Spring 2019

CSEN1083: Data Mining

Linear Algebra Review

Seif Eldawlatly

Linear Algebra Review: Matrices

- Matrix: A set of elements organized in rows and columns

$$\begin{array}{cc} & \begin{array}{cc}\text{Col 1} & \text{Col 2}\end{array} \\ \begin{array}{c}\text{Row 1} \\ \text{Row 2}\end{array} & \begin{bmatrix} a & b \\ c & d \end{bmatrix} \end{array}$$

- Matrix Dimensions: (# of Rows) x (# of Columns)
- Matrix Addition and Subtraction

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \pm \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a \pm e & b \pm f \\ c \pm g & d \pm h \end{bmatrix}$$

Linear Algebra Review: Matrices

- Matrices Multiplication

$$\begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} \times \begin{bmatrix} x & z & k \\ y & g & l \end{bmatrix} = \begin{bmatrix} ax+by & az+bg & ak+bl \\ cx+dy & cz+dg & ck+dl \\ ex+fy & ez+fg & ek+fl \end{bmatrix}$$

(3 x 2) x (2 x 3) = (3x3)

- Multiplying an (n x m) matrix by (m x k) matrix results in (n x k) matrix

- Matrix Transpose

$$\mathbf{M} = \begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} \qquad \mathbf{M}^T = \begin{bmatrix} a & c & e \\ b & d & f \end{bmatrix}$$

(3 x 2) (2 x 3)

Linear Algebra Review: Matrices

- Inverse of matrix $\mathbf{A}$ denoted by $\mathbf{A}^{-1}$

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \qquad \text{where} \qquad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Identity matrix if $\mathbf{A}$ is (3 x 3)

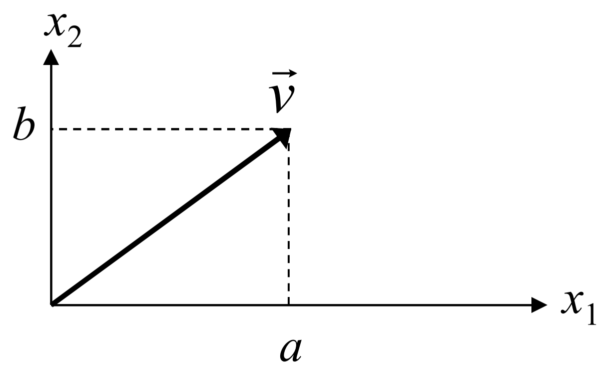
- For a (2 x 2) matrix

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \det(\mathbf{A}) = ad - bc \qquad \mathbf{A}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

