motivating example

10000 good (normal) engines

20 flawed engines (anomalous) 2-50

y=1

Training set: 6000 good engines

CV: 2000 good engines ($y = 0$), 10 anomalous ($y = 1$)

Test: 2000 good engines ($y = 0$), 10 anomalous ($y = 1$)

Algorithm evaluation

→ Fit model $p(x)$ on training set $\{x^{(1)}, \ldots, x^{(m)}\}$ $(x_{test}^{(i)}, y_{test}^{(i)})$

→ On a cross validation/test example x, predict

$$y = \begin{cases} 1 & \text{if } p(x) < \varepsilon \text{ (anomaly)} \\ 0 & \text{if } p(x) \geq \varepsilon \text{ (normal)} \end{cases}$$

$y = 0$

Possible evaluation metrics:

→ - True positive, false positive, false negative, true negative

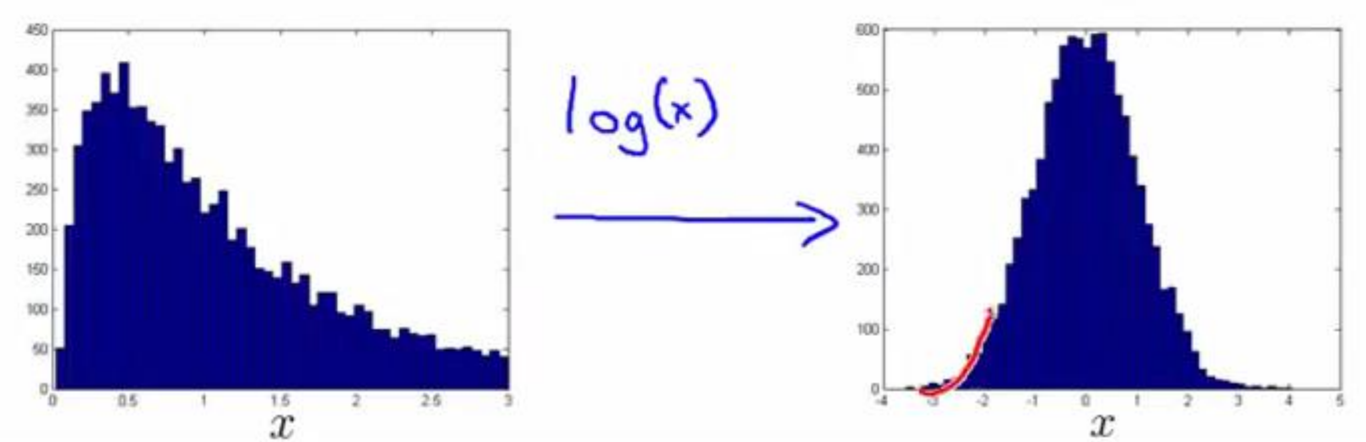
→ - Precision/Recall

→ - F_1-score

Can also use cross validation set to choose parameter ε

Features engineering

When a feature is not normal



Log(x+c) or x^1/c

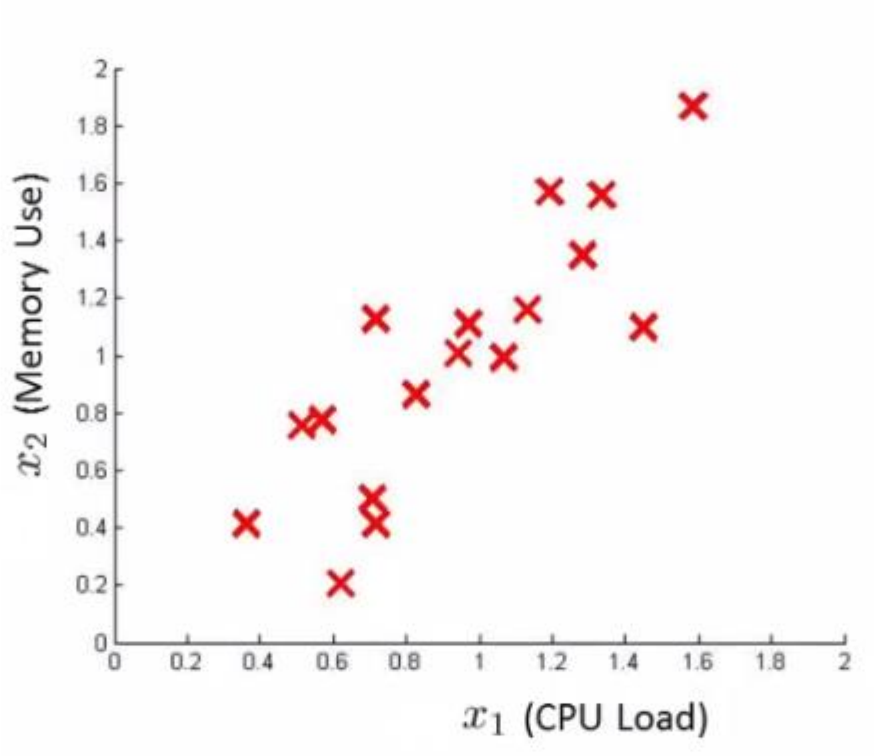
Play with c to get a normal histogram

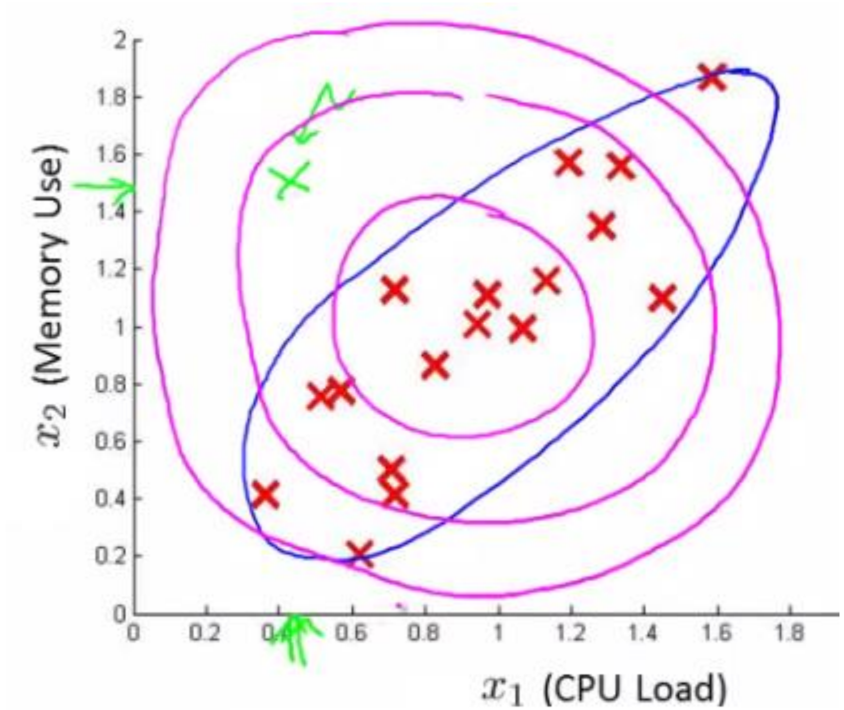
When both bad and good samples have same p

