

Essential Mathematics for Machine Learning

Dr. Dinesh Kumar Vishwakarma

Professor,

Department of Information Technology,

Delhi Technological University, Delhi-110042

dinesh@dtu.ac.in

<http://www.dtu.ac.in/Web/Departments/InformationTechnology/faculty/dkvishwakarma.php>

Areas of math essential to machine learning

- Machine learning is part of both ***statistics*** and computer science
 - Probability
 - Statistical inference
 - Validation
 - Estimates of error, confidence intervals
- ***Linear algebra***
 - Hugely useful for compact representation of linear transformations on data
 - Dimensionality reduction techniques
- ***Optimization*** theory

Why worry about the math?

- There are lots of easy-to-use machine learning packages out there.
- After this course, you will know how to apply several of the most general-purpose algorithms.

HOWEVER

- To get really useful results, you need good mathematical intuitions about certain general machine learning principles, as well as the inner workings of the individual algorithms.

Why worry about the math?

These intuitions will allow you to:

- Choose the right algorithm(s) for the problem
- Make good choices on parameter settings, validation strategies
- Recognize over- or underfitting
- Troubleshoot poor / ambiguous results
- Put appropriate bounds of confidence / uncertainty on results
- Do a better job of coding algorithms or incorporating them into more complex analysis pipelines

Notation

- $a \in A$ *set membership: a is member of set A*
- $| B |$ *cardinality: number of items in set B*
- $\| \mathbf{v} \|$ *norm: length of vector v*
- Σ *summation*
- \int *integral*
- \mathbb{R} the set of *real* numbers
- \mathbb{R}^n *real number space of dimension n*
 - n = 2 : plane or 2-space
 - n = 3 : 3- (dimensional) space
 - n > 3 : *n*-space or *hyperspace*

Notation

- **$\mathbf{x}, \mathbf{y}, \mathbf{z}$** , *vector* (bold, lower case)
 \mathbf{u}, \mathbf{v}
- **$\mathbf{A}, \mathbf{B}, \mathbf{X}$** *matrix* (bold, upper case)
- $v = f(x)$ *function (map)*: assigns unique value in range of y to each value in domain of x
- dy / dx *derivative* of y with respect to single variable x
- $y = f(\mathbf{x})$ *function* on multiple variables, i.e. a vector of variables; *function* in n -space
- $\partial y / \partial x_i$ *partial derivative* of y with respect to element i of vector \mathbf{x}

The concept of probability

Intuition:

- In some process, several outcomes are possible. When the process is repeated a large number of times, each outcome occurs with a characteristic *relative frequency*, or *probability*. If a particular outcome happens more often than another outcome, we say it is more probable.

The concept of probability

Arises in two contexts:

- ***In actual repeated experiments.***
 - Example: You record the color of 1000 cars driving by. 57 of them are green. You *estimate* the probability of a car being green as $57 / 1000 = 0.057$.
- ***In idealized conceptions of a repeated process.***
 - Example: You consider the behavior of an unbiased six-sided die. The *expected* probability of rolling a 5 is $1 / 6 = 0.1667$.
 - Example: You need a model for how people's heights are distributed. You choose a normal distribution (bell-shaped curve) to represent the *expected* relative probabilities.

Probability spaces

A *probability space* is a *random process* or *experiment* with three components:

- Ω , the set of possible *outcomes* O
 - ◆ number of possible outcomes = $|\Omega| = N$
- F , the set of possible *events* E
 - ◆ an event comprises 0 to N outcomes
 - ◆ number of possible events = $|F| = 2^N$
- P , the *probability distribution*
 - ◆ function mapping each outcome and event to real number between 0 and 1 (the *probability* of O or E)
 - ◆ probability of an event is *sum* of probabilities of possible outcomes in event

Axioms of probability

1. Non-negativity:

for any event $E \in \mathcal{F}$, $p(E) \geq 0$

2. All possible outcomes:

$$p(\Omega) = 1$$

3. Additivity of disjoint events:

for all events $E, E' \in \mathcal{F}$ where $E \cap E' = \emptyset$,

$$p(E \cup E') = p(E) + p(E')$$

Types of probability spaces

Define $|\Omega|$ = number of possible outcomes

- **Discrete space** $|\Omega|$ is finite
 - Analysis involves *summations* (Σ)
- **Continuous space** $|\Omega|$ is infinite
 - Analysis involves *integrals* (\int)

Example of discrete probability space

Single roll of a six-sided die

- 6 possible outcomes: $O = 1, 2, 3, 4, 5, \text{ or } 6$
- $2^6 = 64$ possible events
 - ◆ example: $E = (O \in \{ 1, 3, 5 \})$, i.e. outcome is odd
- If die is fair, then probabilities of outcomes are equal

$$p(1) = p(2) = p(3) =$$

$$p(4) = p(5) = p(6) = 1 / 6$$

- ◆ example: probability of event $E = (\text{outcome is odd})$ is

$$p(1) + p(3) + p(5) = 1 / 2$$

Example of discrete probability space

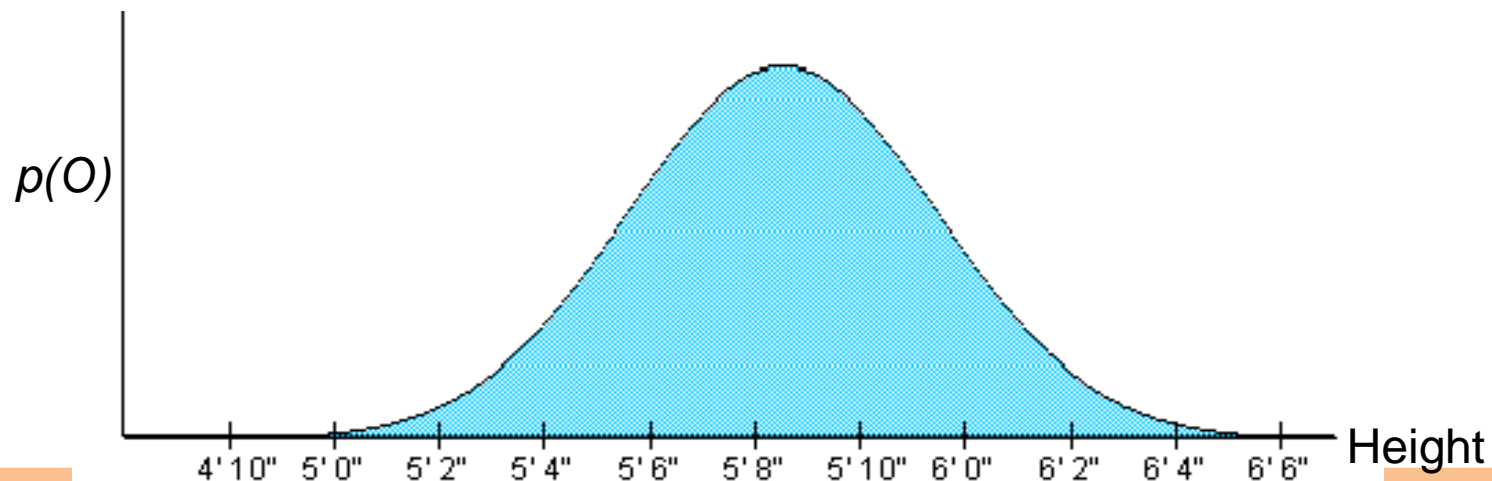
Three consecutive flips of a coin

- 8 possible outcomes: $O = \text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}$
- $2^3 = 8$ possible events
 - ◆ example: $E = (O \in \{ \text{HHT, HTH, THH} \})$, i.e. exactly two flips are heads
 - ◆ example: $E = (O \in \{ \text{THT, TTT} \})$, i.e. the first and third flips are tails
- If coin is fair, then probabilities of outcomes are equal
$$p(\text{HHH}) = p(\text{HHT}) = p(\text{HTH}) = p(\text{HTT}) =$$
$$p(\text{THH}) = p(\text{THT}) = p(\text{TTH}) = p(\text{TTT}) = 1 / 8$$
 - ◆ example: probability of event $E = (\text{exactly two heads})$ is
$$p(\text{HHT}) + p(\text{HTH}) + p(\text{THH}) = 3 / 8$$

Example of continuous probability space

Height of a randomly chosen Indian male

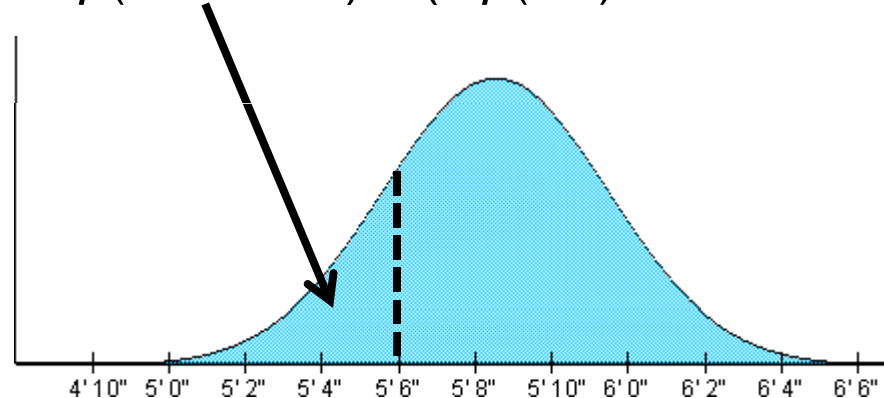
- Infinite number of possible outcomes: O has some single value in range 2 feet to 8 feet
- Infinite number of possible events
 - ◆ example: $E = (O \mid O < 5.5 \text{ feet})$, i.e. individual chosen is less than 5.5 feet tall
- Probabilities of outcomes are not equal, and are described by a continuous function, $p(O)$



Example of continuous probability space

Height of a randomly chosen Indian male

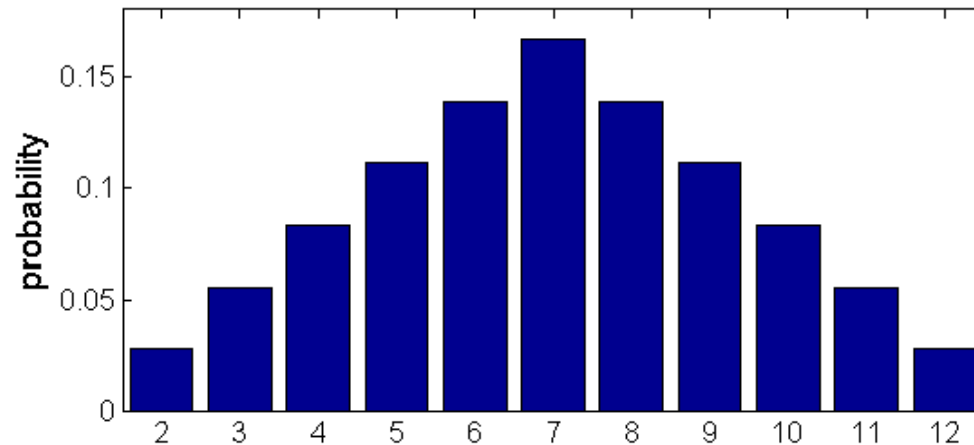
- Probabilities of outcomes O are not equal, and are described by a continuous function, $p(O)$
- $p(O)$ is a *relative*, not an *absolute* probability
 - ◆ $p(O)$ for any particular O is zero
 - ◆ $\int p(O)$ from $O = -\infty$ to ∞ (i.e. area under curve) is 1
 - ◆ example: $p(O = 5'8") > p(O = 6'2")$
 - ◆ example: $p(O < 5'6") = (\int p(O) \text{ from } O = -\infty \text{ to } 5'6") \approx 0.25$