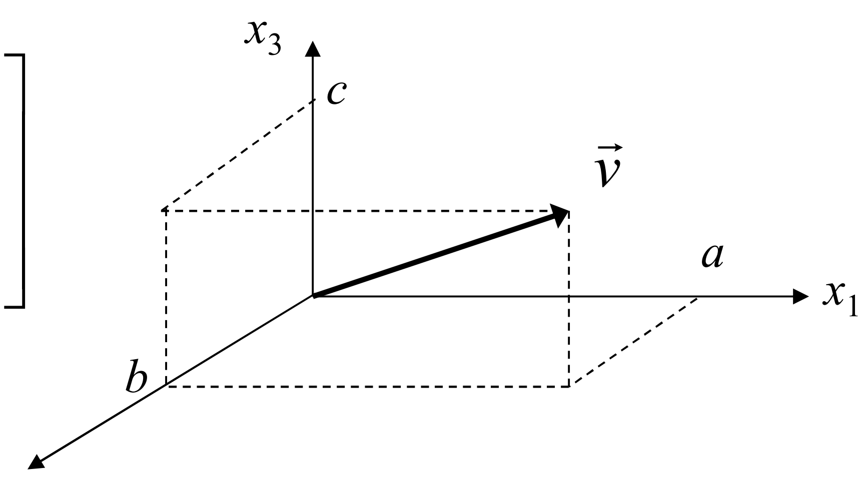
Linear Algebra Review: Vectors

- Vector = n x 1 matrix
- Represents a straight line in n-dimensional space

$$\vec{v} = \begin{bmatrix} a \\ b \end{bmatrix}$$



$$\vec{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

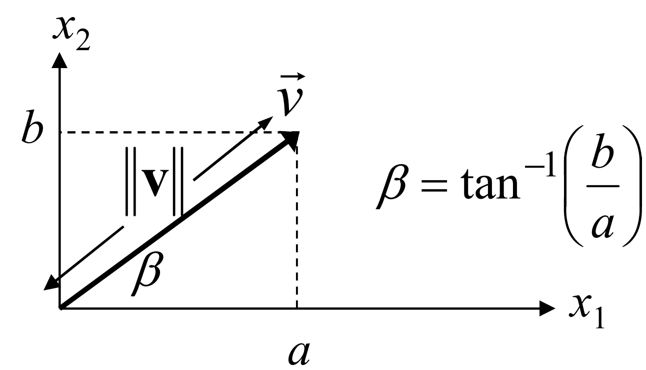


- A vector is denoted as $\vec{v}$ or $\mathbf{v}$

Linear Algebra Review: Vectors

- Vector Magnitude (Norm): Gives the length of a vector

$$\mathbf{v} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad \|\mathbf{v}\| = \sqrt{a^2 + b^2}$$



$$\beta = \tan^{-1}\left(\frac{b}{a}\right)$$

- Unit Vector: Vector with norm = 1
- For an n-dimensional vector

$$\mathbf{v} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \qquad \|\mathbf{v}\| = \sqrt{\sum_{i=1}^{n} a_i^2}$$

- The norm of the vector squared is equivalent to

$$\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a_1^2 + a_2^2 + \ldots + a_n^2 = \sum_{i=1}^{n} a_i^2$$

(1 x n) (n x 1) (1 x 1)

Linear Algebra Review: Vectors

- Vectors Dot Product

$$\mathbf{v} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} c \\ d \end{bmatrix}$$

$$\mathbf{v} \cdot \mathbf{u} = \mathbf{v}^T \mathbf{u} = \mathbf{u}^T \mathbf{v} = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = ac + bd$$

(Always (1 x 1))

- For an n-dimensional vector

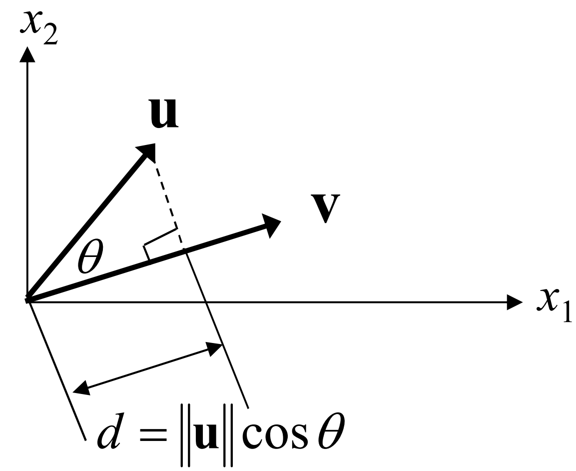
$$\mathbf{v} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\mathbf{v} \cdot \mathbf{u} = \mathbf{v}^T \mathbf{u} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \sum_{i=1}^{n} a_i b_i$$

Linear Algebra Review: Vectors

- Dot product can be expressed as

$$\mathbf{v} \cdot \mathbf{u} = \|\mathbf{v}\| . \|\mathbf{u}\| \cos\theta$$



- $\|\mathbf{u}\| \cos\theta$ is the projection of the vector $\mathbf{u}$ on the vector $\mathbf{v}$

- If $\mathbf{v}$ is a unit vector, then the dot product is equivalent to the projection of the vector $\mathbf{u}$ on the vector $\mathbf{v}$

- If both vectors are unit vectors, the dot product will be maximum if both vectors are perfectly aligned