Can you define a new attribute with anomalous value

Multivariate normal distribution

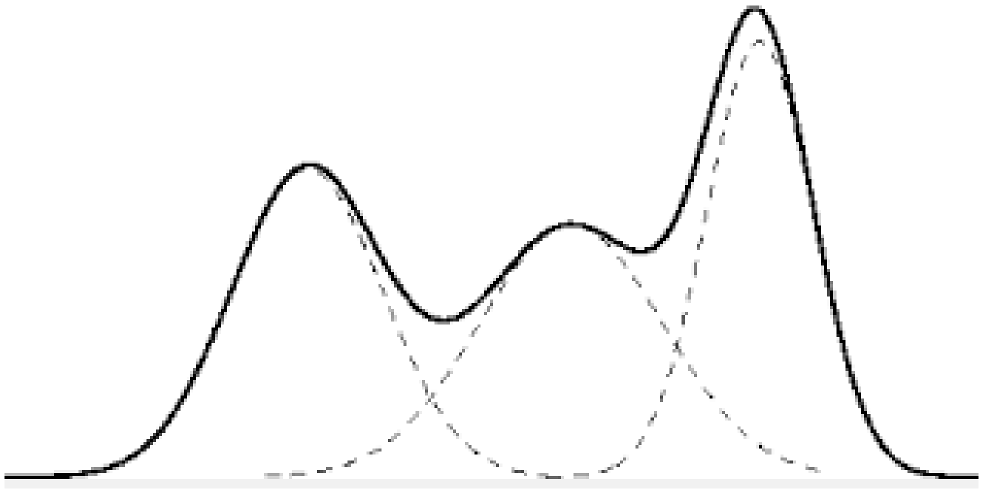
- When features are not independent

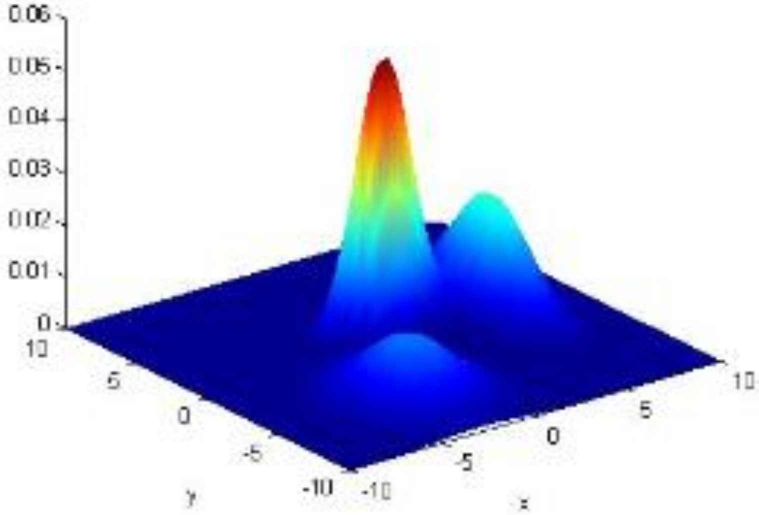




Gaussian mixtures and clusters

Clusters





Gaussian mixture model

The probability given in a mixture of *K* Gaussians is:

$$p(x) = \sum_{j=1}^{K} w_j \cdot N(x \mid \mu_j, \Sigma_j)$$

where w_j is the prior probability (weight) of the jth Gaussian.

$$\sum_{j=1}^{K} w_j = 1 \qquad \text{and} \qquad 0 \le w_j \le 1$$

Expectation maximization

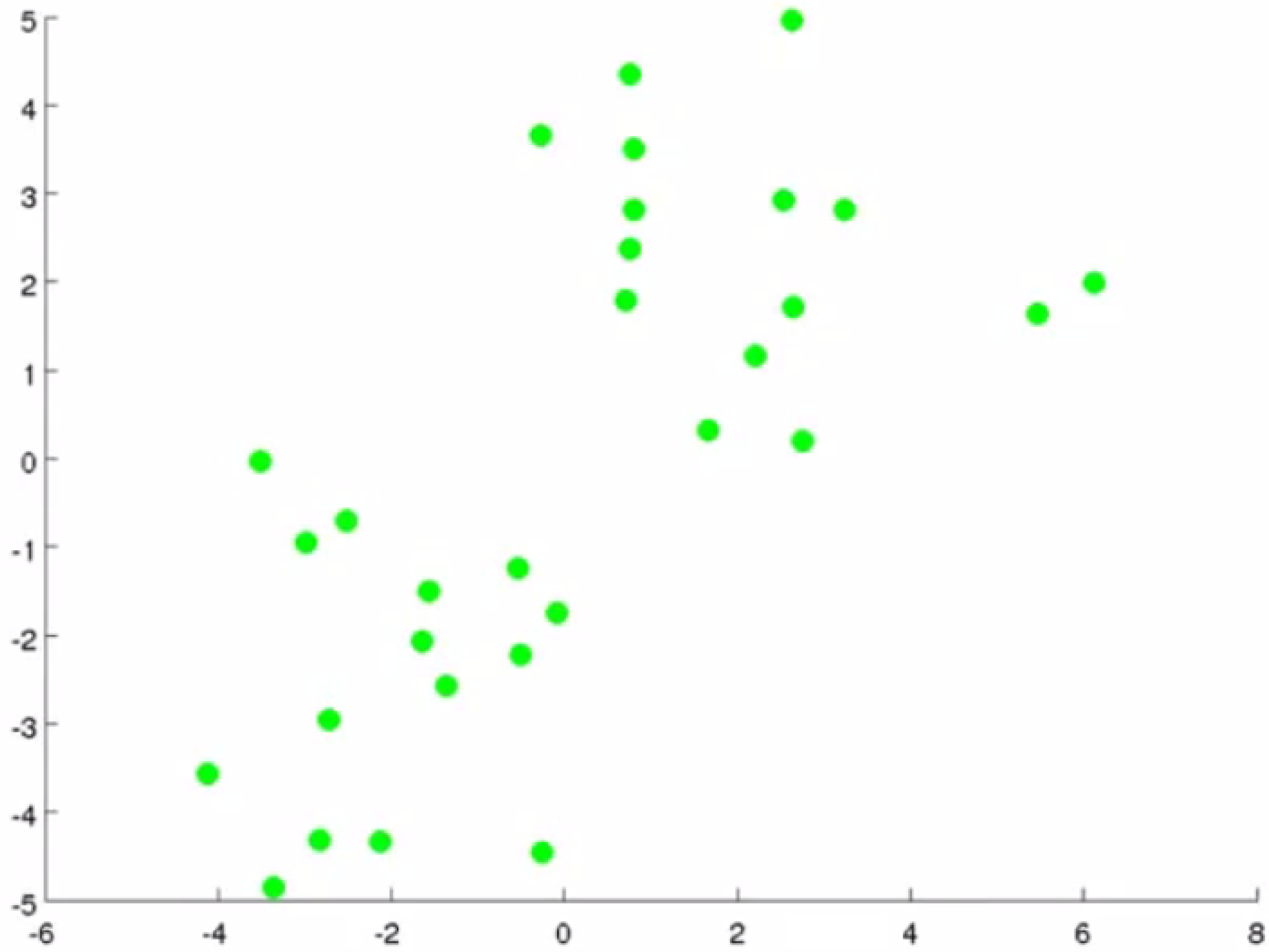
- A point belongs to a center in proportion to its Gaussian probability