Probability distributions

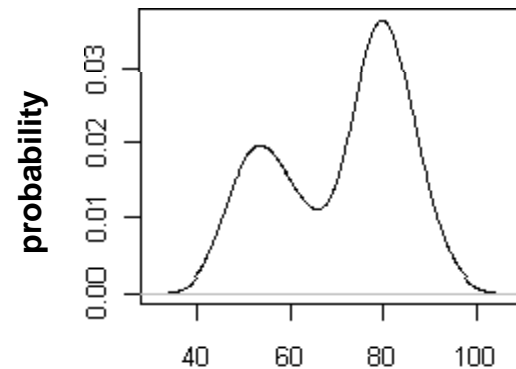
- Discrete: *probability mass function (pmf)*

example:
sum of two
fair dice



- Continuous: *probability density function (pdf)*

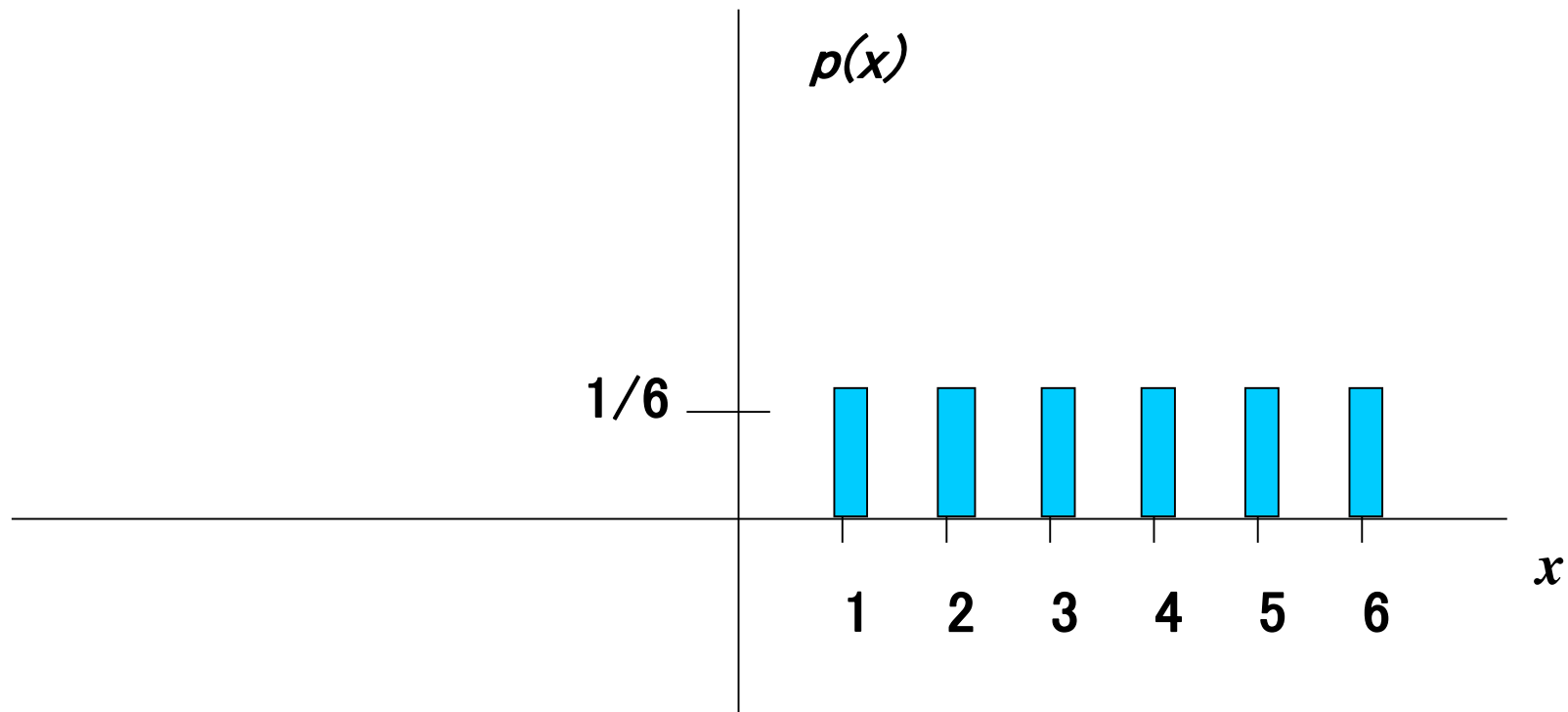
example:
waiting time between
eruptions of Old Faithful
(minutes)



Probability Distribution functions

- A probability function maps the possible values of x against their respective probabilities of occurrence, $p(x)$
 - $p(x)$ is a number from 0 to 1.0.
 - The area under a probability function is always 1.
-

Discrete example: roll of a die



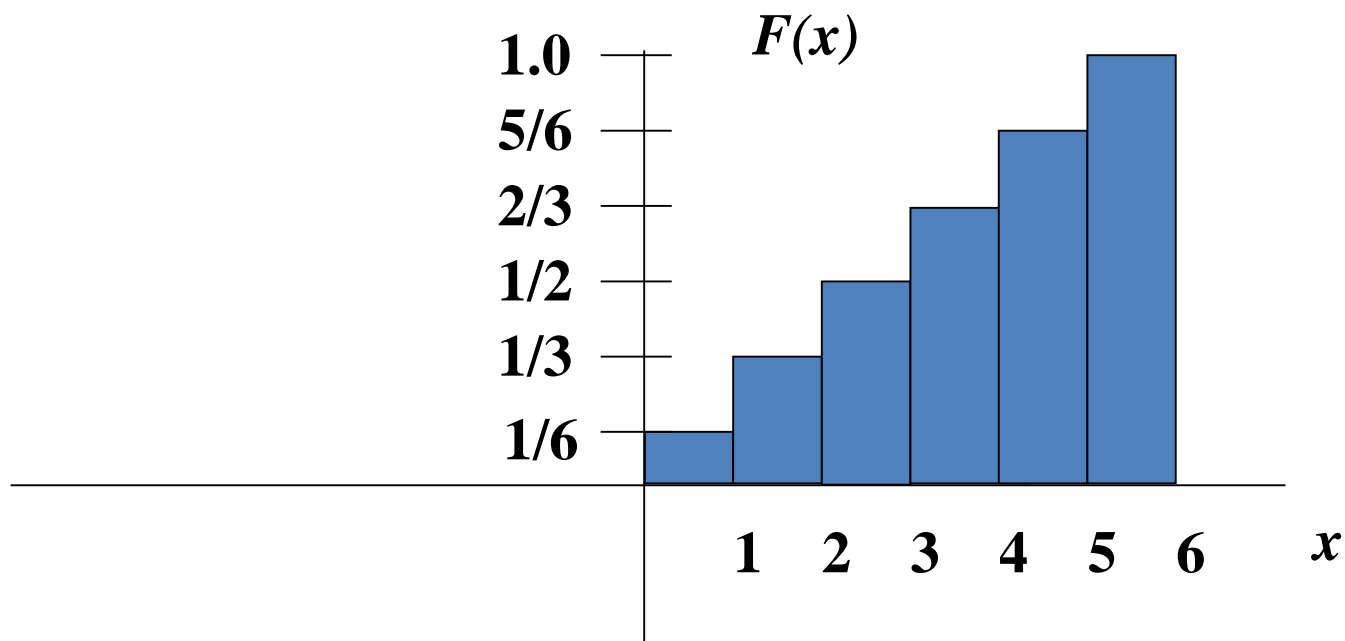
$$\sum_{\text{all } x} P(x) = 1$$

Probability Mass Function (pmf)

x	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	<u>$p(x=6)=1/6$</u>

1.0

Cumulative distribution function (CDF)



Cumulative distribution function

x	$P(x \leq A)$
1	$P(x \leq 1) = 1/6$
2	$P(x \leq 2) = 2/6$
3	$P(x \leq 3) = 3/6$
4	$P(x \leq 4) = 4/6$
5	$P(x \leq 5) = 5/6$
6	$P(x \leq 6) = 6/6$

Examples

1. What's the probability that you roll a 3 or less?

$$P(x \leq 3) = 1/2$$

2. What's the probability that you roll a 5 or higher?

$$P(x \geq 5) = 1 - P(x \leq 4) = 1 - 2/3 = 1/3$$

Practice Problem

Which of the following are probability functions?

- a. $f(x) = .25$ for $x = 9, 10, 11, 12$
 - b. $f(x) = (3-x)/2$ for $x = 1, 2, 3, 4$
 - c. $f(x) = (x^2 + x + 1)/25$ for $x = 0, 1, 2, 3$
-

Answer (a)

a. $f(x) = .25$ for $x = 9, 10, 11, 12$

x	$f(x)$
9	.25
10	.25
11	.25
12	<u>.25</u>

1.0

Yes, probability
function!

Answer (b)

b. $f(x) = (3-x)/2$ for $x=1,2,3,4$

x	$f(x)$
1	$(3-1)/2=1.0$
2	$(3-2)/2=.5$
3	$(3-3)/2=0$
4	$(3-4)/2=-.5$

Though this sums to 1, you can't have a negative probability; therefore, it's not a probability function.

Answer (c)

c. $f(x) = (x^2 + x + 1)/25$ for $x=0,1,2,3$

x	f(x)
0	1/25
1	3/25
2	7/25
3	<u>13/25</u>

24/25

Doesn't sum to 1. Thus,
it's not a probability
function.

Random variables

- A random variable X is a function that associates a number x with each outcome O of a process
 - Common notation: $X(O) = x$, or just $X = x$
- Basically a way to redefine (usually simplify) a probability space to a new probability space
 - X must obey axioms of probability (over the possible values of x)
 - X can be discrete or continuous
- Example: X = number of heads in three flips of a coin
 - Possible values of X are 0, 1, 2, 3
 - $p(X = 0) = p(X = 3) = 1/8$ $p(X = 1) = p(X = 2) = 3/8$
 - Size of space (number of “outcomes”) reduced from 8 to 4
- Example: X = average height of five randomly chosen Indian men
 - Size of space unchanged (X can range from 2 feet to 8 feet), but pdf of X different than for single man

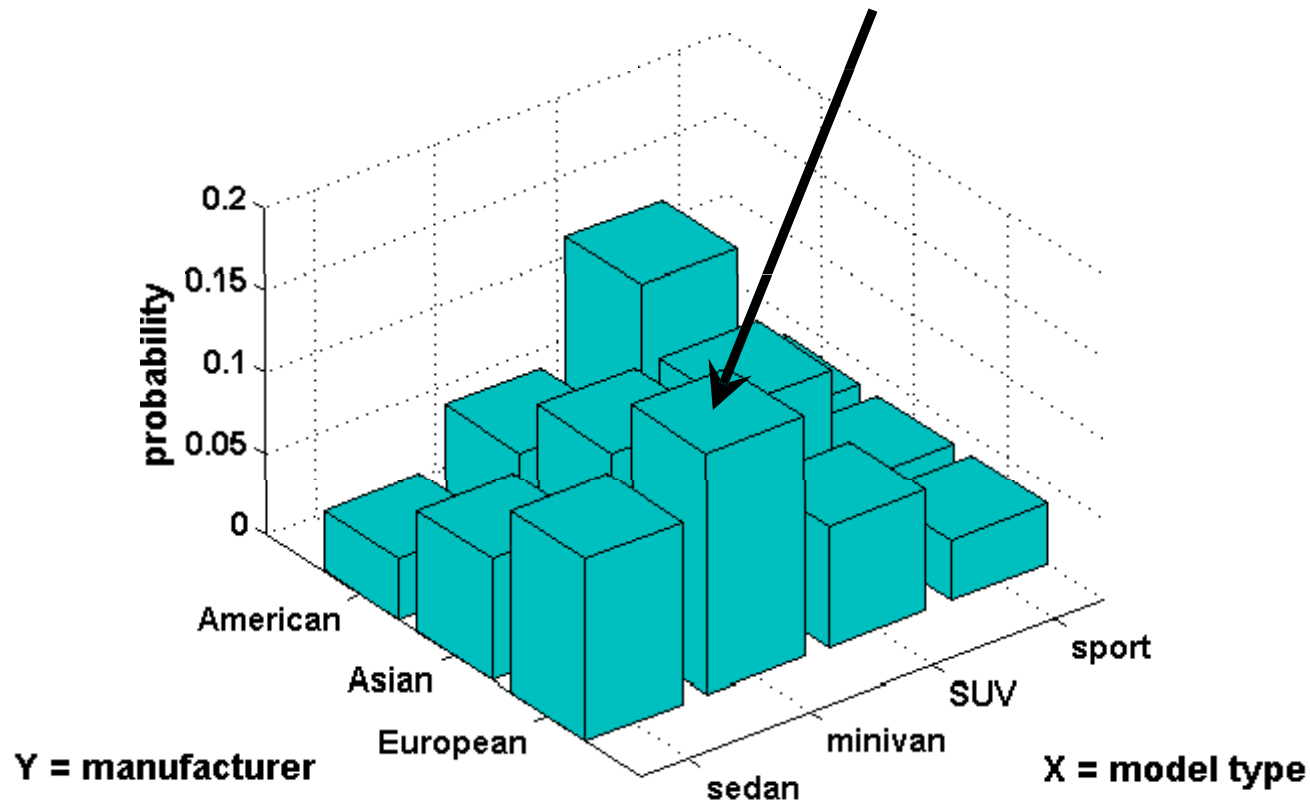
Multivariate probability distributions

- Scenario
 - Several random processes occur (doesn't matter whether in parallel or in sequence)
 - Want to know probabilities for each possible combination of outcomes
- Can describe as *joint probability* of several random variables
 - Example: two processes whose outcomes are represented by random variables X and Y . Probability that process X has outcome x and process Y has outcome y is denoted as

$$p(X = x, Y = y)$$

Example of multivariate distribution

joint probability: $p(X = \text{minivan}, Y = \text{European}) = 0.1481$



Multivariate probability distributions

- *Marginal* probability

- Probability distribution of a single variable in a joint distribution
- Example: two random variables X and Y :