- If two vectors are orthogonal, the dot product will equal 0

Linear Algebra Review

- Matrix Calculus

$$\frac{\partial(\mathbf{a}^T\mathbf{x})}{\partial\mathbf{x}} = \frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial\mathbf{x}} = \mathbf{a}$$

$$\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} = (\mathbf{A}+\mathbf{A}^T)\mathbf{x} \longrightarrow \frac{\partial\mathbf{x}^T\mathbf{x}}{\partial\mathbf{x}} = 2\mathbf{x}$$

$$\frac{\partial(\mathbf{a}^T\mathbf{X}\mathbf{b})}{\partial\mathbf{X}} = \mathbf{a}\mathbf{b}^T$$

Function Optimization

- To find the minimum (maximum) of a function, take the derivative and equate with zero

$$\min f(x) = f(x^*) \quad \text{where} \quad \frac{df(x)}{dx}\Big|_{x^*} = 0$$

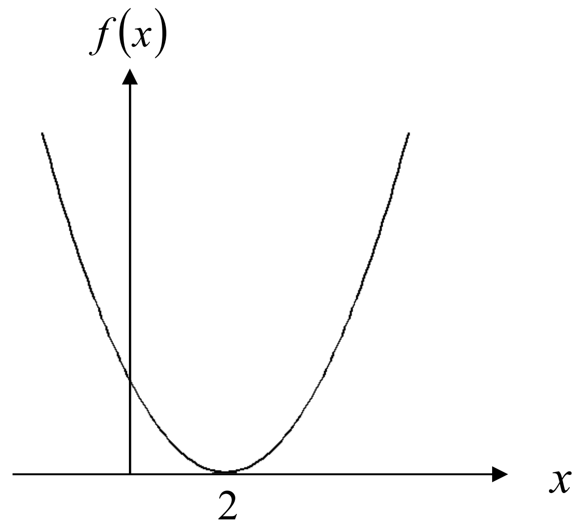
- Example

$$f(x) = (x-2)^2$$

$$f'(x) = 2(x-2) = 2x - 4 = 0$$

$$x^* = 2$$

$$f(x^*) = 0$$



Spring 2019

CSEN1083: Data Mining

Probability Theory Review

Seif Eldawlatly

Some parts are adapted from Prof. Rong Jin’s slides

Definition of Probability

- ***Experiment***: toss a coin twice
- ***Sample space***: possible outcomes of an experiment
  - S = {HH, HT, TH, TT}
- ***Event***: a subset of possible outcomes
  - A={HH}, B={HT, TH}
- ***Probability of an event***: a number assigned to an event Pr(A)
  - Axiom 1: Pr(A) $\geq$ 0
  - Axiom 2: Pr(S) = 1
  - Axiom 3: For every sequence of disjoint events

$$\Pr\left(\bigcup_i A_i\right) = \sum_i \Pr(A_i)$$

  - Example: Pr(A) = n(A)/N: frequentist statistics

    (If we repeat an experiment N times, and denote by n(A) the number of times we observe A, then Pr(A) = n(A)/N)

Joint Probability

- For events A and B, **joint probability** Pr(AB) (or Pr(A, B)) stands for the probability that both events happen

- Example: What is the probability that the first toss is H and the second toss is H?

  Pr(1^{st} is H and 2^{nd} is H)=Pr(1^{st} is H)Pr(2^{nd} is H)=0.5x0.5=0.25

- Example: A={HH}, B={HT, TH}, what is the joint probability Pr(AB)?

  Answer: 0

Independence

- Two events ***A and B are independent*** if

$$\Pr(AB) = \Pr(A)\Pr(B)$$

- A set of events $\{A_i\}$ is independent if

$$\Pr\left(\bigcap_i A_i\right) = \prod_i \Pr(A_i)$$

- Example: Drug test

	Women	Men
Success	200	1800
Failure	1800	200

A = {Patient is a Woman}

B = {Drug fails}

Is event A independent of event B?

- Pr(AB)=1800/4000, Pr(A)=2000/4000, Pr(B)=2000/4000

$\because$ $\Pr(AB) \neq \Pr(A)\Pr(B) \rightarrow$ A and B are dependent

Conditioning

- If A and B are events with Pr(A) > 0, the ***conditional probability of B given A*** is

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)}$$

- Example: Drug test

	Women	Men
Success	200	1800
Failure	1800	200

A = {Patient is a Woman}

B = {Drug fails}

Pr(B|A) = ?