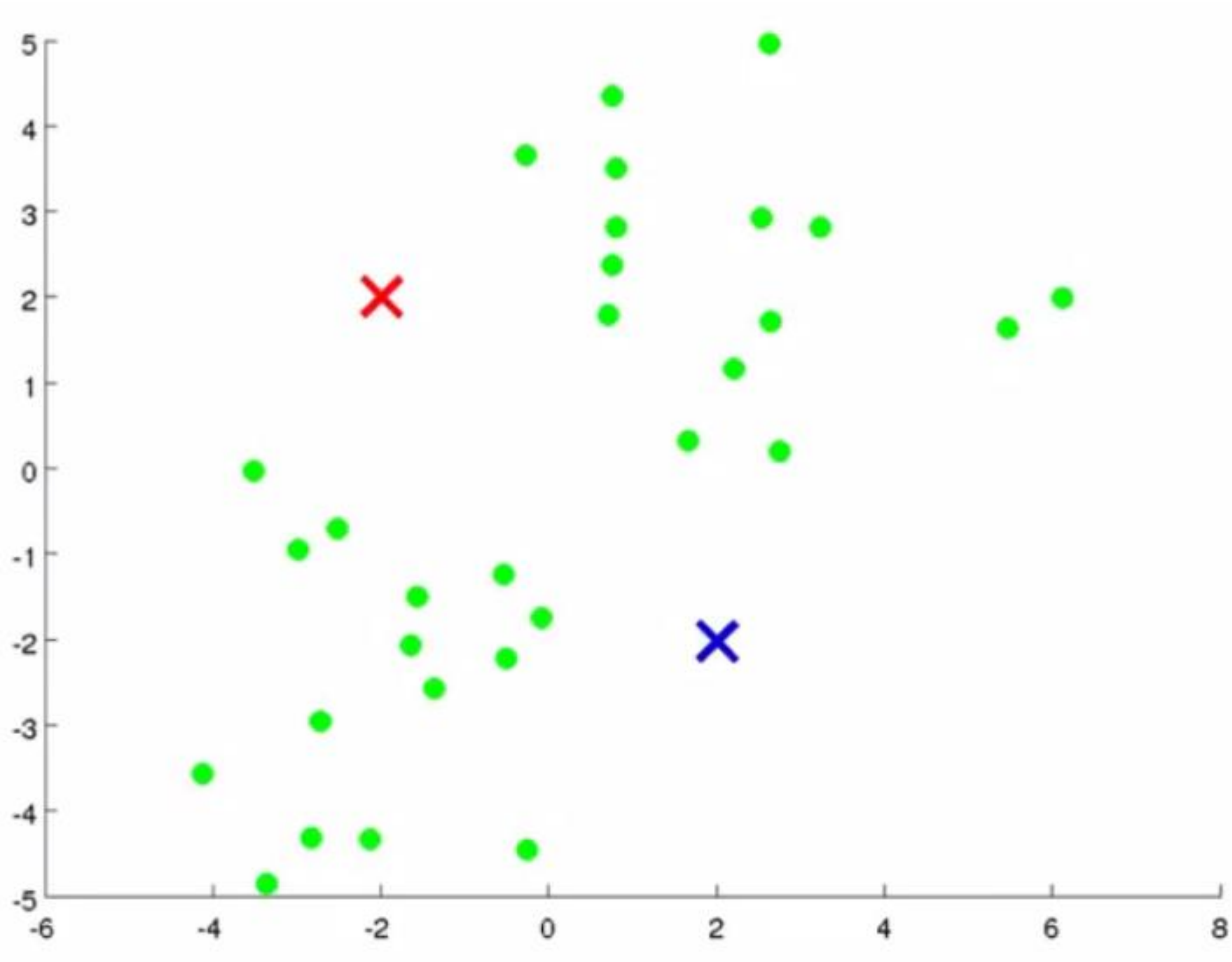
- Cluster centers are optimized to maximize the probability of all points (so they do not move as drastically as K-means)

- $P(x) = \sum_{i=1}^{k} p(c = i).p(x|c = i)$

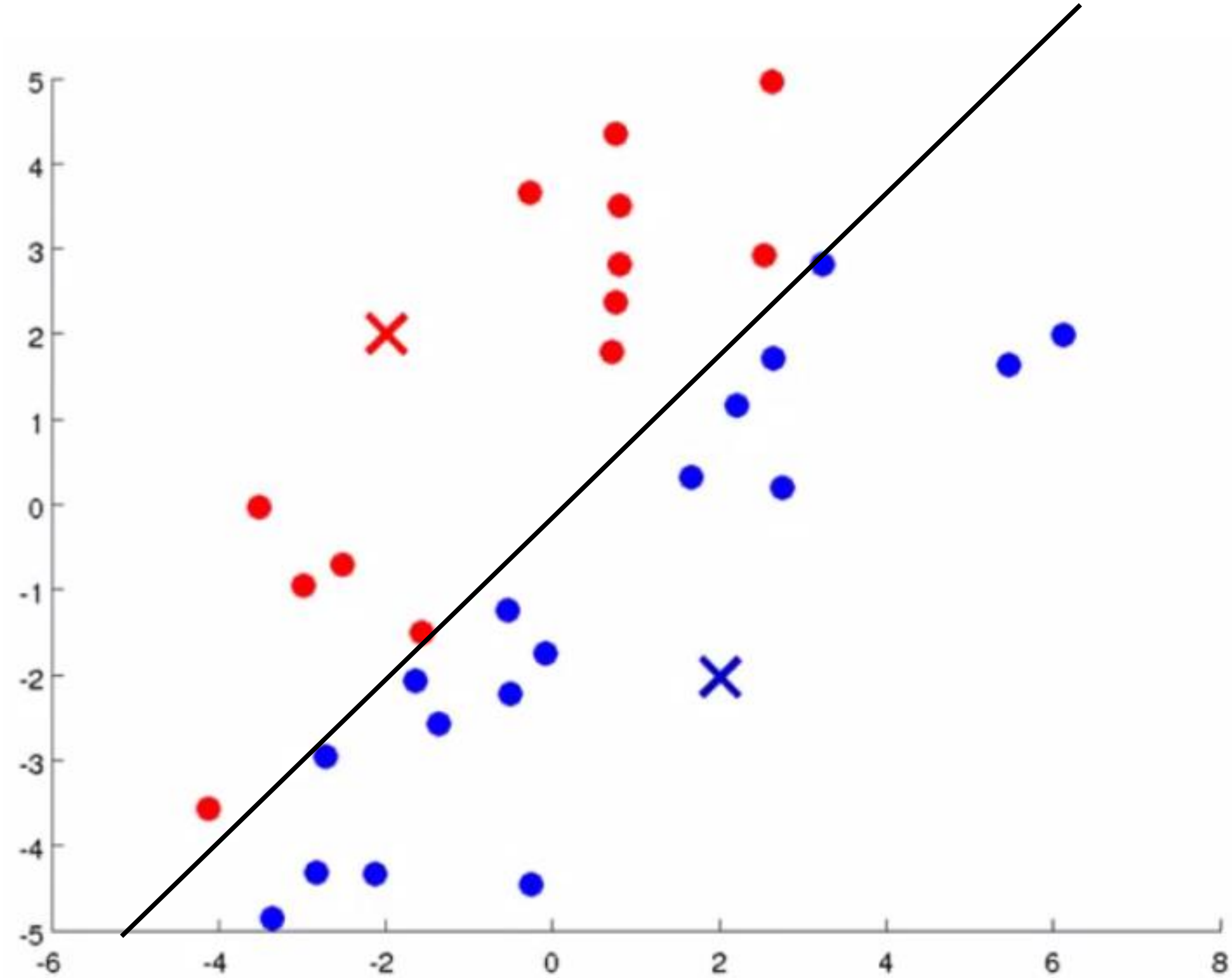
- We assume a probability and μ *and* σ of the distribution