$$p(X = x) = \sum_{b=\text{all values of } Y} p(X = x, Y = b)$$

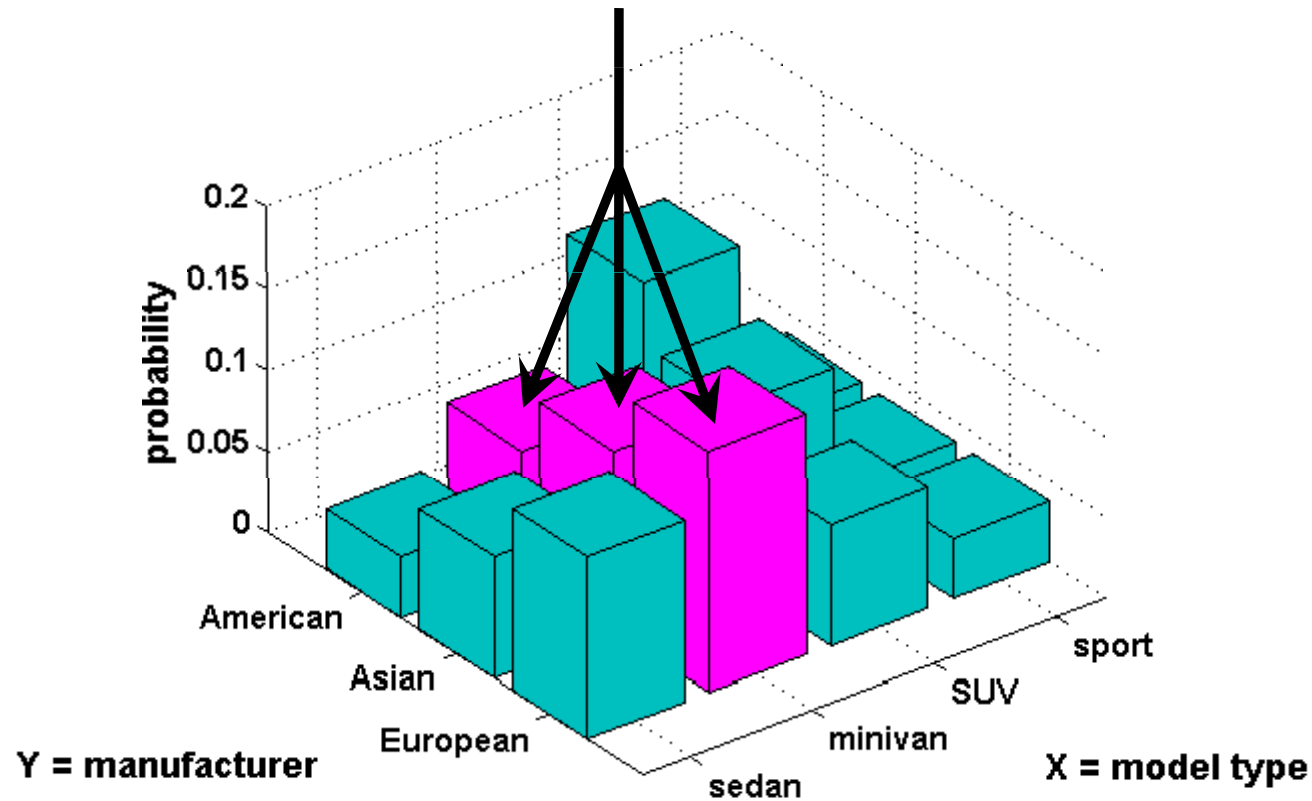
- *Conditional* probability

- Probability distribution of one variable *given* that another variable takes a certain value
- Example: two random variables X and Y :

$$p(X = x \mid Y = y) = p(X = x, Y = y) / p(Y = y)$$

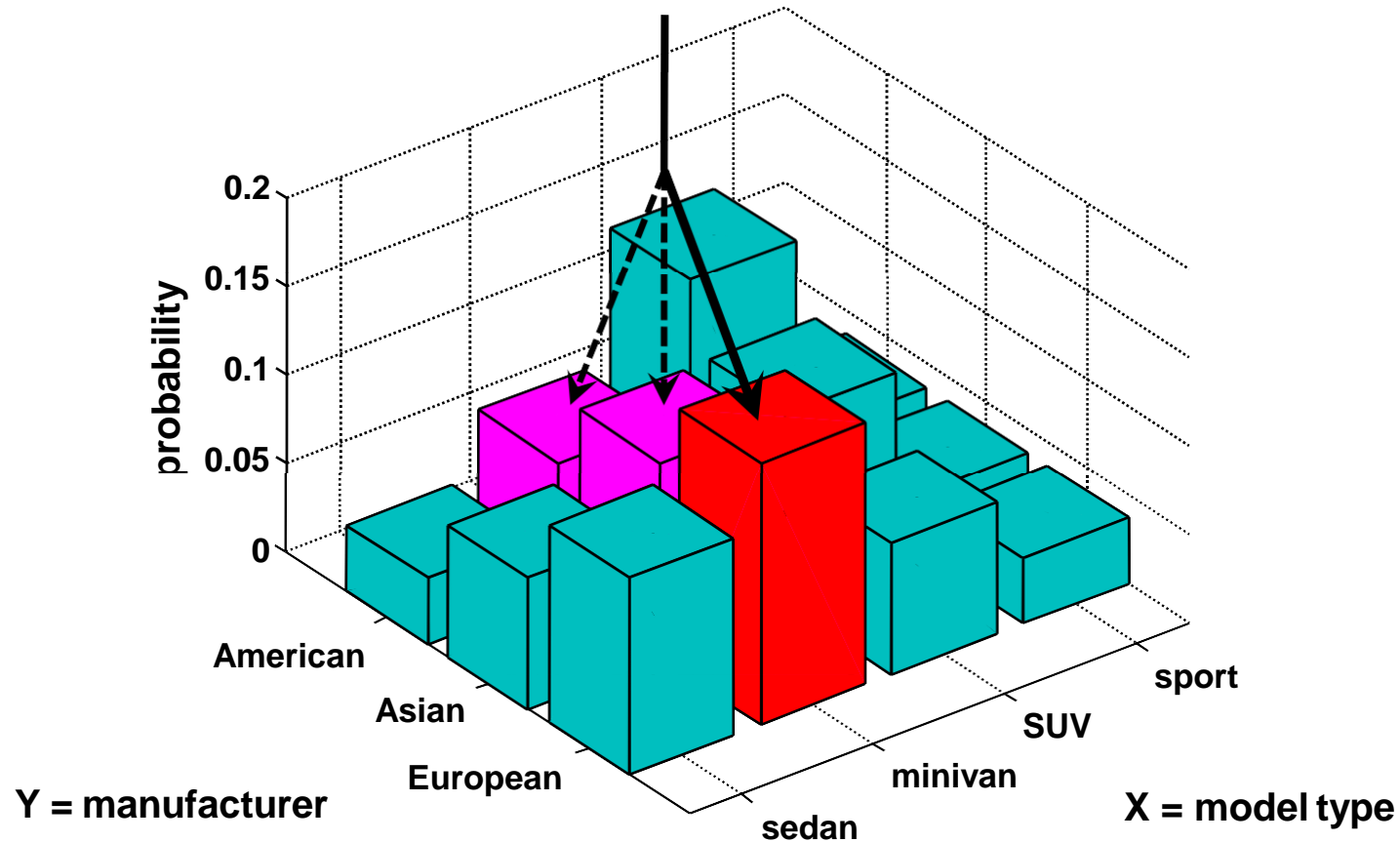
Example of marginal probability

marginal probability: $p(X = \text{minivan}) = 0.0741 + 0.1111 + 0.1481 = 0.3333$



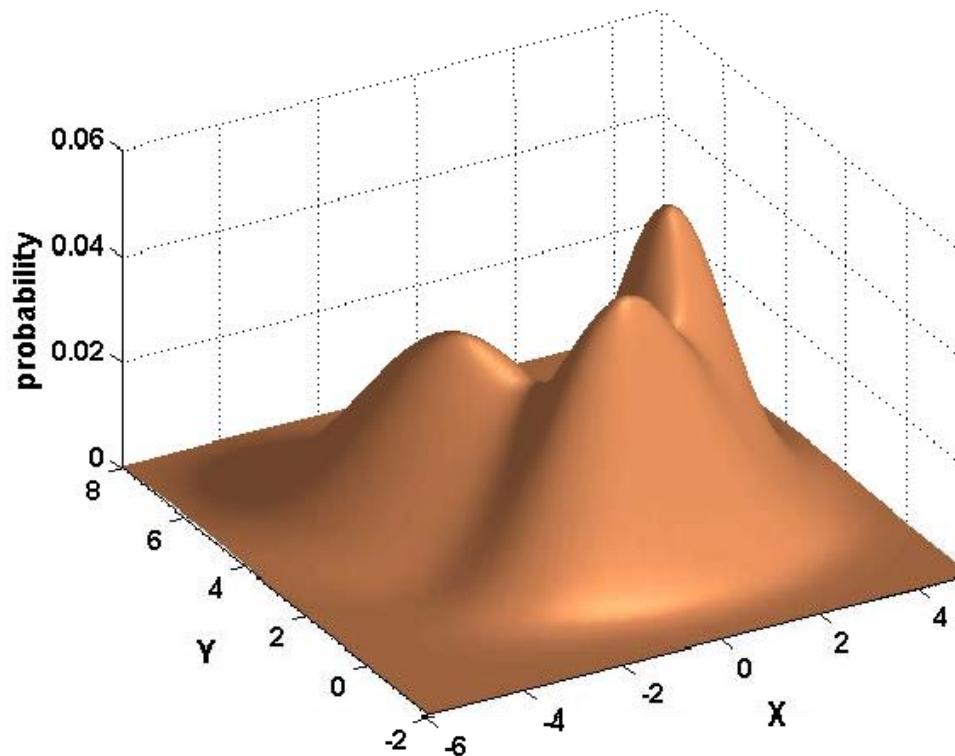
Example of conditional probability

conditional probability: $p(Y = \text{European} \mid X = \text{minivan}) = 0.1481 / (0.0741 + 0.1111 + 0.1481) = 0.4433$



Continuous multivariate distribution

- Same concepts of joint, marginal, and conditional probabilities apply (except use integrals)
- Example: three-component Gaussian mixture in two dimensions



Expected value

- A value calculated before the experiment.
- Consider a discrete random variable X , with possible values $x = x_1, x_2, \dots, x_n$
- Probabilities $p(X = x_i)$ that X takes on the various values of x_i
- A function $y_i = f(x_i)$ defined on X
- The *expected value* of f is the probability-weighted “average” value of $f(x_i)$:

$$E(f) = \sum_i p(x_i) \cdot f(x_i)$$

Expected value is used when we want to calculate the mean of a probability distribution.

E.g.

- Let X represent the outcome of a roll of an unbiased six-sided die.
 - The possible values for X are 1, 2, 3, 4, 5, and 6, each having the probability of occurrence of $1/6$.
 - The expectation value (or expected value) of X is:
 - $$E(X) = \frac{1}{6} * 1 + \frac{1}{6} * 2 + \frac{1}{6} * 3 + \frac{1}{6} * 4 + \frac{1}{6} * 5 + \frac{1}{6} * 6 = 3.5$$
-

Practice questions

Q1. Suppose an individual thinks that if they quit their job and work for themselves that there is a 60% chance they could earn \$20,000 in their first year, a 30% chance they could earn \$60,000, and a 10% chance they would earn \$0. Calculate the expected value for their income in the first year of entrepreneurship.

Solution: Expected value = $0.6 * \$20,000 + 0.3 * \$60,000 + 0.1 * \$0 = \$30,000$

Q2. You are a financial analyst in a development company. Your manager just asked you to assess the viability of future development projects and select the most promising one. According to estimates, Project A, upon completion, shows a probability of 0.4 to achieve a value of \$2 million and a probability of 0.6 to achieve a value of \$500,000. Project B shows a probability of 0.3 to be valued at \$3 million and a probability of 0.7 to be valued at \$200,000 upon completion. Which project will be the most promising to select?

Practice questions

$$\text{EV (Project A)} = [0.4 \times \$2,000,000] + [0.6 \times \$500,000] = \$1,100,000$$

$$\text{EV (Project B)} = [0.3 \times \$3,000,000] + [0.7 \times \$200,000] = \$1,040,000$$

The EV of Project A is greater than the EV of Project B. Therefore, your company should select Project A.

Mean

- A calculated value after experiment. Observed value are required.
- Suppose that in a sequence of ten rolls of the die, if the outcomes are 5, 2, 6, 2, 2, 1, 2, 3, 6, 1, then the average (arithmetic mean) value is calculated using mean.
- Which is defined using following formula.
- $\mu = \frac{1}{N} \sum_{I=1}^n x_i$
- $mean = \frac{5+2+6+2+2+1+2+3+6+1}{10} = 3.0$

Mean is typically used when we want to calculate the average value of a given sample

Variance of a discrete random variable

- The variance of a random variable X is defined by
- $Var(X) = E[(X - \mu)^2] = \sum_i (x_i - \mu)^2 \cdot p(x_i), \mu = E(X)$.
- Average value of squared deviation of $X = x_i$ from mean μ , taking into account probability of the various x_i
 - Most common measure of “spread” of a distribution
 - σ is the *standard deviation* $= \sqrt{Var(X)}$
 - Compare to formula for variance of an actual sample

Example

Consider the rolling of a fair six-sided die, with X the number on the uppermost face. We know that the pf of X is

$$p_X(x) = \frac{1}{6}, \quad x = 1, 2, 3, 4, 5, 6,$$

and that $\mu_X = 3.5$. The variance of X is given by

$$\begin{aligned} \text{var}(X) &= E[(X - \mu_X)^2] \\ &= \sum (x - \mu_X)^2 p_X(x) \\ &= \sum (x - \mu_X)^2 \frac{1}{6} \\ &= \frac{1}{6} \left((1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2 \right) \\ &= \frac{1}{6} \times 17.5 \\ &= \frac{35}{12} \approx 2.9167. \end{aligned}$$