Pr(A|B) = ?

- Pr(B|A)=Pr(AB)/Pr(A)=(1800/4000)/(2000/4000)=0.9
- Pr(A|B)=Pr(AB)/Pr(B)=(1800/4000)/(2000/4000)=0.9

- Given A is independent from B, what is the relationship between Pr(A|B) and Pr(A)?

  Pr(A|B) = Pr(A)

Bayes' Rule

- Given two events A and B and suppose that Pr(A) > 0, then

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)} = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}$$

- Example:

$\Pr(R) = 0.8$

$\Pr(W\|R)$	R	$\neg R$
W	0.7	0.4
$\neg W$	0.3	0.6

R: It is a rainy day

W: The grass is wet

$\Pr(R|W)$ = ?

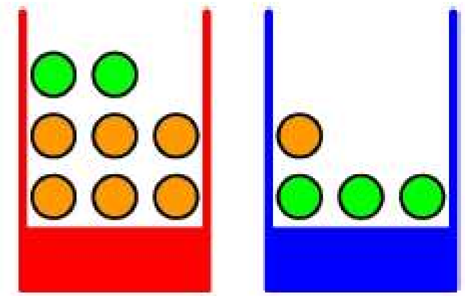
$$\Pr(R \mid W) = \frac{\Pr(W \mid R)\Pr(R)}{\Pr(W)} = \frac{0.7 \times 0.8}{\Pr(W)}$$

$$\Pr(W) = \Pr(WR) + \Pr(W\neg R) = \Pr(W \mid R)\Pr(R) + \Pr(W \mid \neg R)\Pr(\neg R)$$

$$= 0.7 \times 0.8 + 0.4 \times 0.2 = 0.64$$

Bayes' Rule

- Example:
  Two boxes: one red and one blue
  Two kinds of fruit: apples and oranges
  Task: randomly pick one of the boxes
  and then select a fruit



Let B represent the box (B = r or B = b)
Let F represent the fruit (F = a or F = o)

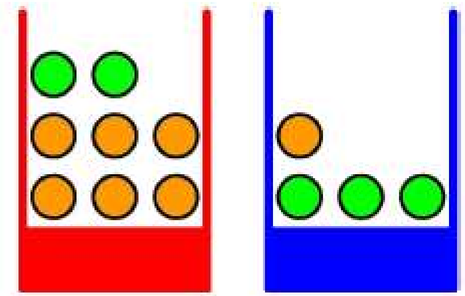
Let the probability of picking the red box be 40% and the blue box be 60%

$$\begin{aligned} p(B=r) &= 4/10 \\ p(B=b) &= 6/10 \end{aligned}$$

Bayes' Rule

- Conditional Probabilities

$$
\begin{aligned}
p(F=a|B=r) &= 1/4 \\
p(F=o|B=r) &= 3/4 \\
p(F=a|B=b) &= 3/4 \\
p(F=o|B=b) &= 1/4.
\end{aligned}
$$



- What is the probability of picking an apple? ($p(F = a)$)