Pictures courtesy Andrew Ngs course on Coursera

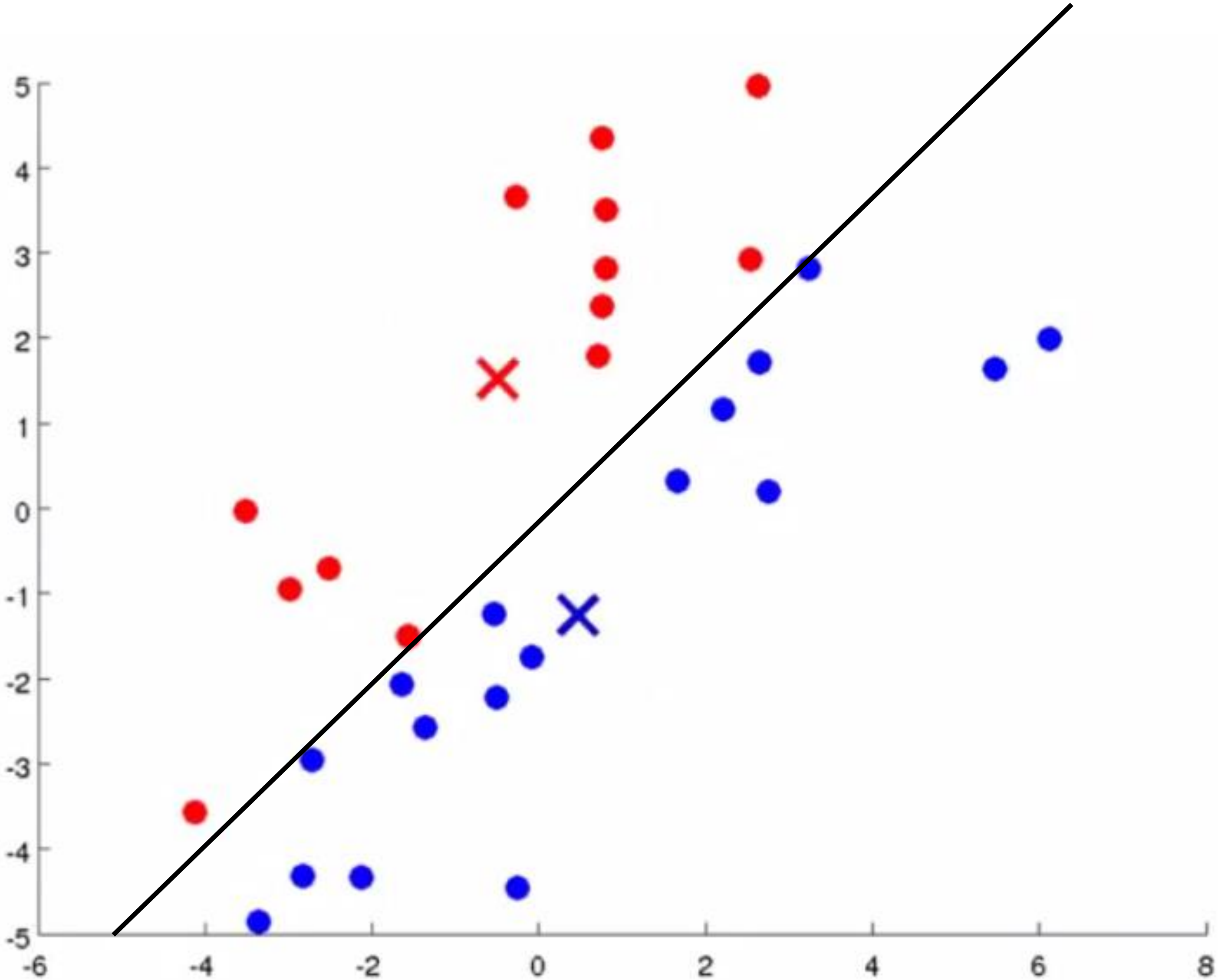


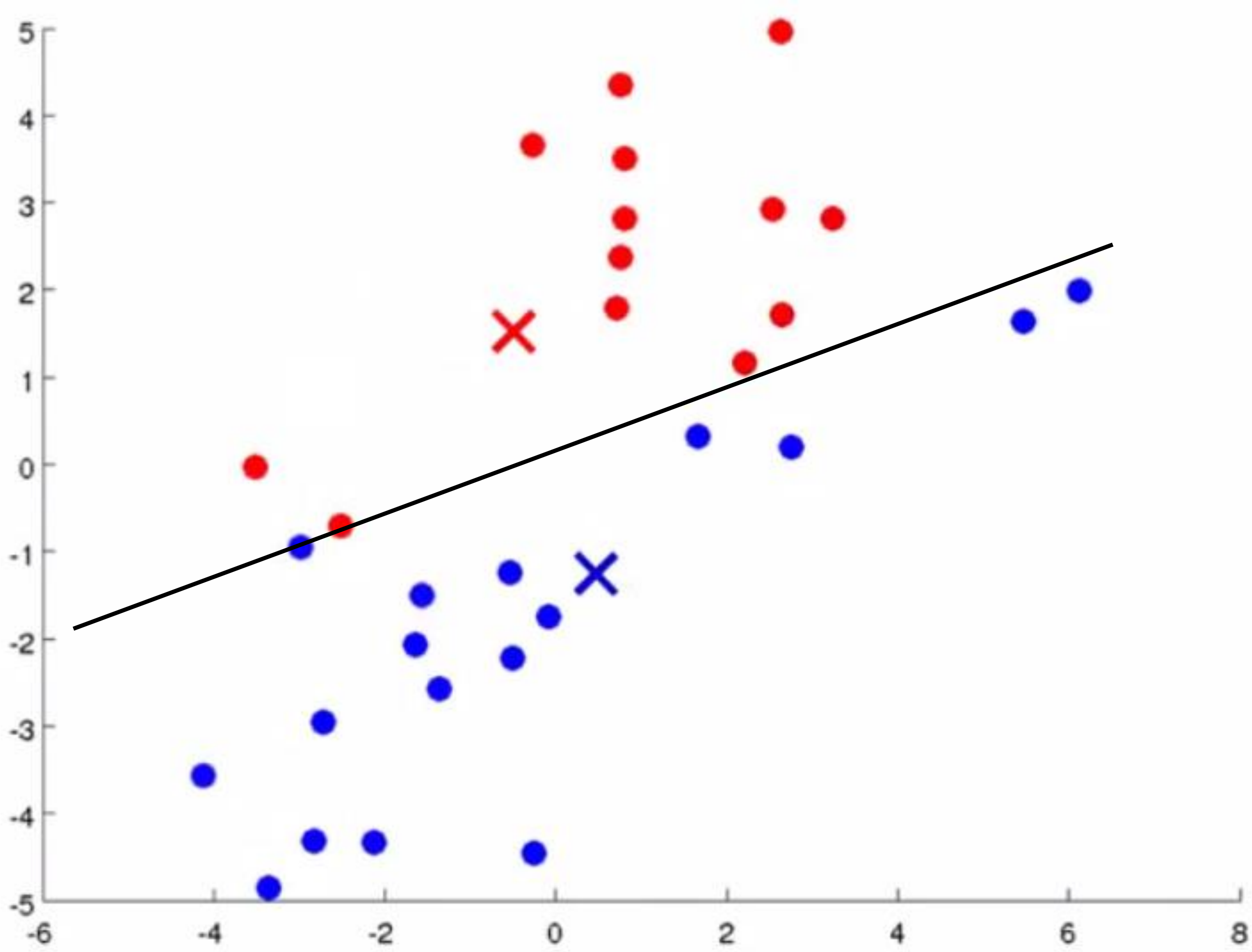