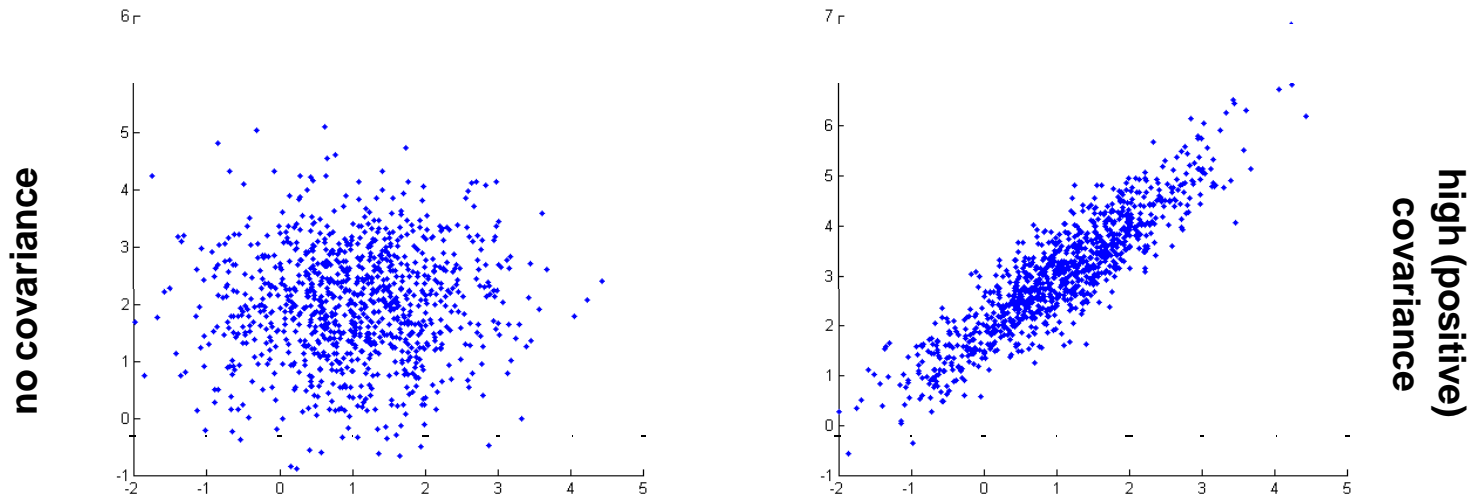
Hence, the standard deviation of X is $\sigma_X = \sqrt{\frac{35}{12}} \approx 1.7078$.

Common forms of expected value (3)

- Covariance

$$f(x_i) = (x_i - \mu_x), \quad g(y_i) = (y_i - \mu_y) \Rightarrow \\ \text{cov}(x, y) = \sum_i p(x_i, y_i) \cdot (x_i - \mu_x) \cdot (y_i - \mu_y)$$

- Measures tendency for x and y to deviate from their means in same (or opposite) directions at same time



- Compare to formula for covariance of actual samples

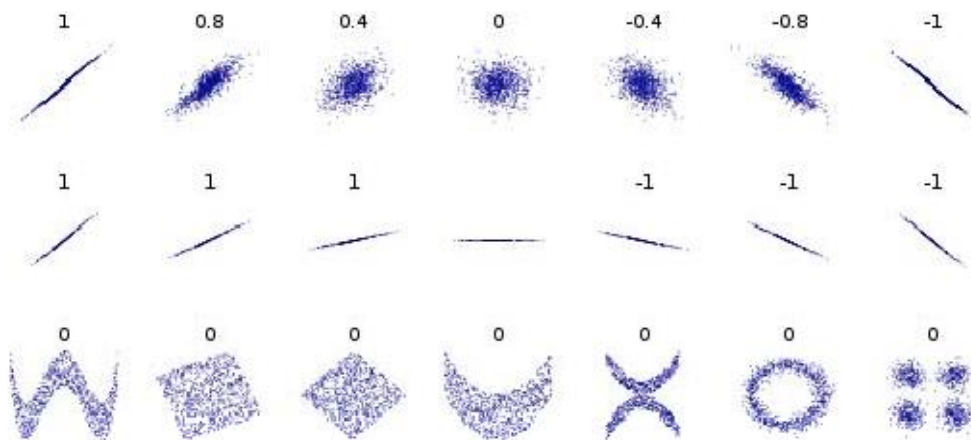
$$\text{cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)$$

Correlation

- Pearson's correlation coefficient is covariance normalized by the standard deviations of the two variables

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

- Always lies in range -1 to 1
- Only reflects *linear dependence* between variables



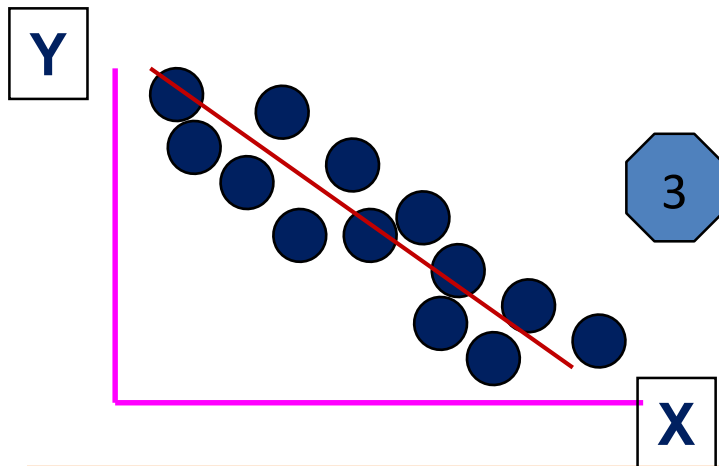
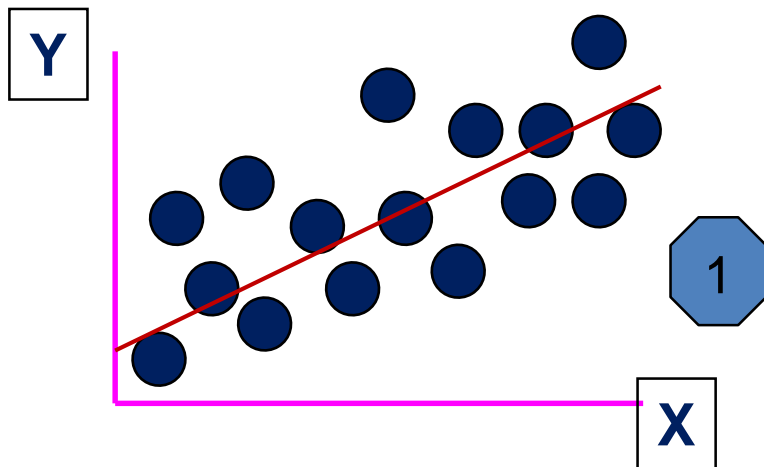
Linear dependence
with noise

Linear dependence
without noise

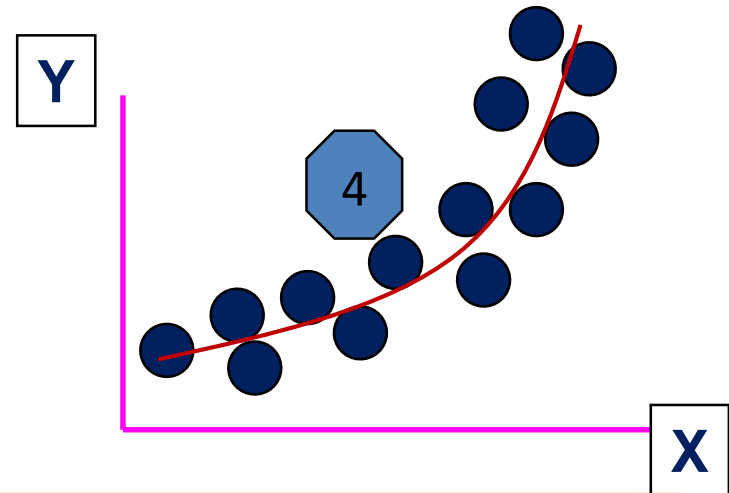
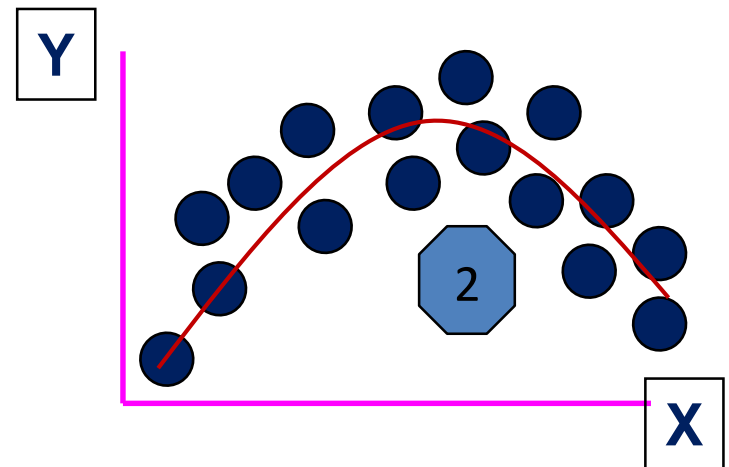
Various nonlinear
dependencies

Types of Relationship

Linear relationships

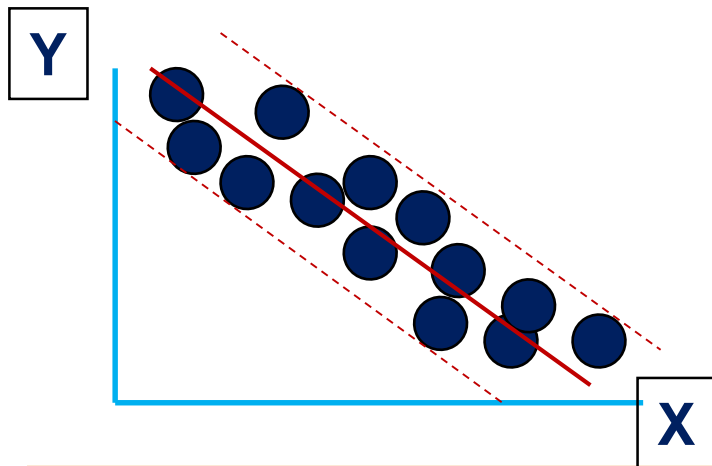
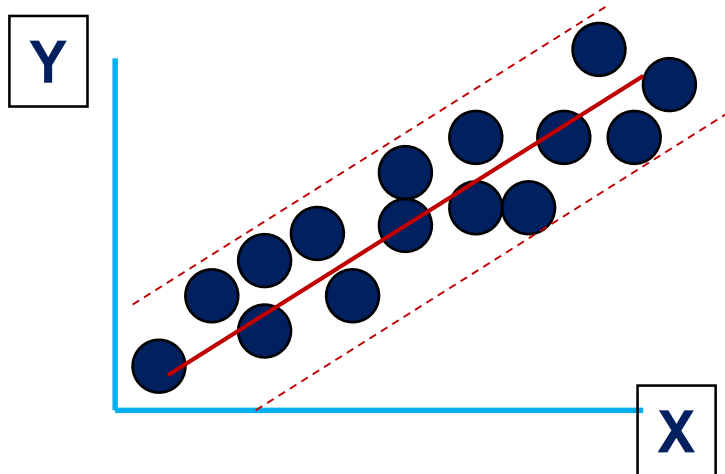


Curvilinear relationships

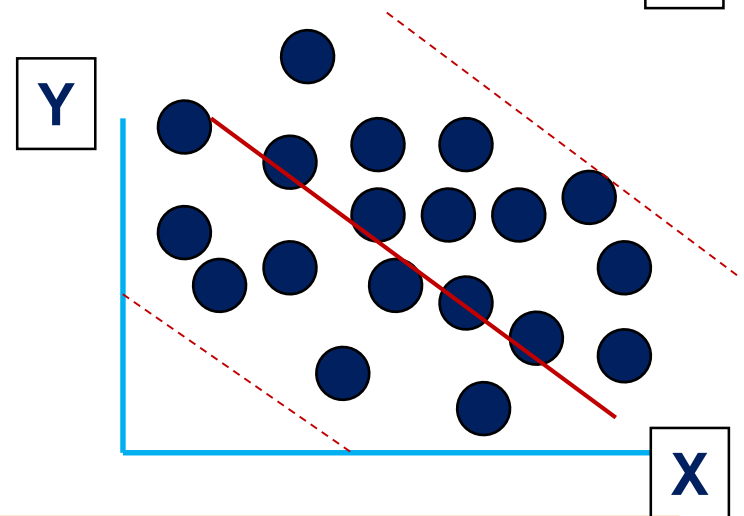
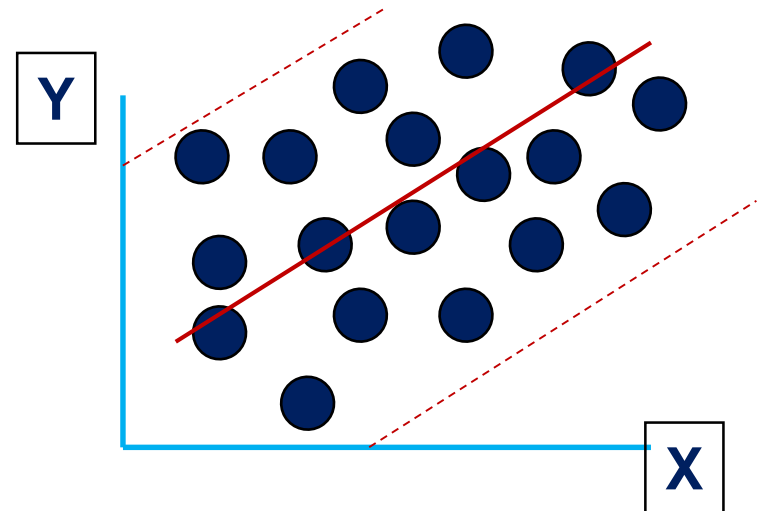


Types of Relationship...