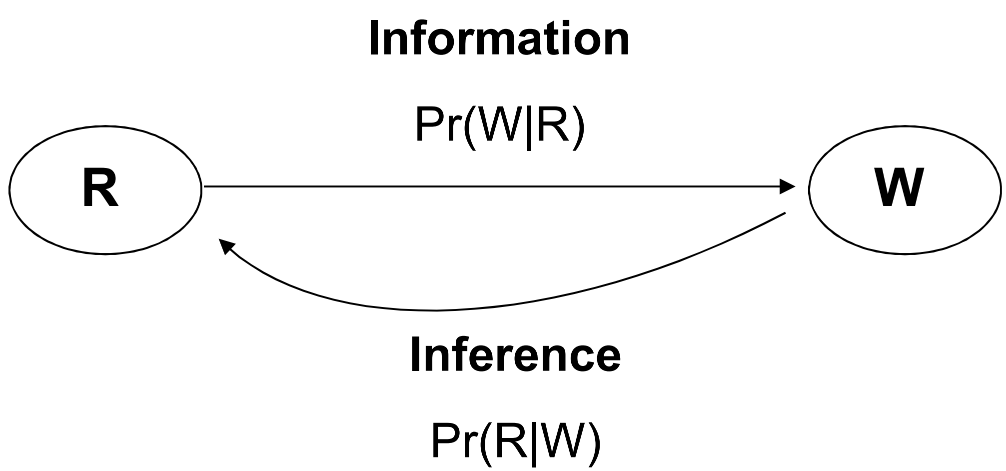
$$
\begin{aligned}
p(F=a) &= p(F=a|B=r)p(B=r) + p(F=a|B=b)p(B=b) \\
&= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20}
\end{aligned}
$$

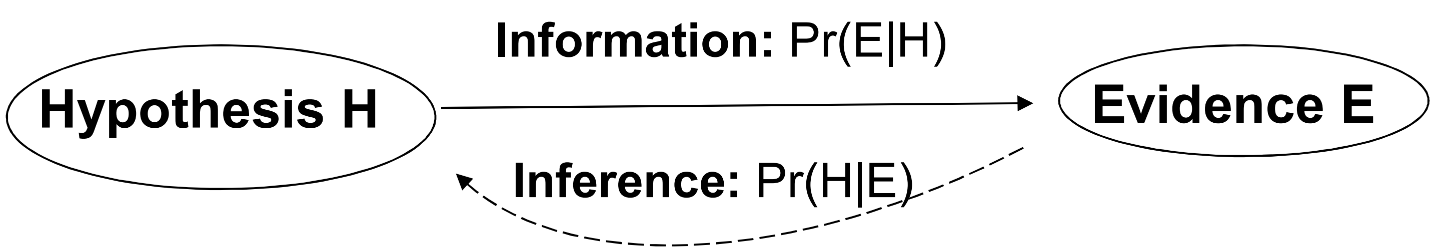
Bayes' Rule

	R	¬R
W	0.7	0.4
¬W	0.3	0.6

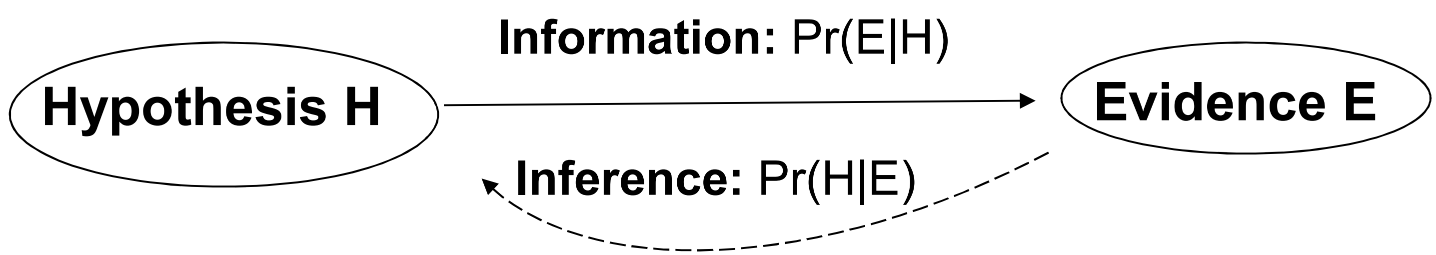
R: It rains

W: The grass is wet





Bayes' Rule



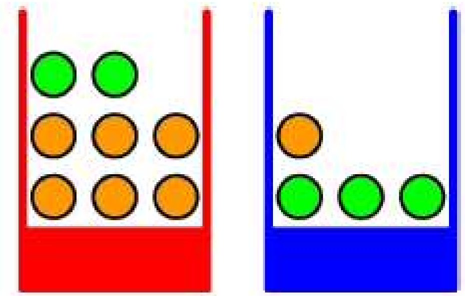
Posterior **Likelihood** **Prior**

$$\boxed{\Pr(H \mid E)} = \frac{\boxed{\Pr(E \mid H)}\,\boxed{\Pr(H)}}{\Pr(E)}$$

- Prior Probability: Probability available ***before*** observing the evidence
- Posterior Probability: Probability obtained ***after*** observing the evidence