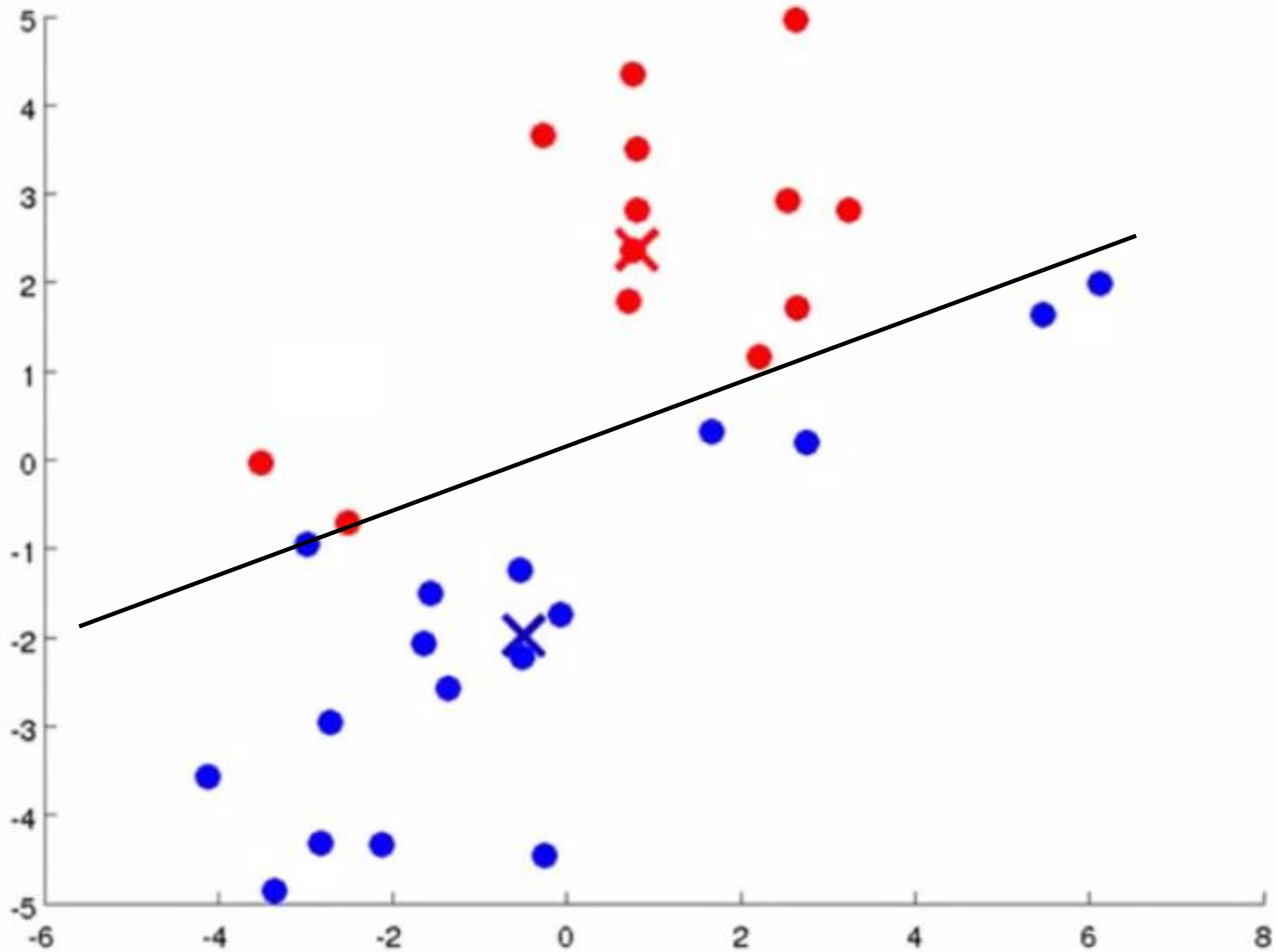
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