Strong relationships

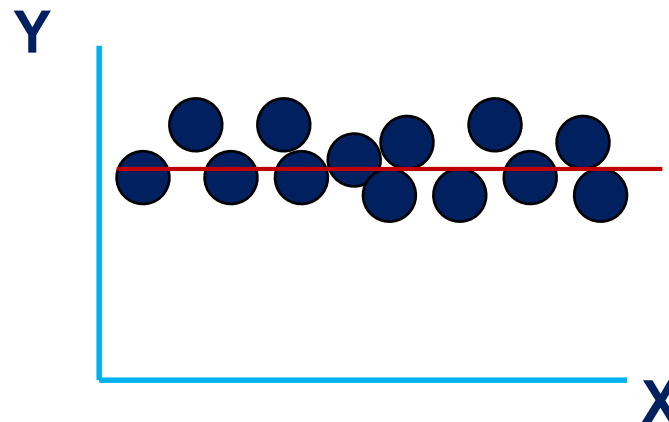
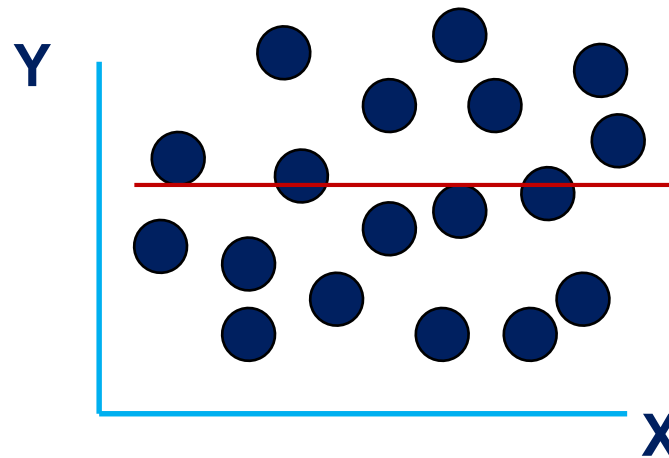


Weak relationships



Types of Relationship...

No relationship



$$\sigma^2_x = \frac{\Sigma(X - \bar{X})^2}{N} = \frac{\Sigma X^2}{N} - \bar{X}^2$$

Example

$$\text{Correlation } (X, Y) = \frac{\frac{1}{N-1} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \cdot \sigma_y}$$

$$\sigma^2_y = \frac{\Sigma(Y - \bar{Y})^2}{N} = \frac{\Sigma Y^2}{N} - \bar{Y}^2$$

$$r = \frac{42}{7 \sqrt{\frac{112}{7}} \sqrt{\frac{38}{7}}} = 0.644$$

No. of years of schooling of farmers	Annual yield per acre in '000 (Rs)
0	4
2	4
4	6
6	10
8	10
10	8
12	7

Calculation of r between years of schooling of farmers and annual yield

Years of Education (X)	$(X - \bar{X})$	$(X - \bar{X})^2$	Annual yield per acre in '000 Rs (Y)	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
0	-6	36	4	-3	9	18
2	-4	16	4	-3	9	12
4	-2	4	6	-1	1	2
6	0	0	10	3	9	0
8	2	4	10	3	9	6
10	4	16	8	1	1	4
12	6	36	7	0	0	0
$\Sigma X=42$	$\Sigma (X - \bar{X})^2=112$		$\Sigma Y=49$	$\Sigma (Y - \bar{Y})^2=38$		$\Sigma (X - \bar{X})(Y - \bar{Y})=42$

Practice question

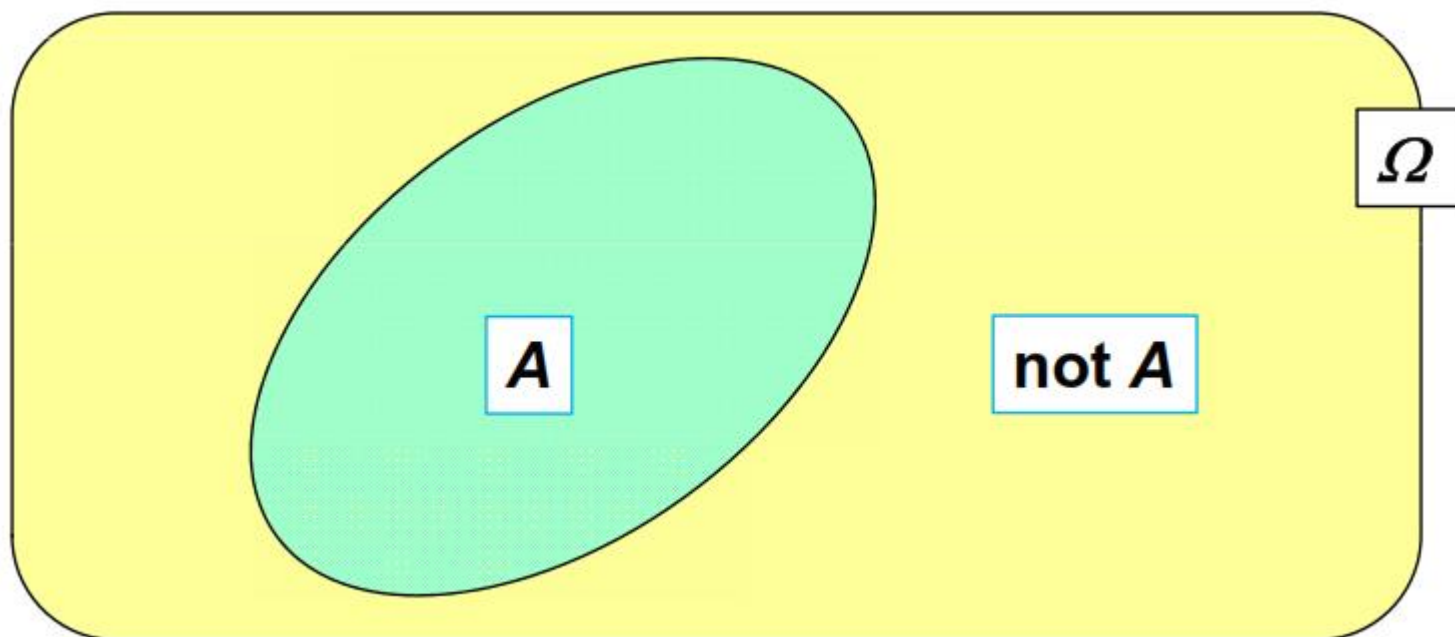
Q1. Calculate and analyze the correlation coefficient between the number of study hours and the number of sleeping hours of different students using the following data.

Number of Study Hours	2	4	6	8	10
Number of Sleeping Hours	10	9	8	7	6

Complement rule

Given: event A , which can occur or not

$$p(\text{not } A) = 1 - p(A)$$



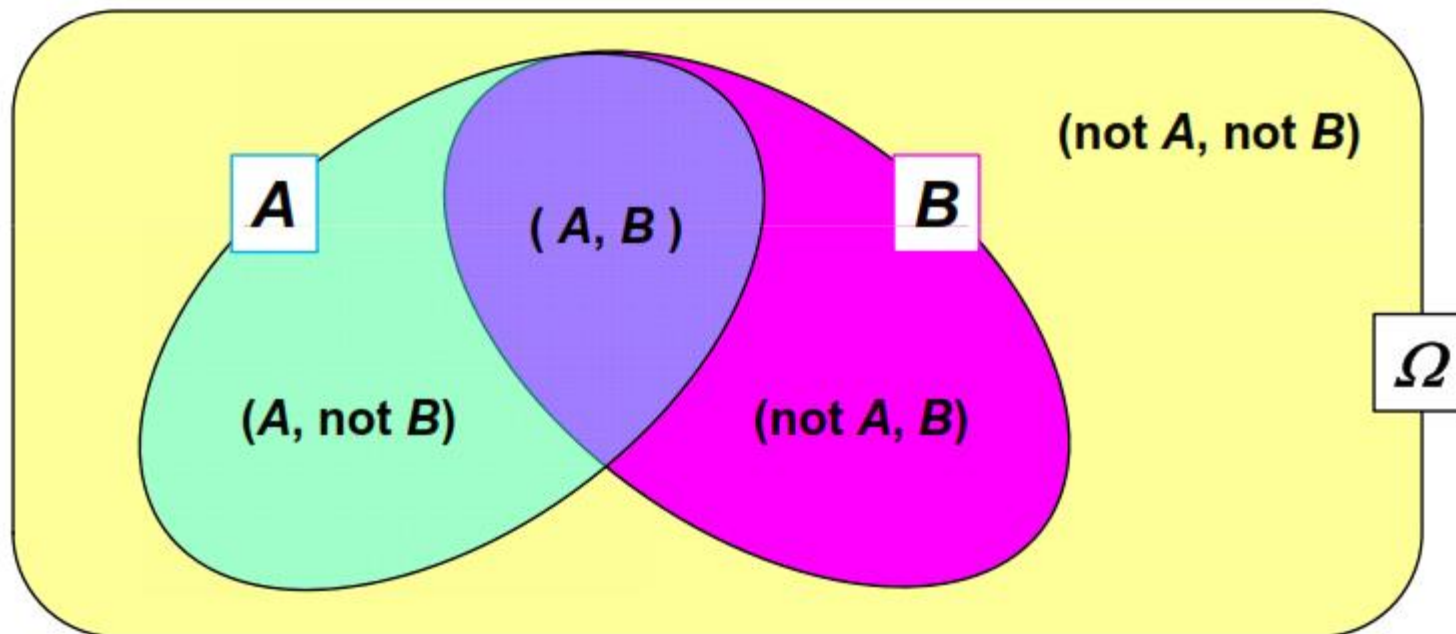
areas represent relative probabilities

Product rule

Given: events A and B , which can co-occur (or not)

$$p(A, B) = p(A | B) \cdot p(B)$$

(same expression given previously to define conditional probability)



areas represent relative probabilities

Example of product rule

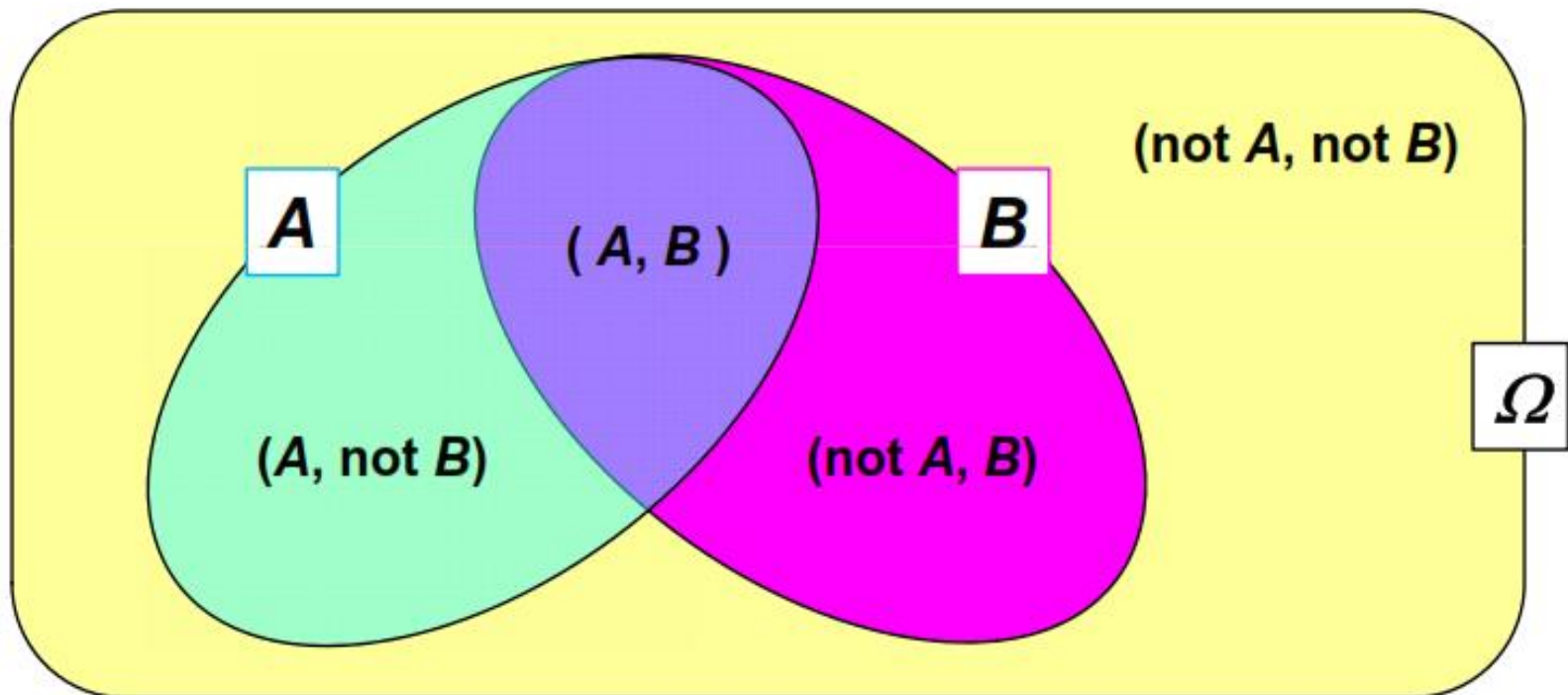
- Probability that a man has white hair (event A) and is over 65 (event B)
 - $p(B) = 0.18$
 - $p(A | B) = 0.78$
 - $p(A, B) = p(A | B) \cdot p(B) = 0.78 \cdot 0.18 = 0.14$

Rule of total probability

Given: events A and B , which can co-occur (or not)

$$p(A) = p(A, B) + p(A, \text{not } B)$$

(same expression given previously to define marginal probability)

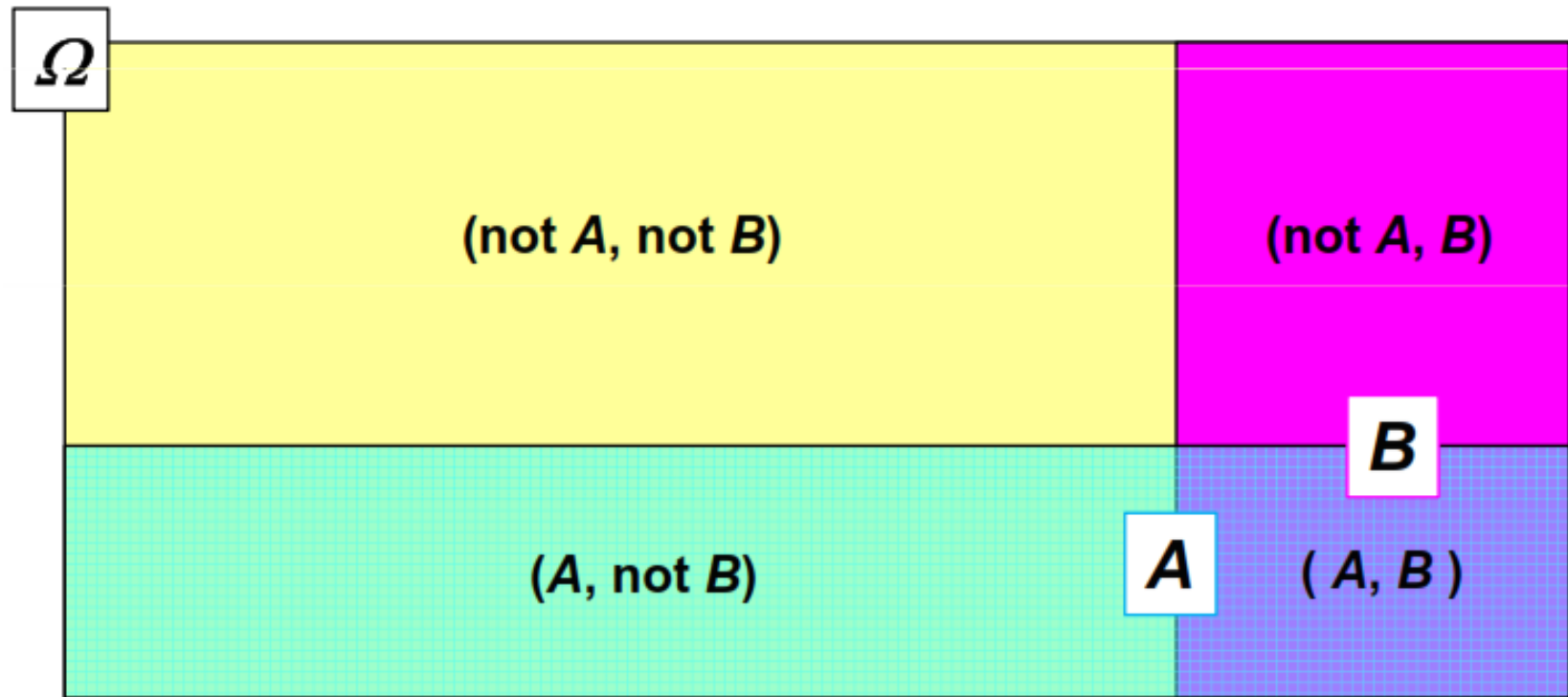


areas represent relative probabilities

Independence

Given: events A and B , which can co-occur (or not)

$$p(A | B) = p(A) \quad \text{or} \quad p(A, B) = p(A) \cdot p(B)$$



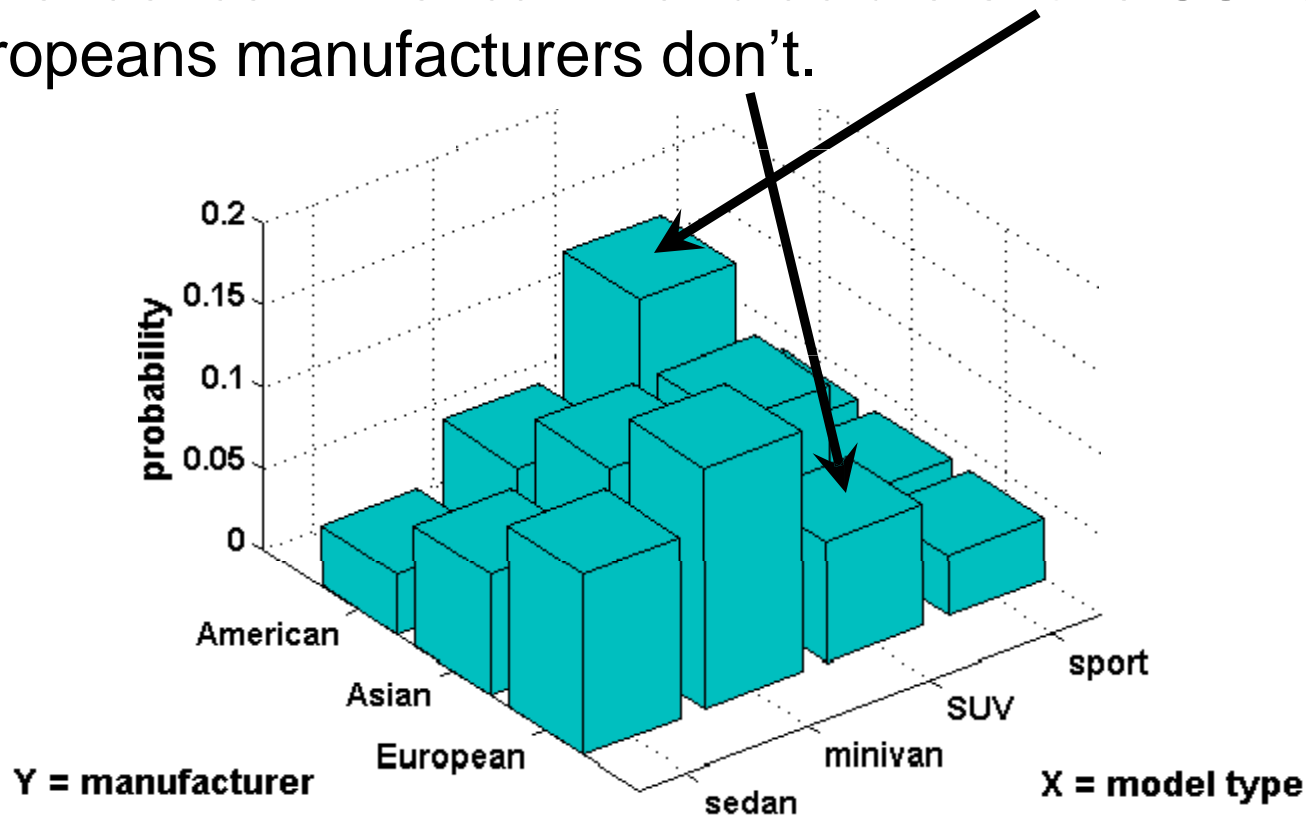
areas represent relative probabilities

Examples of independence / dependence

- Independence:
 - Outcomes on multiple rolls of a die
 - Outcomes on multiple flips of a coin
 - Height of two unrelated individuals
 - Probability of getting a king on successive draws from a deck, if card from each draw is *replaced*
- Dependence:
 - Height of two related individuals
 - Duration of successive eruptions of Old Faithful
 - Probability of getting a king on successive draws from a deck, if card from each draw is *not replaced*

Example of independence vs. dependence

- Independence: All manufacturers have identical product mix. $p(X = x \mid Y = y) = p(X = x)$.
- Dependence: American manufacturers love SUVs, Europeans manufacturers don't.



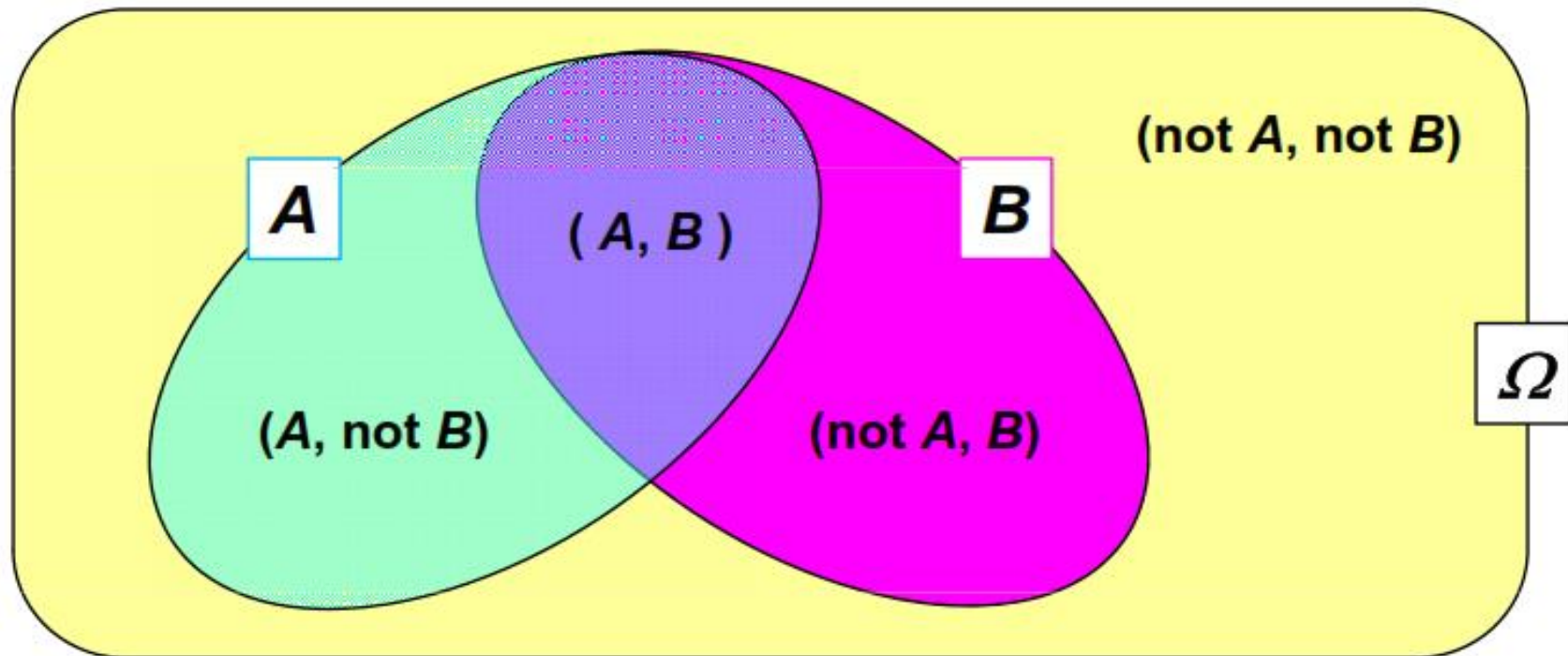
Bayes rule

A way to find conditional probabilities for one variable when conditional probabilities for another variable are known.

$$p(B | A) = p(A | B) \cdot p(B) / p(A)$$

where $p(A) = p(A, B) + p(A, \text{not } B)$

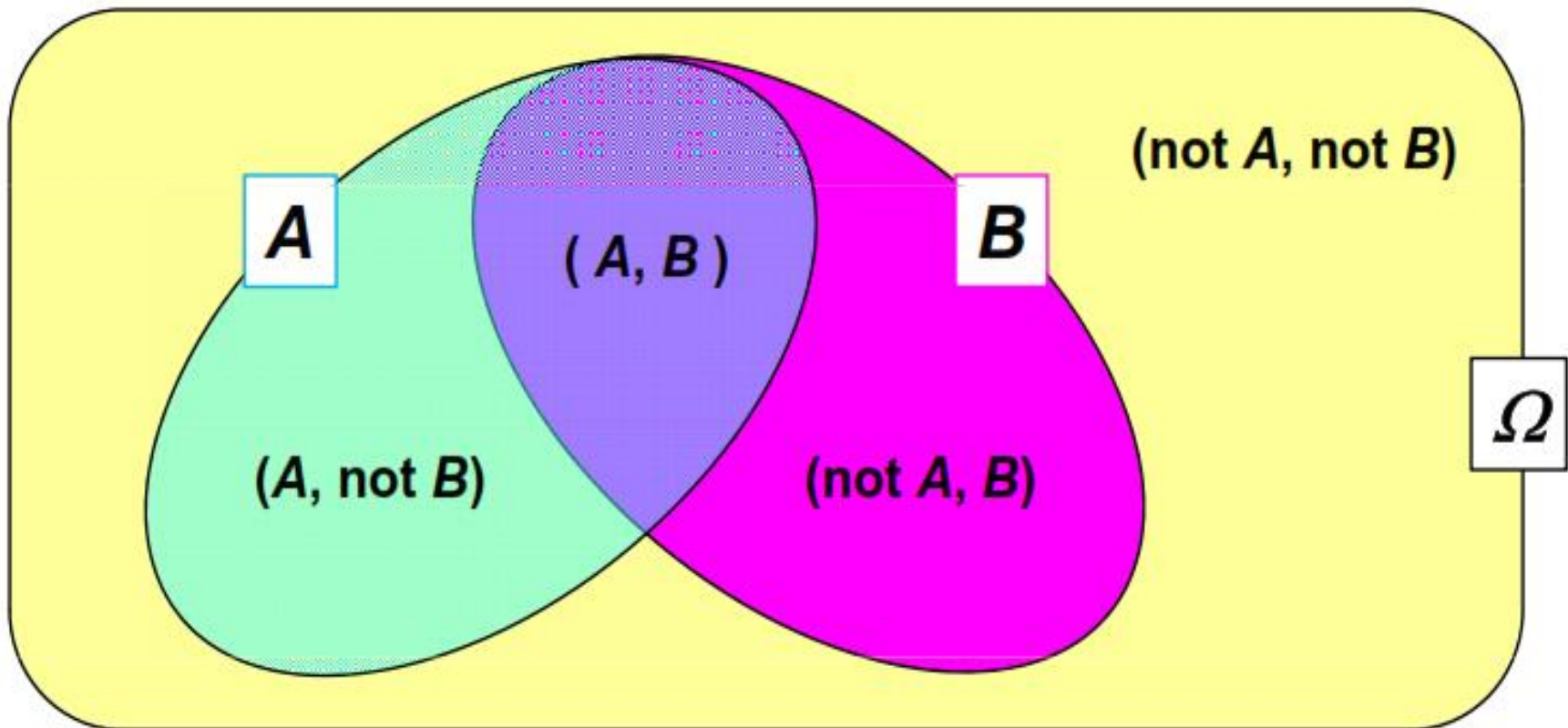
$$p(B \cap A) = p(B | A) \cdot p(A) \text{ if } p(A) > 0$$



Bayes rule

posterior probability \propto likelihood \times prior probability

$$p(B | A) = p(A | B) \cdot p(B) / p(A)$$



Example of Bayes rule

- Marie is getting married tomorrow at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman is forecasting rain for tomorrow. When it actually rains, the weatherman has forecast rain 90% of the time. When it doesn't rain, he has forecast rain 10% of the time. What is the probability it will rain on the day of Marie's wedding?
- **Event A: The weatherman has forecast rain.**
- **Event B: It rains.**
- We know:
 - $p(B) = 5 / 365 = 0.0137$ [It rains 5 days out of the year.]
 - $p(\text{not } B) = 360 / 365 = 0.9863$
 - $p(A | B) = 0.9$ [When it rains, the weatherman has forecast rain 90% of the time.]
 - $p(A | \text{not } B) = 0.1$ [When it does not rain, the weatherman has forecast rain 10% of the time.]

Example of Bayes rule, cont'd.

- We want to know $p(B | A)$, the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes rule:
 1. $p(B | A) = p(A | B) \cdot p(B) / p(A)$
 2. $p(A) = p(A | B) \cdot p(B) + p(A | \text{not } B) \cdot p(\text{not } B) = (0.9)(0.014) + (0.1)(0.986) = 0.111$
 3. $p(B | A) = (0.9)(0.0137) / 0.111 = 0.111$
- The result seems unintuitive but is correct. Even when the weatherman predicts rain, it only rains only about 11% of the time. Despite the weatherman's gloomy prediction, it is unlikely Marie will get rained on at her wedding.

Probabilities: when to add, when to multiply

- **ADD:** When you want to allow for occurrence of any of several possible outcomes of a *single* process. Comparable to logical OR.
- **MULTIPLY:** When you want to allow for simultaneous occurrence of *particular* outcomes from *more than one* process. Comparable to logical AND.
 - But only if the processes are *independent*.

Practice questions

Q1. 3 persons A, B, C independently fire at a target. What is the probability that (i) Exactly one of them hits the target, (ii) At least one of them hits the target? Given: Probability of hitting the target. $P(A) = 1/6$, $P(B) = 1/4$, $P(C) = 1/3$.

Solution: $P(A)=1/6$, $P(A')=5/6$, $P(B)=1/4$, $P(B')=3/4$,
 $P(C)=1/3$ and $P(C')=2/3$

(i) The probability that exactly one of them hits the target (it means only one will hit the target while others two cannot)

$$P(A)P(B')P(C') + P(A')P(B)P(C') + P(A')P(B')P(C) \\ 1/6 \times 3/4 \times 2/3 + 5/6 \times 1/4 \times 2/3 + 5/6 \times 3/4 \times 1/3 = 31/72$$

(ii) The probability that atleast one of them hits the target will be:

$$1 - P(A')P(B')P(C') = 1 - [(5/6) \times (3/4) \times (2/3)] = 1 - 30/72 = 42/72$$

Practice questions

Q2. A covid test is 99% accurate (both ways). Only 0.3% of the population is covid+. What is the probability that a random person is covid+ given that the person tests+?

Solution: Let person is covid+ is taken as A and person tests+ as B hence we have to find $P(A | B)=?$

By Bayes theorem: $P(A|B)=\frac{P(A)P(B|A)}{P(B)}$

$$P(A | B) = \frac{P(\text{person is covid+}) \times P[(\text{person tests+}) | (\text{person is covid+})]}{P(\text{person tests+})}$$

Remember :

If $p(B | A) = p(A | B) \cdot p(B) / p(A)$ -----equation1

Then $p(A) = p(A | B) \cdot p(B) + p(A | \text{not } B) \cdot p(\text{not } B)$ --equation2

By using equation2

$P(\text{person tests+}) = P[(\text{person tests+}) | (\text{person is covid+})] \times P(\text{person is covid+}) + P[(\text{person tests+}) | (\text{person is not covid+})] \times P(\text{person is not covid+})$

Practice questions

Solution:
$$\frac{0.3\% \times 0.99\%}{(0.3\% \times 0.99\%) + (99.7\% \times 1\%)} = 22.95\%$$

[P(is Covid+) = 0.3% and P(is not covid+) = 99.7%]

Practice questions

Q3. Two dice are rolled. Consider the events $A = \{\text{sum of two dice equals } 3\}$, $B = \{\text{sum of two dice equals } 7\}$, and $C = \{\text{at least one of the dice shows a } 1\}$.

(a) What is $P(A | C)$?

(b) What is $P(B | C)$?

(c) Are A and C independent? What about B and C ?

Solution:

Note that the sample space is $S = \{(i, j) \mid i, j = 1, 2, 3, 4, 5, 6\}$ with each outcome equally likely.

Then, $A = \{(1, 2), (2, 1)\}$

$B = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$

$C = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1)\}$

Hence, $P(A | C) = P(A \cap C) / P(C) = \frac{2/36}{11/36} = 2/11$.

$P(B | C) = P(B \cap C) / P(C) = \frac{2/36}{11/36} = 2/11$.

Note that $P(A) = 2/36 \neq P(A | C)$, so they are not independent. Similarly, $P(B) = 6/36 \neq P(B | C)$, so they are not independent.

Linear algebra applications

- 1) Operations on or between vectors and matrices
- 2) Coordinate transformations
- 3) Dimensionality reduction
- 4) Linear regression
- 5) Solution of linear systems of equations
- 6) Many others

Applications 1) – 4) are directly relevant to this course.

Why vectors and matrices?

- Most common form of data organization for machine learning is a 2D array, where
 - *rows* represent samples (records, items, datapoints)
 - *columns* represent attributes (features, variables)
- Natural to think of each sample as a *vector* of attributes, and whole array as a *matrix*

vector

Refund	Marital Status	Taxable Income	Cheat
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

matrix

Vectors

- Definition: an n -tuple of values (usually real numbers).
 - n referred to as the *dimension* of the vector
 - n can be any positive integer, from 1 to infinity
- Can be written in column form or row form
 - Column form is conventional
 - Vector elements referenced by subscript

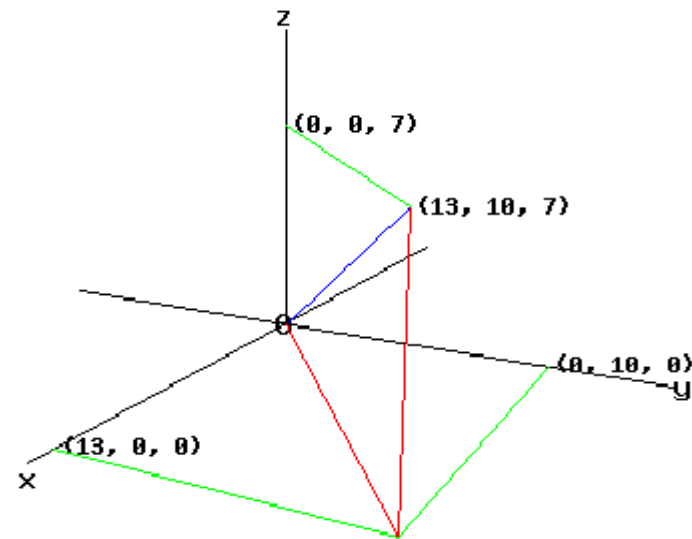
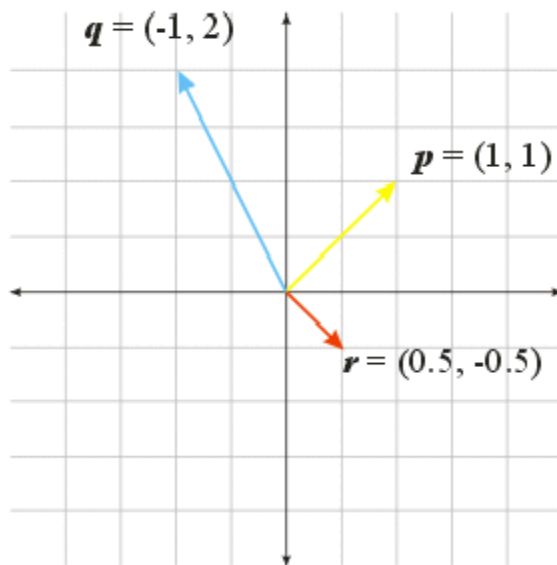
$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$$\mathbf{x}^T = (x_1 \quad \cdots \quad x_n)$$

^T means "transpose"

Vectors

- Can think of a vector as:
 - a point in space *or*
 - a directed line segment with a magnitude and direction



Vector arithmetic

- Addition of two vectors

- add corresponding elements

$$\mathbf{z} = \mathbf{x} + \mathbf{y} = (x_1 + y_1 \quad \text{L} \quad x_n + y_n)^T$$

- result is a vector

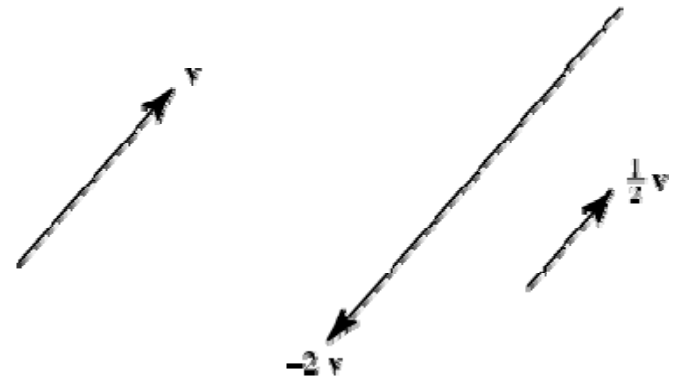


- Scalar multiplication of a vector

- multiply each element by scalar

$$\mathbf{y} = a\mathbf{x} = (ax_1 \quad \text{L} \quad ax_n)^T$$

- result is a vector



Vector arithmetic

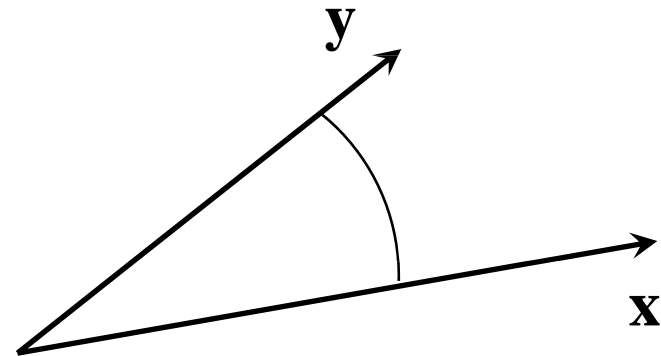
- Dot product of two vectors
 - multiply corresponding elements, then add products

$$a = \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$$

- result is a scalar

- Dot product alternative form

$$a = \mathbf{x} \cdot \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos(\theta)$$



Matrices

- Definition: an $m \times n$ two-dimensional array of values (usually real numbers).
 - m rows
 - n columns
- Matrix referenced by two-element subscript
 - first element in subscript is row
 - second element in subscript is column
 - example: \mathbf{A}_{24} or a_{24} is element in second row, fourth column of \mathbf{A}

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

Matrices

- A vector can be regarded as special case of a matrix, where one of matrix dimensions = 1.
- Matrix *transpose* (denoted T)
 - swap columns and rows
 - ◆ row 1 becomes column 1, etc.
 - $m \times n$ matrix becomes $n \times m$ matrix
 - example:

$$\mathbf{A} = \begin{pmatrix} 2 & 7 & -1 & 0 & 3 \\ 4 & 6 & -3 & 1 & 8 \end{pmatrix} \quad \mathbf{A}^T = \begin{pmatrix} 2 & 4 \\ 7 & 6 \\ -1 & -3 \\ 0 & 1 \\ 3 & 8 \end{pmatrix}$$

Matrix arithmetic

- Addition of two matrices

- matrices must be same size
- add corresponding elements:

$$c_{ij} = a_{ij} + b_{ij}$$

- result is a matrix of same size

$$\mathbf{C} = \mathbf{A} + \mathbf{B} =$$

$$\begin{pmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{pmatrix}$$

- Scalar multiplication of a matrix

- multiply each element by scalar:

$$b_{ij} = d \cdot a_{ij}$$

- result is a matrix of same size

$$\mathbf{B} = d \cdot \mathbf{A} =$$

$$\begin{pmatrix} d \cdot a_{11} & \cdots & d \cdot a_{1n} \\ \vdots & \ddots & \vdots \\ d \cdot a_{m1} & \cdots & d \cdot a_{mn} \end{pmatrix}$$

Matrix arithmetic

- Matrix-matrix multiplication
 - vector-matrix multiplication just a special case

- Multiplication is associative

$$\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}) = (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C}$$

- Multiplication is *not* commutative

$$\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A} \quad (\text{generally})$$

- Transposition rule:

$$(\mathbf{A} \cdot \mathbf{B})^T = \mathbf{B}^T \cdot \mathbf{A}^T$$

Matrix arithmetic

- *RULE*: In any chain of matrix multiplications, the *column* dimension of one matrix in the chain must match the *row* dimension of the *following* matrix in the chain.
- Examples

A 3 x 5

B 5 x 5

C 3 x 1

Right:

A · B · A^T

C^T · A · B

A^T · A · B

C · C^T · A

Wrong:

A · B · A

C · A · B

A · A^T · B

C^T · C · A

Vector projection

- Orthogonal projection of \mathbf{y} onto \mathbf{x}
 - Can take place in any space of dimensionality ≥ 2
 - Unit vector in direction of \mathbf{x} is

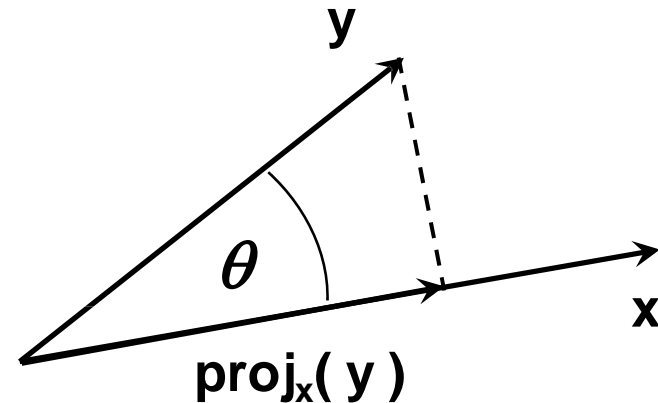
$$\mathbf{x} / \|\mathbf{x}\|$$

- Length of projection of \mathbf{y} in direction of \mathbf{x} is

$$\|\mathbf{y}\| \cos(\theta)$$

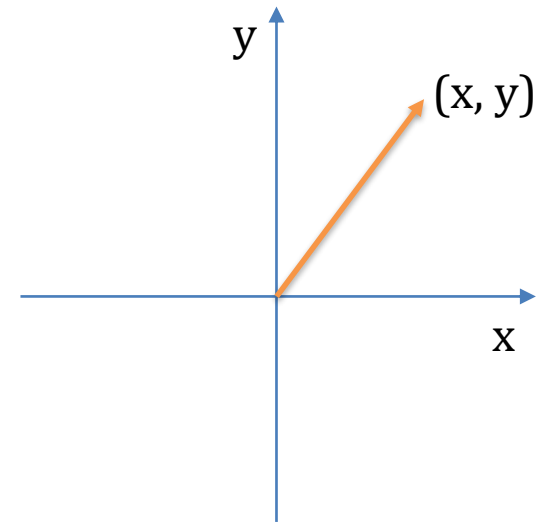
- Orthogonal projection of \mathbf{y} onto \mathbf{x} is the vector

$$\begin{aligned} \mathbf{proj}_x(\mathbf{y}) &= \mathbf{x} \cdot \|\mathbf{y}\| \cdot \cos(\theta) / \|\mathbf{x}\| = \\ &= [(\mathbf{x} \cdot \mathbf{y}) / \|\mathbf{x}\|^2] \mathbf{x} \quad (\text{using dot product alternate form}) \end{aligned}$$



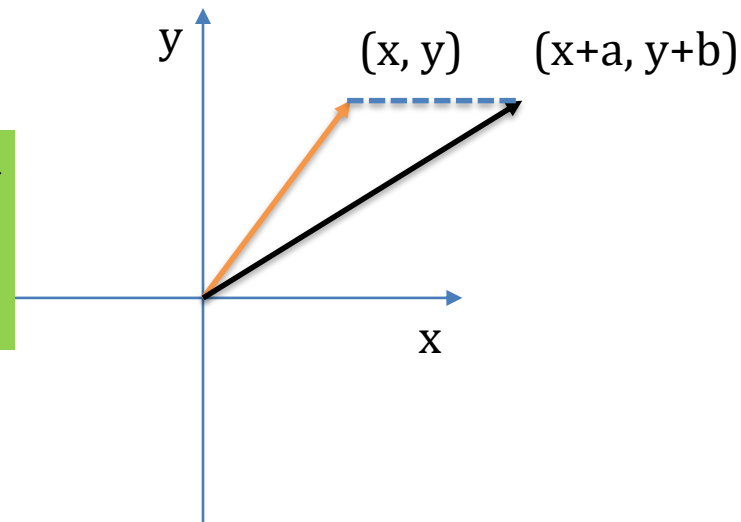
Vector and Matrix Operation

- Consider a vector $[x, y]^T$
- Translation** of vector by (a, b) is
- $$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} x + a \\ y + b \end{bmatrix}$$



E.g. of Translation

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

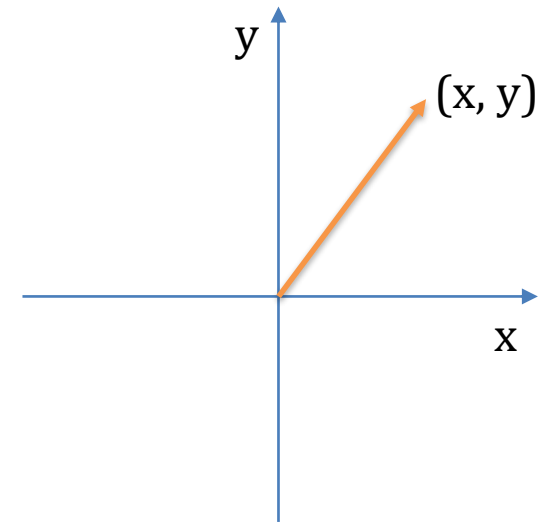


Vector and Matrix Operation...

- Consider a vector $[x, y]^T$

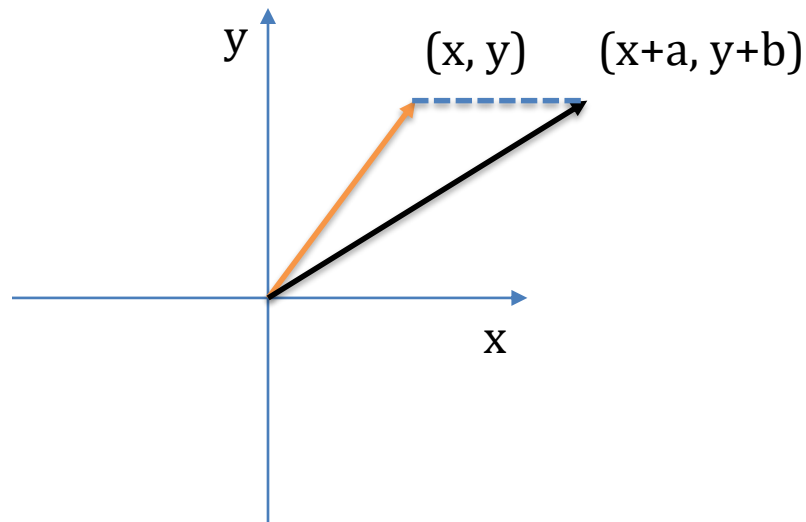
- Scaling** of vector by (a) is

- $$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} x + a \\ y + b \end{bmatrix}$$



E. g. of Scaling

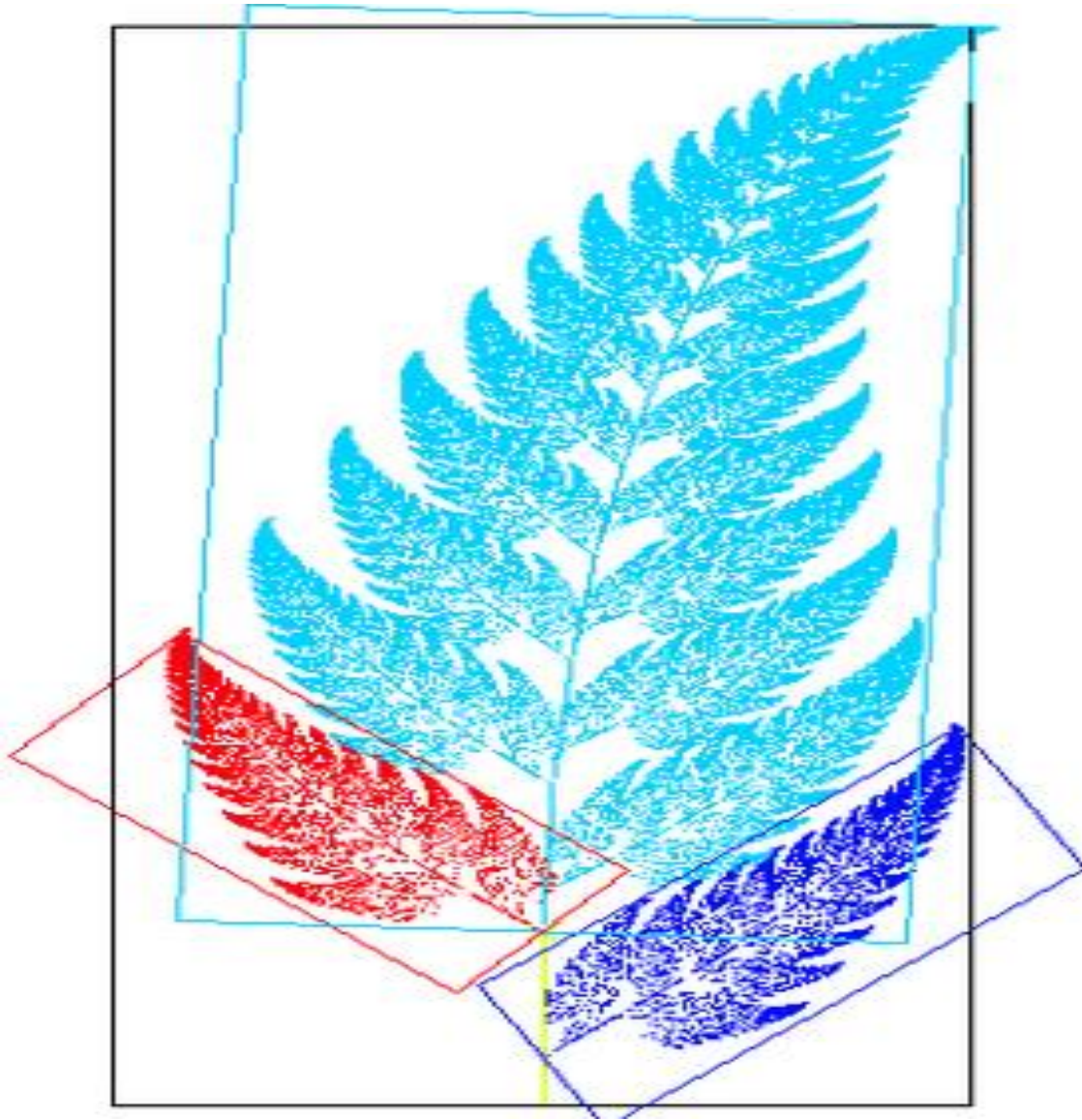
$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} * \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$



Matrix Transformation

- Translation: $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} tx \\ ty \end{bmatrix} = \begin{bmatrix} S_x & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} x \\ y \end{bmatrix}$
 - Rotation: $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} * \begin{bmatrix} x \\ y \end{bmatrix}$
 - Scaling: $\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} S_x & 0 \\ 0 & S_y \end{bmatrix} * \begin{bmatrix} x \\ y \end{bmatrix}$
-

Transformation of an Image



- An image of a fern-like fractal (Barnsley's fern) that exhibits affine self-similarity.
- Each of the leaves of the fern is related to each other leaf by an **affine transformation**.
- For instance, the **red leaf** can be **transformed** into both the **dark blue** leaf and any of the **light blue** leaves by a combination of **reflection**, **rotation**, **scaling**, and **translation**.

Optimization theory topics

- Maximum likelihood
- Expectation maximization
- Gradient descent