Bayes' Rule

- $p(B)$: Prior probability: Probability available before observing the identity of the fruit



- $p(B|F)$: Posterior probability: Probability obtained after observing the picked fruit

- Since $p(B = r)$ is 0.4, we are more likely to pick the blue box

- However, once we observe that the picked fruit is an orange we find that it's more likely that we picked from the red box (posterior probability)

$$p(B=r|F=o) = \frac{p(F=o|B=r)p(B=r)}{p(F=o)} = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3}$$

Bayes' Rule: More Complicated

- Suppose that B_1, B_2, … B_k form a partition of S:

$$B_i \bigcap B_j = \varnothing; \quad \bigcup_i B_i = S$$

Suppose that $\Pr(B_i) > 0$ and $\Pr(A) > 0$. Then

$$\Pr\left(B_i \middle| A\right) = \frac{\Pr\left(A \middle| B_i\right)\Pr\left(B_i\right)}{\Pr\left(A\right)}$$

$$= \frac{\Pr\left(A \middle| B_i\right)\Pr\left(B_i\right)}{\sum_{j=1}^{k} \Pr\left(AB_j\right)}$$

$$= \frac{\Pr\left(A \middle| B_i\right)\Pr\left(B_i\right)}{\sum_{j=1}^{k} \Pr\left(B_j\right)\Pr\left(A \middle| B_j\right)}$$

Random Variable and Distribution

- A ***random variable*** X is a numerical outcome of a random experiment

- The ***distribution*** of a random variable is the collection of possible outcomes along with their probabilities:

  – Discrete case: $\Pr(X = x) = p_\theta(x)$

  – Continuous case: $\Pr(a \le X \le b) = \int_a^b p_\theta(x)dx$

  θ represents the parameter(s) of the distribution

Random Variable: Example and Distribution

- Let S be the set of all sequences of three rolls of a die. Let X be the sum of the number of dots on the three rolls

- What are the possible values of X?

  Answer: 3, 4, 5, 6, …, 18

- $\Pr(X = 5) = ?$

  To get $X = 5$: (1,1,3), (1,3,1), (3,1,1),

  (1,2,2), (2,1,2), (2,2,1)

  $\Pr(X = 5) = 6/6^3$

Expectation

- A random variable X ~ Pr(X = x). Then, its expectation is

$$E[X] = \sum_x x \Pr(X = x)$$

  - In an empirical sample, x_1, x_2,…, x_N, the sample mean is

$$E[X] = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Continuous case: $E[X] = \int_{-\infty}^{\infty} x p_\theta(x) dx$

- Expectation of sum of random variables

$$E[X_1 + X_2] = E[X_1] + E[X_2]$$

- Expectation of product of random variables

$$E[X_1 X_2] = \sum_{x_1}\sum_{x_2} x_1 x_2 \Pr(X_1 = x_1, X_2 = x_2)$$

If X_1 and X_2 are independent, $\Pr(X_1 = x_1, X_2 = x_2) = \Pr(X_1 = x_1)\Pr(X_2 = x_2)$

$$\therefore E[X_1 X_2] = \sum_{x_1} x_1 \Pr(X_1 = x_1) \sum_{x_2} x_2 \Pr(X_2 = x_2) = E[X_1]E[X_2]$$

Expectation: Example

- Let S be the set of all sequences of three rolls of a die. Let X be the sum of the number of dots on the three rolls.

- What is $E[X]$?

  Answer: $X = X_1 + X_2 + X_3$

$$E[X] = E[X_1] + E[X_2] + E[X_3]$$

$$E[X_i] = \sum_{x_i} x_i \Pr(X_i = x_i)$$

$$= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{21}{6}$$

$$\therefore E[X] = \frac{21}{6} + \frac{21}{6} + \frac{21}{6} = 10.5$$

Expectation: Example

- Let S be the set of all sequences of three rolls of a die. Let X be the product of the number of dots on the three rolls
- What is $E[X]$?