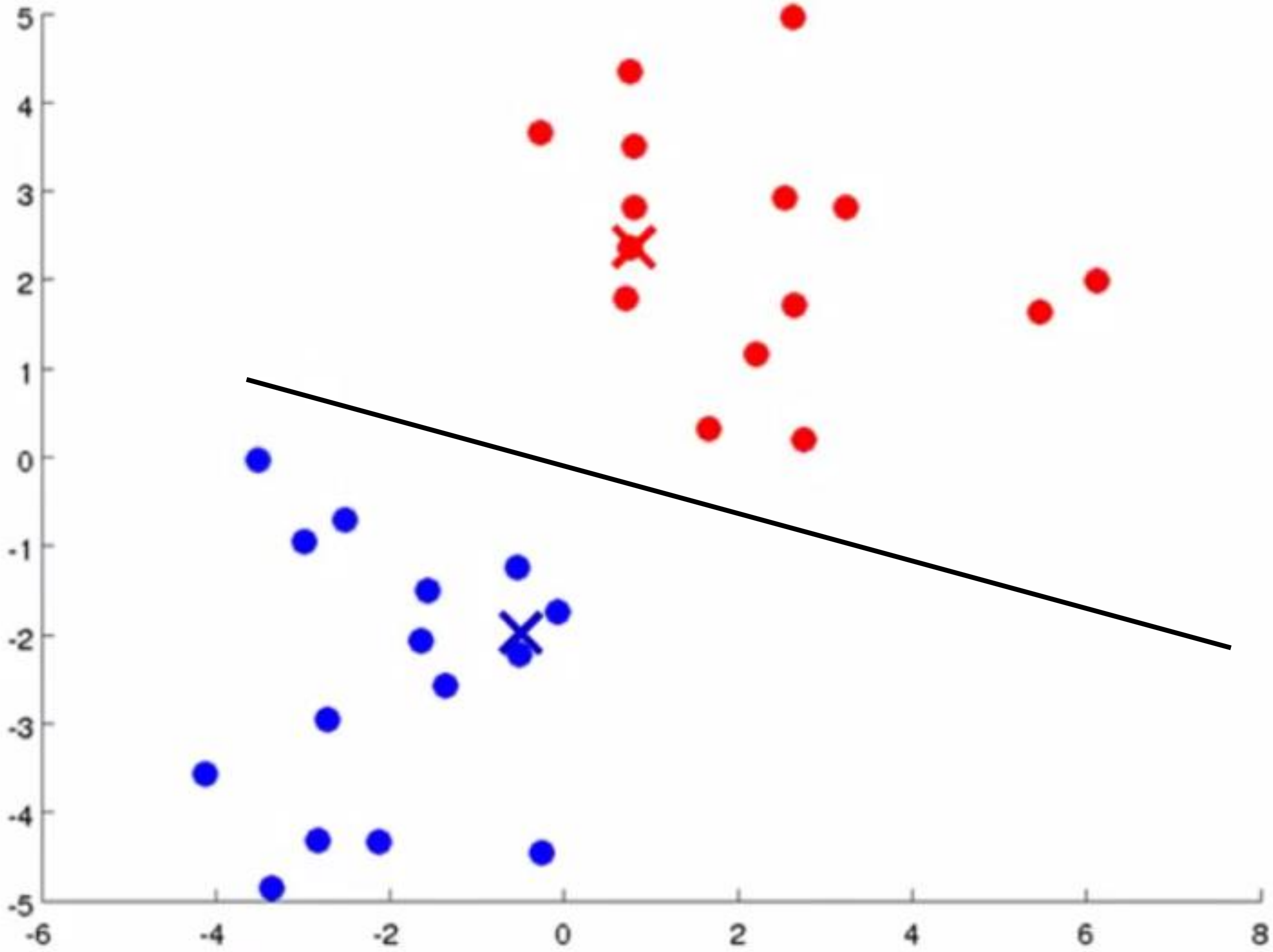


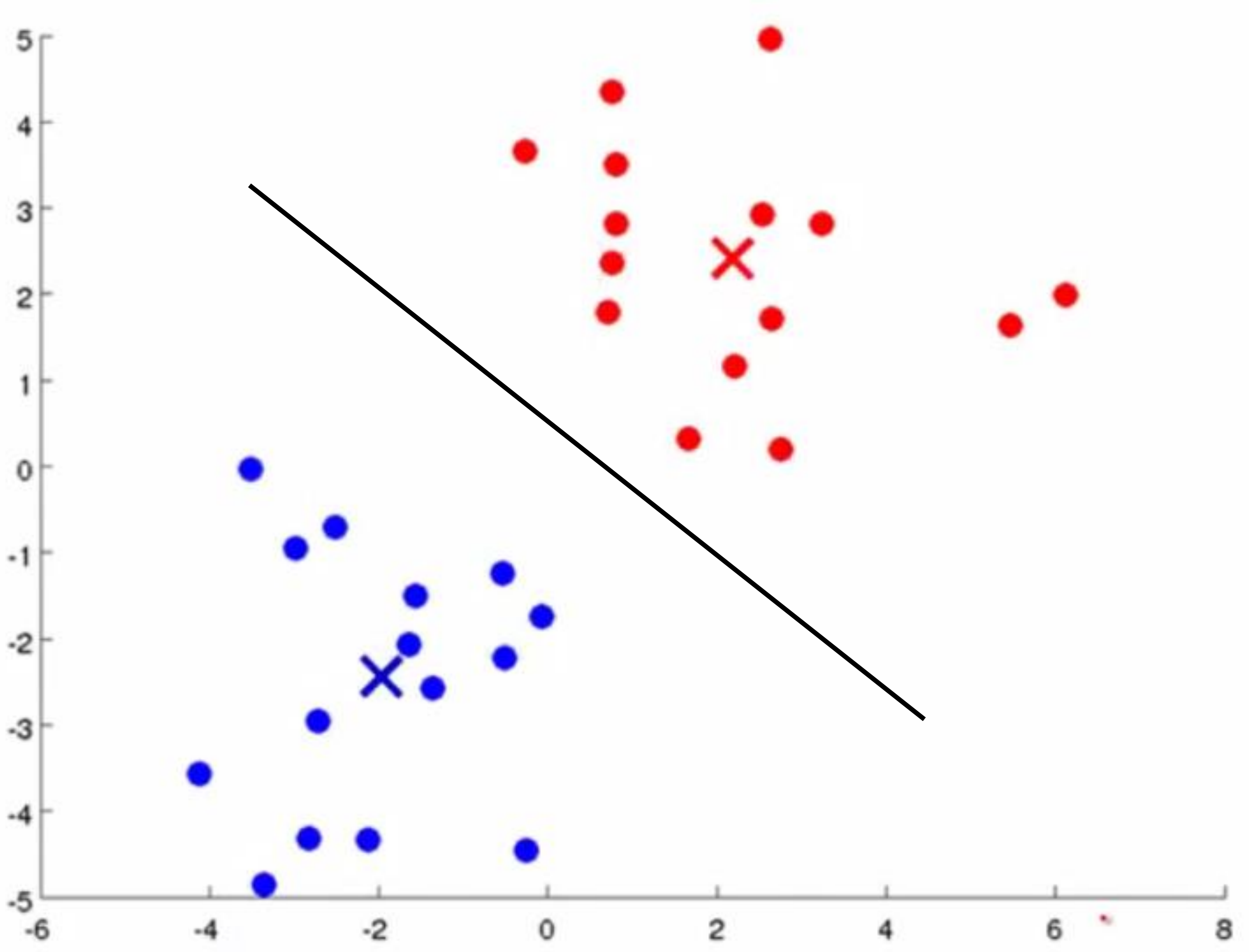
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

E and M steps

E-Step : Assume we know π_i, μ_i, Σ_i

$$r_{ij} = \pi_i \cdot \underbrace{(2\pi)^{-N/2} |\Sigma|^{-1}}_{} \exp -\frac{1}{2}(x_j-\mu_i)^T \Sigma_i^{-1} (x_j-\mu_i)$$