  Answer: $X=X_1X_2X_3$

  Since the three rolls are independent, then

  $$E[X]=E[X_1]E[X_2]E[X_3]$$

  $$=\left(\frac{21}{6}\right)^3$$

Variance

- The variance of a random variable measures how much variability there is in the random variable

$$
\begin{aligned}
Var(X) &= E((X - E[X])^2) \\
&= E(X^2 + E[X]^2 - 2XE[X]) \\
&= E(X^2 - E[X]^2) \\
&= E[X^2] - E[X]^2
\end{aligned}
$$

- Population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$, where μ is the sample mean

- Example

  Let X represent randomly sampled integer numbers in the range (1, 100):

  - {2, 50, 9, 4, 23, 65, 99} → $\sigma^2 = \frac{1}{7}\sum_{i=1}^{7}(x_i - 36)^2 = 1154.85$

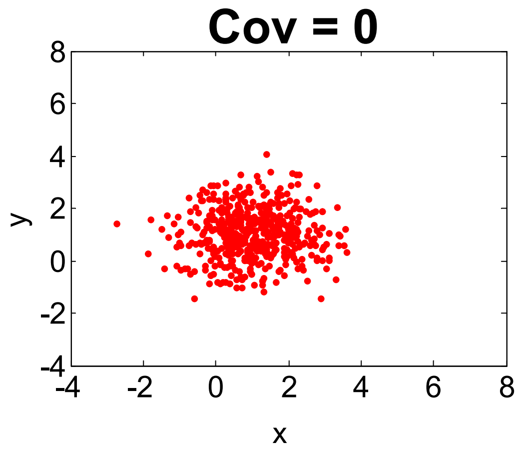
  - {33, 34, 35, 36, 37, 38, 39} → $\sigma^2 = \frac{1}{7}\sum_{i=1}^{7}(x_i - 36)^2 = 4$

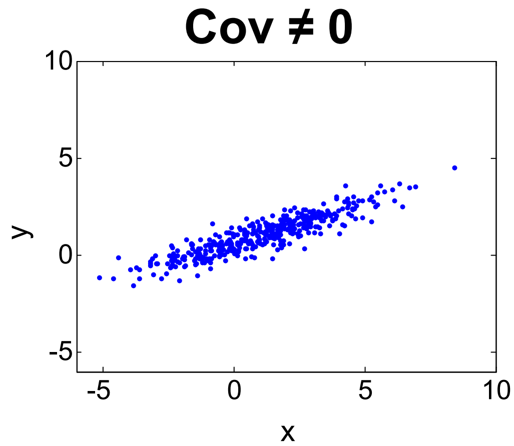
Covariance

- The covariance of two random variables measures the extent to which the two variables vary together

$$
\begin{aligned}
Cov[X,Y] &= E[(X-E[X])(Y-E[Y])] \\
&= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\
&= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\
&= E[XY] - E[X]E[Y]
\end{aligned}
$$

- Sample covariance $C_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$
- Example





Bernoulli Distribution

- The outcome of an experiment can either be success (1) or failure (0)

- $\Pr(X = 1) = \theta$, $\Pr(X = 0) = 1 - \theta$, or

$$p_\theta(x) = \theta^x (1-\theta)^{1-x}$$

- $E[X] = \theta$, $Var(X) = \theta(1 - \theta)$

Gaussian Distribution

- $X \sim N(\mu, \sigma)$

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

$$\Pr(a \le X \le b) = \int_a^b p_\theta(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} dx$$

- $E[X] = \mu$, Var(X)= σ^2

- Mean: μ, Standard deviation: σ

- If $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$, $X = X_1 + X_2$?

  Answer: X is Gaussian as well with mean $\mu_1 + \mu_2$ and variance

$$\sigma^2 = \sigma_1^2 + \sigma_2^2$$

Plots of Gaussian Distribution