M-Step :

$$\pi_i \leftarrow \sum_j r_{ij} / M$$

$$\mu_i \leftarrow \sum_j r_{ij} x_j \Big/ \sum_j r_{ij}$$

$$\Sigma_i \leftarrow \sum_j r_{ij} (x_j-\mu_i)^T (x_j-\mu_i) \Big/ \sum_j r_{ij}$$

MULTIPLE RANDOM VARIABLES

Relations

- Mutually independent
  - Coin toss
- Mutually exclusive
  - Pass, fail
- Dependent
  - High score in the program, salary

Propositions for two variables

$$Cov(X, X) = Var(X)$$
$$Cov(X, Y) = E(XY) - E(X)E(Y)$$
$$Cov(aX, bY) = abCov(X, Y)$$

$$Var(w_1 x_1 + w_2 x_2) = {w_1}^2 Var(x_1) + {w_2}^2 Var(x_2) + w_1 w_2 cov(x_1, x_2)$$

Joint, conditional and marginal probabilities

	Math	English	Total
Female	1	17	18
Male	37	20	57
Total	38	37	75

P(gender=male, subject=English) = 20/75

P(gender=male | subject=English) = 20/37

P(gender= male) = P(G=M|S=Math)+P(G=M|S=Eng)=57/75

Joint probability Distributions

Employee smartness=0,1
Project difficulty = 0, 1
Success = 0, 1

Possible outcomes = 2*2*2=8

Smartness	Difficulty	Success	Joint Probability Distribution
0	0	0	0.1
0	0	1	0.02
0	1	0	0.2
0	1	1	0.02
1	0	0	0.02
1	0	1	0.20
1	1	0	0.2
1	1	1	0.24

Unnormalized conditioning on tough projects

Smartness	Difficulty	Success	Joint Probability Distribution
0	1	0	0.2
0	1	1	0.02
1	1	0	0.2
1	1	1	0.24

P(Smartness, Success | Difficulty=1)

Normalized conditioning on tough projects

Smartness	Difficulty	Success	Joint Probability Distribution
0	1	0	0.2/0.66= 0.30
0	1	1	0.02/0.66=0.03
1	1	0	0.2/0.66=0.30
1	1	1	0.24/0.66=0.37

$$P(\text{Smartness}, \text{Success} \mid \text{Difficulty}=1)$$

Marginalizing smartness

Smartness	Success	Joint Probability Distribution
0	0	0.30
0	1	0.03
1	0	0.30
1	1	0.37

Success	PD
0	0.6
1	0.4

P(Success | Difficulty=1)

Conditional probability

- For independent variables
  - $P(A|B)=P(A)$
  - $P(A,B) = P(A)*P(B)$
- For dependent variables
  - $P(A,B) = P(A)*P(B|A)$

A problem

- 10 people attended an interview for two jobs. What is the probability that you and your friend get the job

$$2/10 * 1/9 = 2/90 = 1/45$$

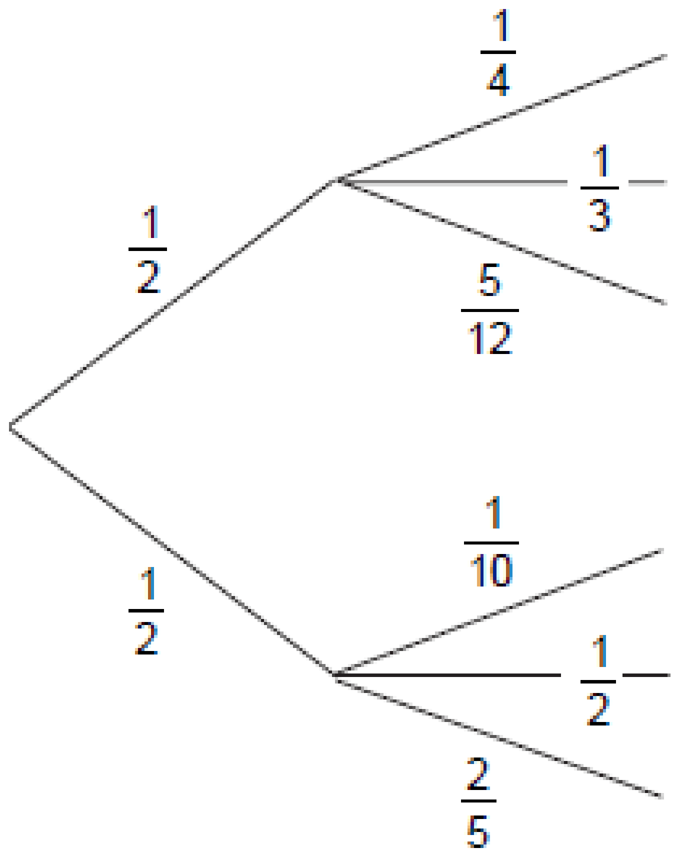
Another problem with a Tree

- There are equal numbers of boys and girls in a school and you know that 1/4 of the boys and 1/10 of the girls walk in every day, 1/3 of the boys and 1/2 of the girls get a lift and the rest come by coach, determine

(a) the proportion of the school population that are girls who go by coach;

(b) the proportion of the school population that go by coach.

Tree diagram



$$\frac{1}{2}\times\frac{1}{4}=\frac{1}{8}=P\left(\text{boy}\cap\text{walk}\right)$$

$$\frac{1}{2}\times\frac{1}{3}=\frac{1}{6}=P\left(\text{boy}\cap\text{lift}\right)$$

$$\frac{1}{2}\times\frac{5}{12}=\frac{5}{24}=P\left(\text{boy}\cap\text{coach}\right)$$

$$\frac{1}{2}\times\frac{1}{10}=\frac{1}{20}=P\left(\text{girl}\cap\text{walk}\right)$$

$$\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}=P\left(\text{girl}\cap\text{lift}\right)$$

$$\frac{1}{2}\times\frac{2}{5}=\frac{1}{5}=P\left(\text{girl}\cap\text{coach}\right)$$

Proportion of the school population that are girls who go by coach;

- $P(C|G) = 2/5$
- $P(C) = (0.5*2/5)+(0.5*5/12)$

Basic Formulas for Probabilities

- Product Rule : probability P(AB) of a conjunction of two events A and B:

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Sum Rule: probability of a disjunction of two events A and B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Theorem of Total Probability : if events A1, …., An are mutually exclusive with

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

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**

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

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

Bayes explained

- There is prior knowledge = P(A)

- There is some test evidence = P(B|A)

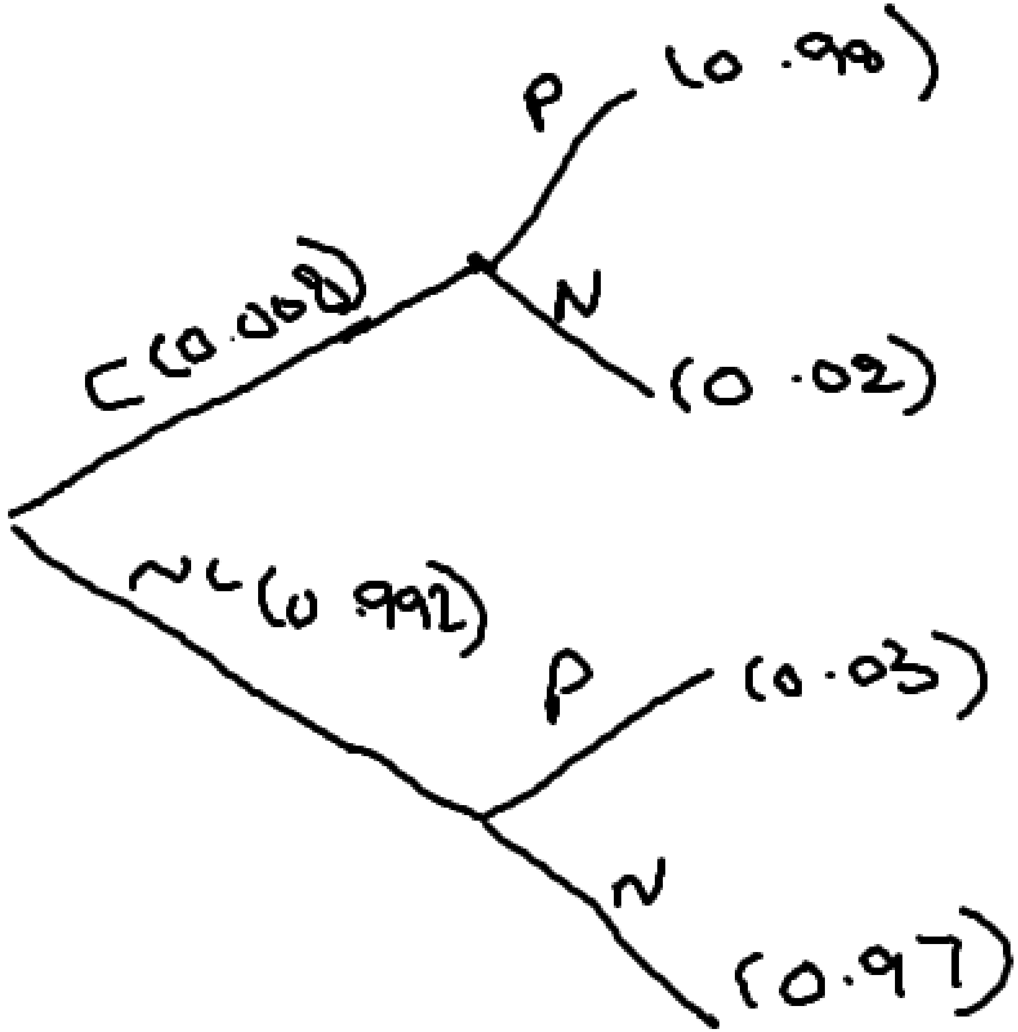
- Posterior = P(A|B)

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

An Example

- A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 98% of the cases. In the remaining 2% cases, the test returns negative even thought the patient has cancer. The test yields a correct negative result in only 97% of the cases. In the other 3% cases, the test is positive even though the patient does not have cancer. Furthermore, only 0.008 of the entire population has this disease.

    1. What is the probability that this patient has cancer?
    2. What is the probability that he does not have cancer?
    3. What is the diagnosis?



C (0.008)
P
(0.98)
N
(0.02)
~C (0.992)
P
(0.03)
N
(0.97)

Does patient have cancer or not?

- Various probabilities are

  $P(cancer) = 0.008$ $P(\sim cancer) = 0.992$

  $P(+|cancer) = 0.98$ $P(-|cancer) = 0.02$

  $P(+|\sim cancer) = 0.03$ $P(-|\sim cancer)$ 0.97

  - Now a new patient, whose test result is positive, Should we diagnose the patient have cancer or not?

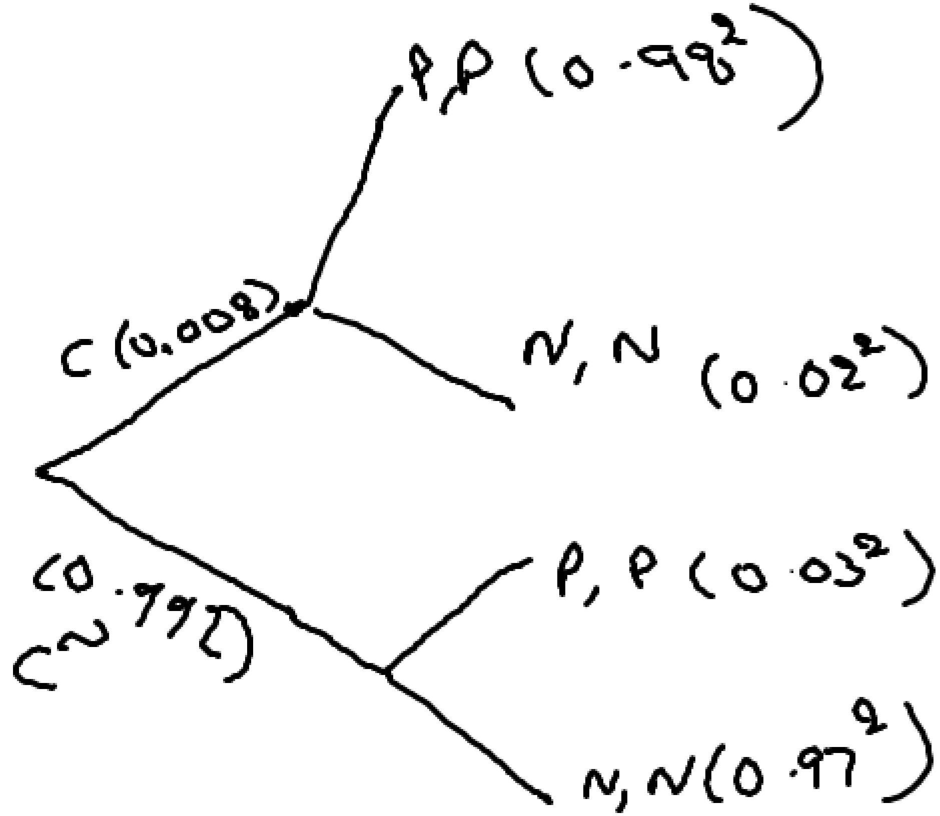
  $P(cancer|p)= P(+|cancer)P(cancer)=0.98*0.008 = 0.0078$

  $P(\sim cancer|p)=P(+|\sim cancer)P(\sim cancer) = 0.03*0.992 = 0.0298$

Diagnosis: ~cancer

How to be convinced then

- If we double test



The renewed probabilities

- P(+|cancer)P(cancer) (twice) = 0.98*0.98*0.008 = 0.0077
- P(+|~cancer)P(~cancer)(twice) = 0.03*0.03*0.992 = 0.0009
- So, the patient has cancer

International School of Engineering

The Oval, Level 2, iLabs Centre, Plot No. 18, Madhapur, Hyderabad - 500 081

For Individuals: +91-9177585755 or 040-65743991

For Corporates: +91-7893866005

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/intl.school.engineering

Twitter: https://twitter.com/IntlSchEngg

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE