Machine Learning

Lesson 03. Intro to Math of Machine Learning

Kirill Svyatov

Ulyanovsk State Technical University,

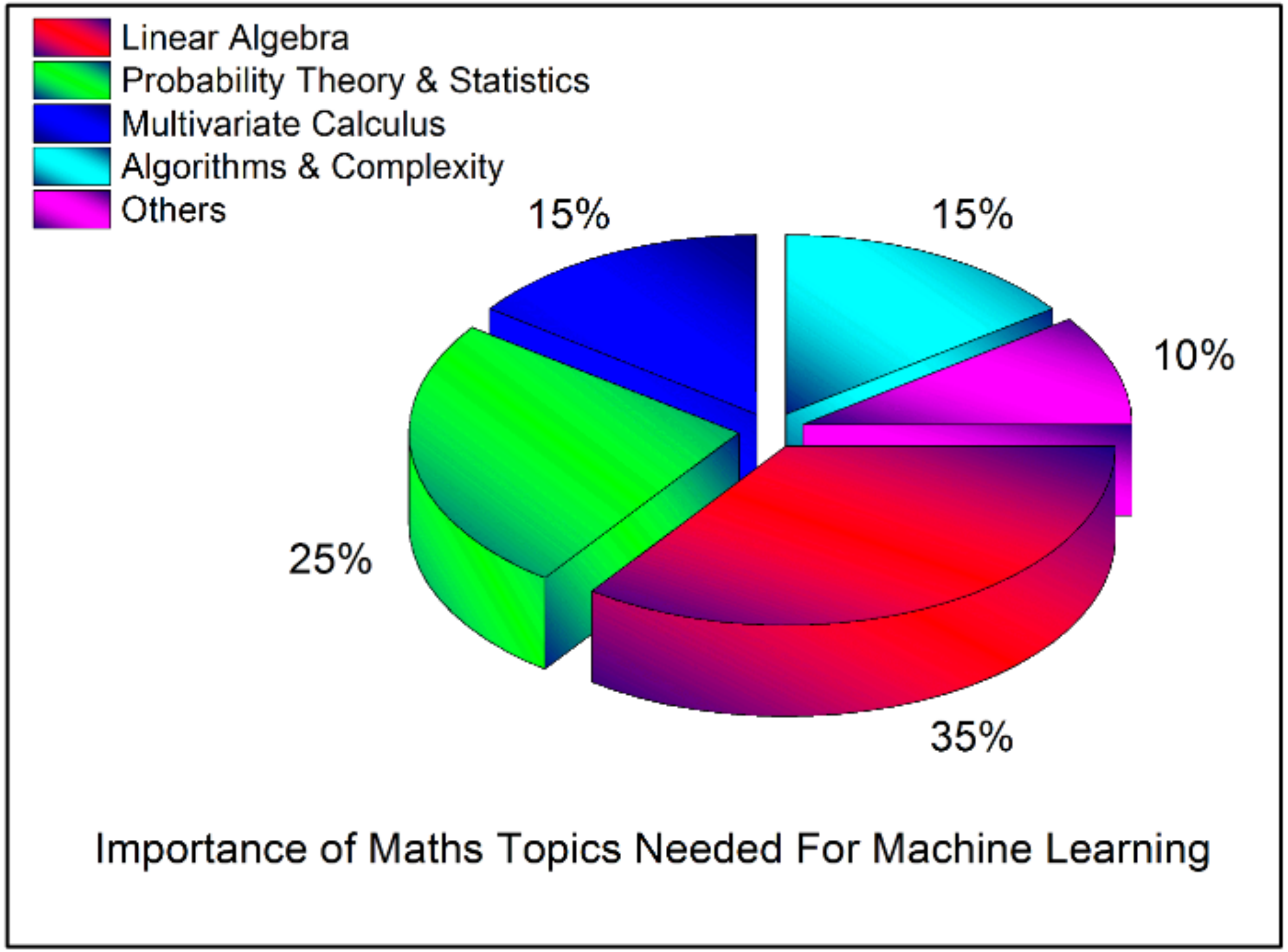
Faculty of Information Systems and Technologies

Ulyanovsk, Russia, 2017

Plan

1. Math models and data
2. Linear algebra
3. Optimization theory. Regression and gradient descend
4. Statistics

What Level of Maths Do You Need?



Math is… models and modeling

Model example - Hooke's law

$$F = -kx$$

$$m\ddot{x} = -kx$$

$$\ddot{x} = \frac{d^2x}{dt^2}$$

Math models

Model in mathematic is just a symbolic formula:

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

X – input data (this information about the object \ task that we know)

Y – output (result) data (this information about the object \ task that we want get)

W – parameters of model (this information about math regularity \ dependence between X and Y which we assume)

Parameters of model can be

In more detail form…

$$Y = P(X, W) = G(X, Wc, Ws, Wd) = F(X, W1, W2 \ldots WN)$$

- Wc – constants. Defined once by human based on experience, knowledge or other model and remain unchanged in the simulation.
- Ws – statics. Defined once by solving an equation and remain unchanged in the simulation.
- Wd – dynamic. Defined by solving an equation in the simulation

Types of model

- Linear vs. non-linear
- Static vs. dynamic
- Explicit vs. implicit
- Discrete vs. continuous
- Deterministic vs. probabilistic (stochastic)
- Deductive vs. inductive

Presentation of information

Information about object is a set of features which describes this concrete object

Example: so apartment can be present as set of next property:

- Address
- Area
- Year of build
- Floor
- Room count
- etc…

All as in programming!

Presentation of information

But all categorical (enum) types of feature must be translate to numeric

(coded as numeric)!

And now we get a vector X:

- Address = (street code)
- Area = (10..100)
- Year of build = (1900..2017)
- Floor = (1..120)
- Room count = (1..30)
- etc…

X = [(street code),
(10..100),
(1900..2017),
(1..120),
(1..30),
…]

Linear algebra

Linear algebra applications

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

Operations on vectors and matrices is also a base of robot navigation

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

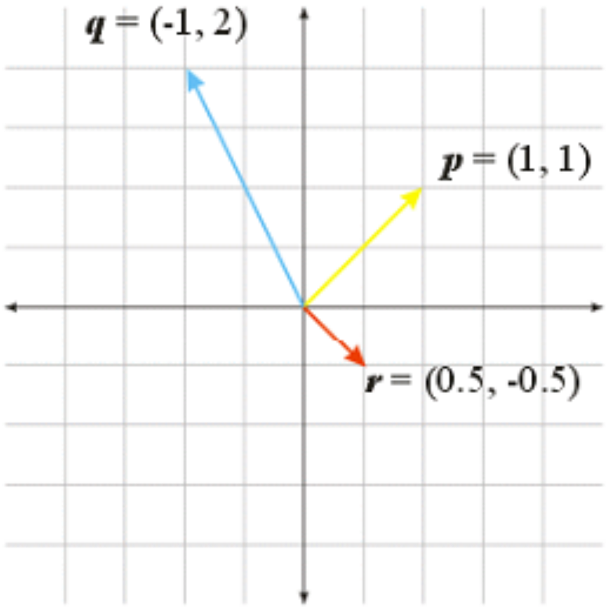
- Definition: an *n*-tuple of values (usually real numbers).
  - *n* referred to as the *dimension* of the vector
  - *n* can be any positive integer, from 1 to infinity
- Can be written in column form or row form
  - Column form is conventional
  - Vector elements referenced by subscript

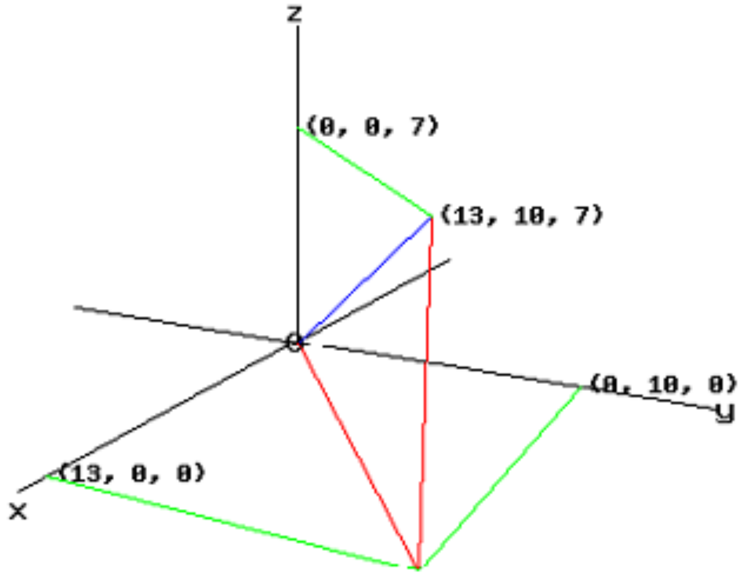
$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \mathbf{x}^{\mathrm{T}} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$$

$^{\mathrm{T}}$ means "transpose"

Vectors

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

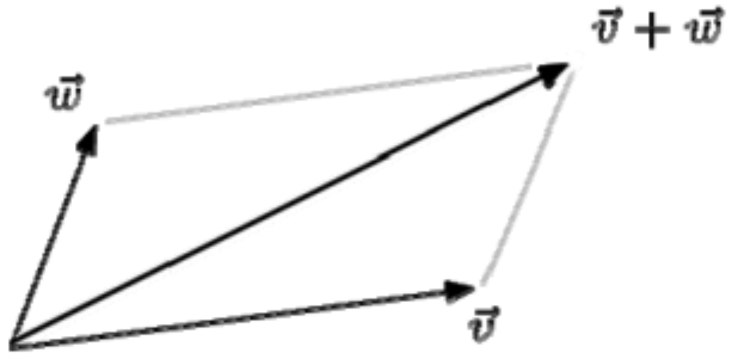




Vector arithmetic

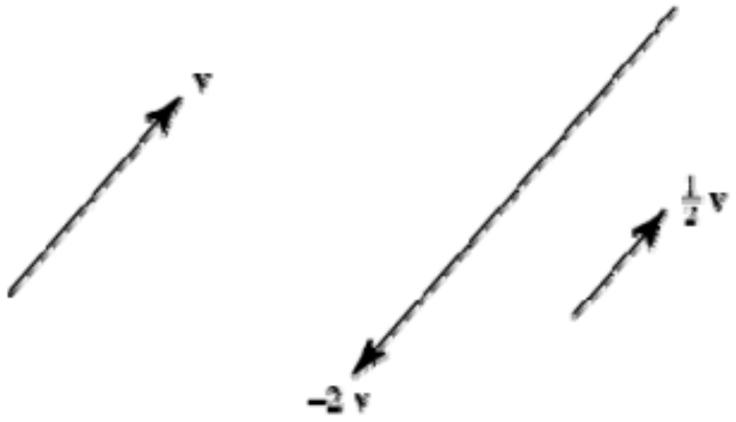
- Addition of two vectors
  - add corresponding elements

    $$\mathbf{z} = \mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 + y_1 & \cdots & x_n + y_n \end{pmatrix}^{\mathrm{T}}$$
  - result is a vector



- Scalar multiplication of a vector
  - multiply each element by scalar

    $$\mathbf{y} = a\mathbf{x} = \begin{pmatrix} a x_1 & \cdots & a x_n \end{pmatrix}^{\mathrm{T}}$$
  - result is a vector



Vector arithmetic

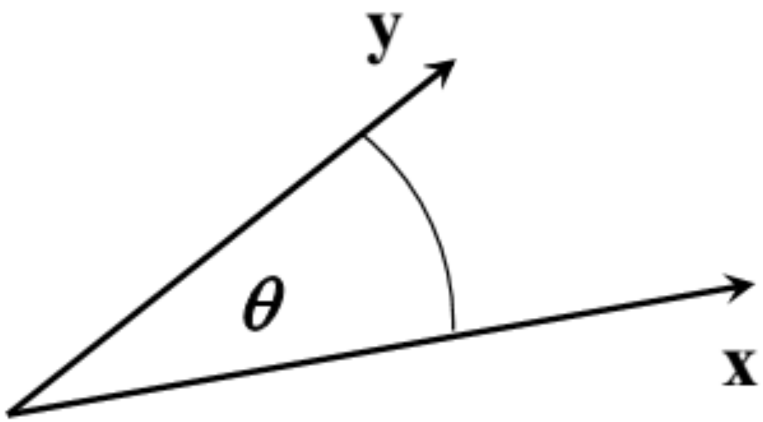
- Dot product of two vectors
  - multiply corresponding elements, then add products

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

  - result is a <u>scalar</u>

- Dot product alternative form

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



Matrices

- Definition: an *m* x *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 **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 x n matrix becomes n x m matrix
  - example:

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

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

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

- 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:

$\mathbf{A} \cdot \mathbf{B} \cdot \mathbf{A}^T \quad \mathbf{C}^T \cdot \mathbf{A} \cdot \mathbf{B} \quad \mathbf{A}^T \cdot \mathbf{A} \cdot \mathbf{B} \quad \mathbf{C} \cdot \mathbf{C}^T \cdot \mathbf{A}$

Wrong:

$\mathbf{A} \cdot \mathbf{B} \cdot \mathbf{A} \quad \mathbf{C} \cdot \mathbf{A} \cdot \mathbf{B} \quad \mathbf{A} \cdot \mathbf{A}^T \cdot \mathbf{B} \quad \mathbf{C}^T \cdot \mathbf{C} \cdot \mathbf{A}$

Vector projection

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

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

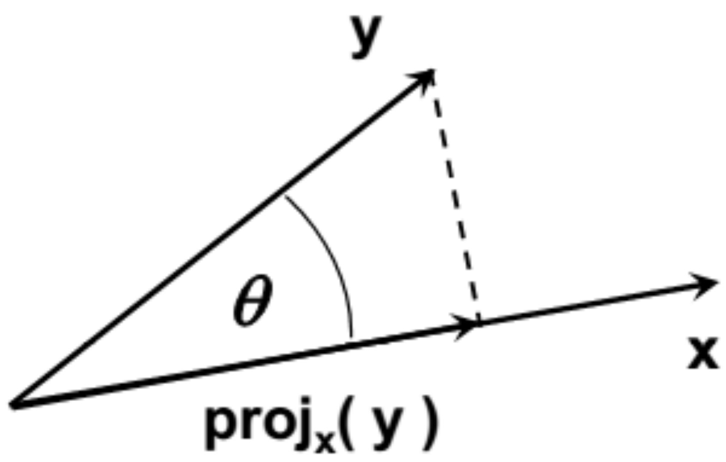
  - Length of projection of **y** in direction of **x** is

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

  - Orthogonal projection of **y** onto **x** is the vector

    $$\mathbf{proj_x}(\, \mathbf{y} \,) \;=\; \mathbf{x} \cdot \|\, \mathbf{y} \,\| \cdot \cos(\theta) \,/\, \|\, \mathbf{x} \,\| \;=$$

    $[\, (\, \mathbf{x} \cdot \mathbf{y} \,) \,/\, \|\, \mathbf{x} \,\|^2 \,]\, \mathbf{x}$ (using dot product alternate form)



Optimization theory

Idea of error

- The formula for obtaining the exact value is not always known, but the optimal solution of the problem nevertheless exists
- Estimating the error or deviation from the desired value. A very important idea in the theory of optimization!

Example: When we go on a bicycle, the brain does not know the formula for the motion of such a complex system in the gravitational field. Nevertheless he successfully solves the problem by using only the idea of the desired and deviating from the desired (error)

Idea of error and iterative procedure

Not all tasks can be solved analytically.

Sometimes it is the output parameters which are known, and the corresponding inputs must be solved for by an iterative procedure, such as Newton's method (if the model is linear) or Broyden's method (if non-linear)

Example: n-body problem in physics not solved yet

Idea of error and iterative procedure

Representing the error as a function, we can set the task of minimizing the error in an iterative way:

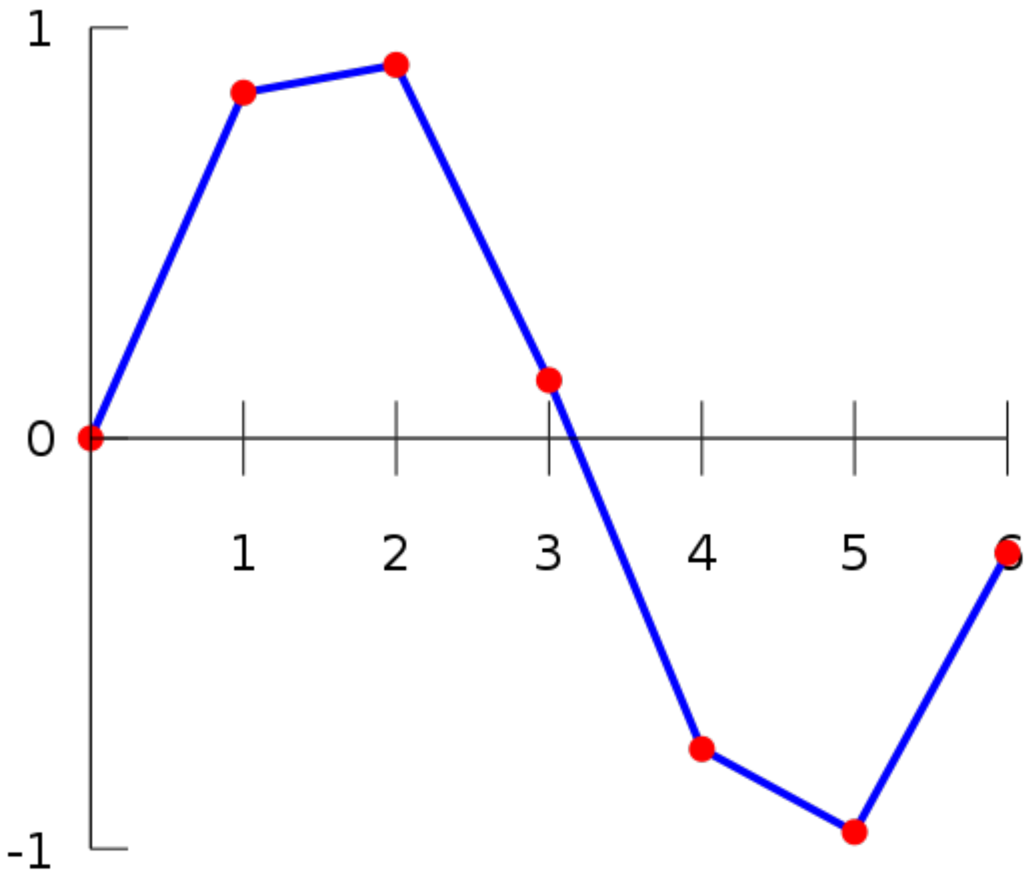
- Simplex algorithm
- Least squares
- Newton's method
- Gradient descent
- etc…

$$f(x) \to min$$

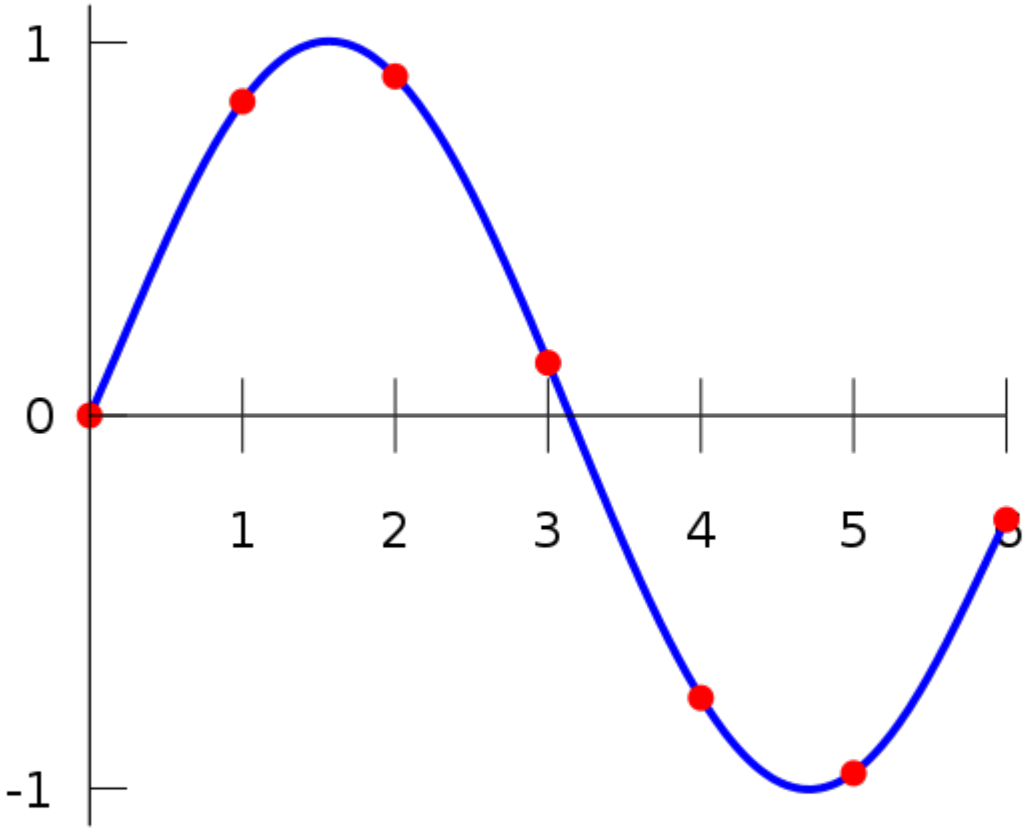
$$E = |y_r - Y_m|$$

$$Y_m = wx + b$$

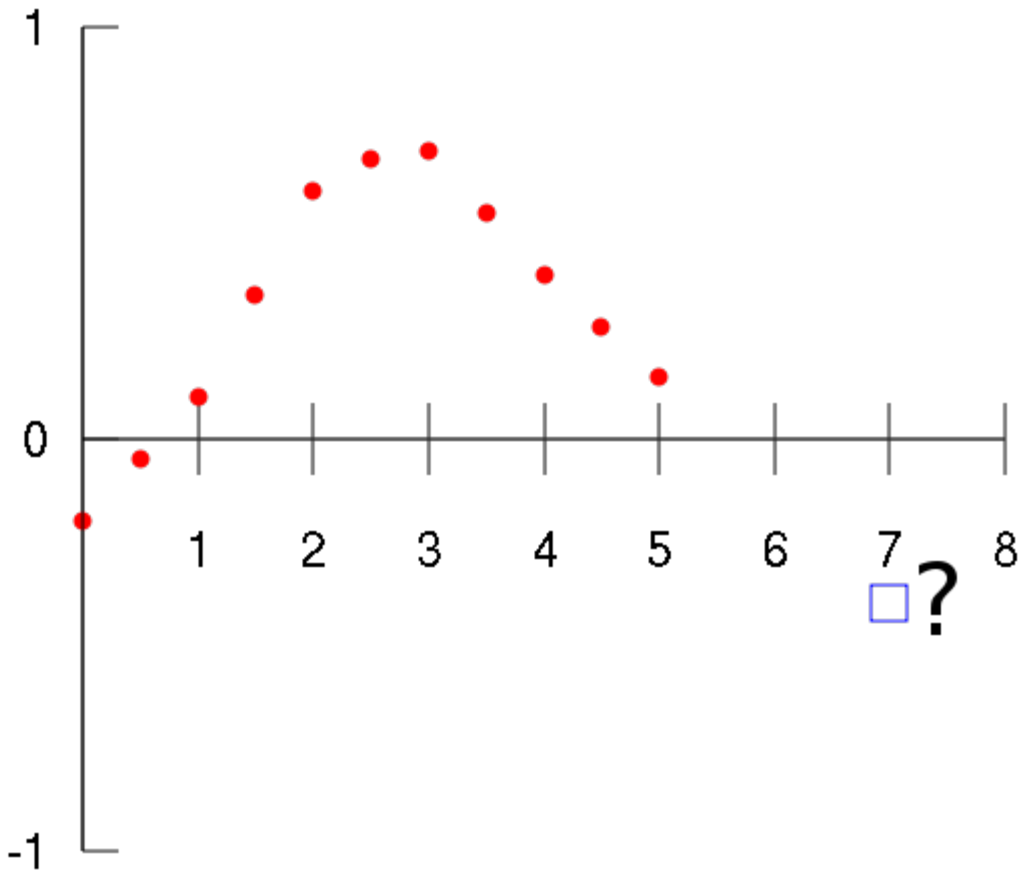
Linear interpolation



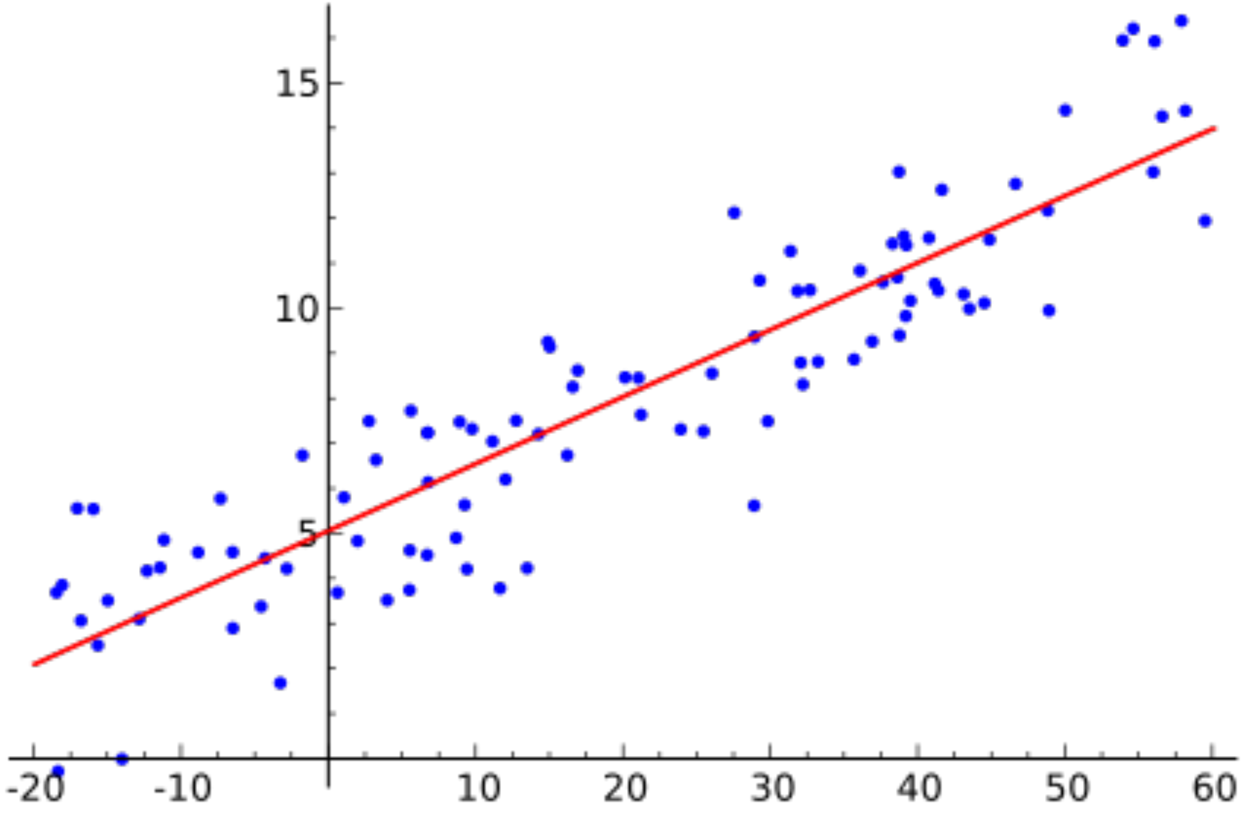
Polynomic interpolation



Extrapolation



Linear regression

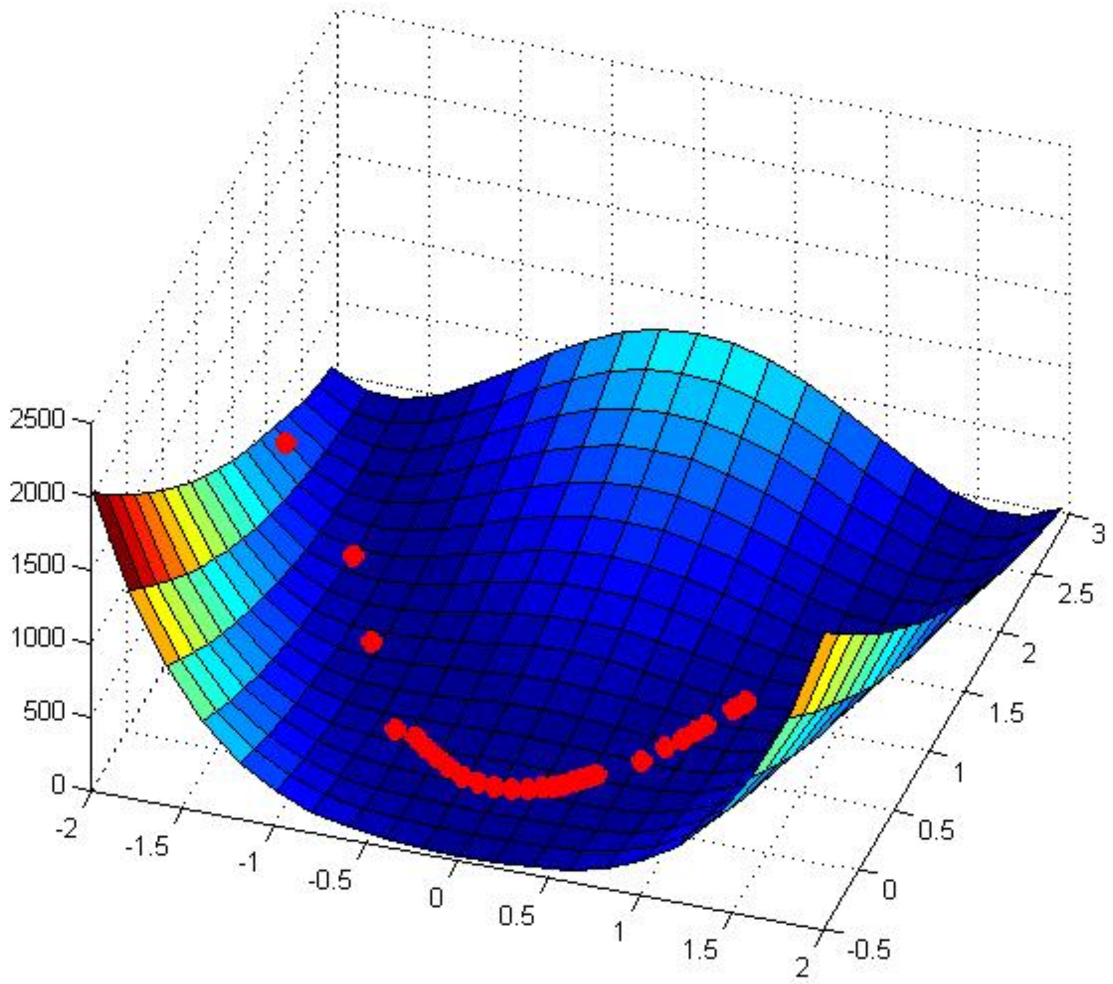


Gradient descent

is a first-order iterative optimization algorithm for finding the minimum of a function.

To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or of the approximate gradient) of the function at the current point. If instead one takes steps proportional to the positive of the gradient, one approaches a local maximum of that function; the procedure is then known as gradient ascent.

Gradient descent



Gradient descent

$$w_{i+1} = w_i - \lambda_i \nabla F(w_i)$$

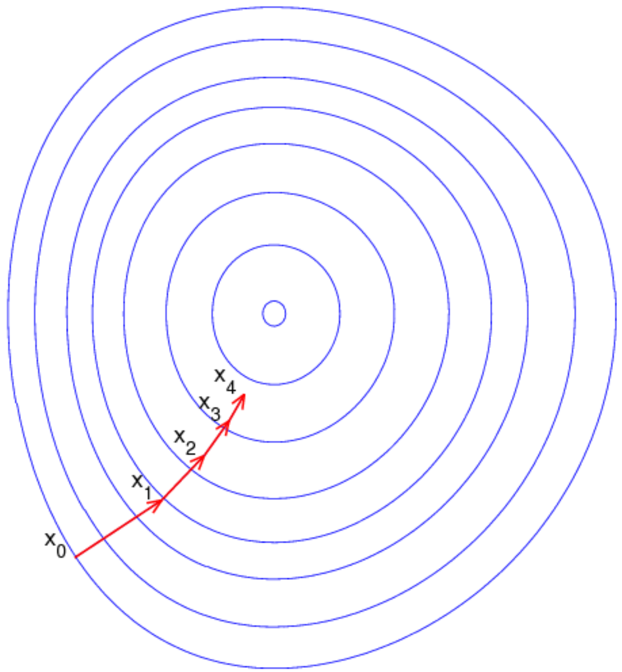
$$E = (y_r - Y_m)^2 = y_r^2 - 2y_r Y_m + Y_m^2$$

$$Y_m = wx + b$$

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial Y_m}\frac{\partial Y_m}{\partial w} = (-2y_r + 2Y_m)\frac{\partial Y_m}{\partial w} = -2y_r x + 2Y_m x$$

$$= -2y_r x + 2(w_i x + b)x \ = -y_r x + w_i x^2 + bx$$

Gradient descent



- For some of the above examples, gradient descent is relatively slow close to the minimum: technically, its asymptotic rate of convergence is inferior to many other methods. For poorly conditioned convex problems, gradient descent increasingly 'zigzags' as the gradients point nearly orthogonally to the shortest direction to a minimum point. For more details, see the comments below.

Gradient descent

- For non-differentiable functions, gradient methods are ill-defined.

There are many modifications of the basic algorithm

- Conjugate gradient method
- Stochastic gradient descent
- Rprop
- Delta rule
- Wolfe conditions
- Preconditioning
- BFGS method
- etc…

Statistics

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 = 0.0057.
- 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 $E \in F$, $p(E) \geq 0$

2. All possible outcomes:

   $p(\Omega) = 1$

3. Additivity of disjoint events:

   for all events $E, E' \in F$ where $E \cap E' = \varnothing$,
   $p(E \cup E') = p(E) + p(E')$

Types of probability spaces

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

- Discrete space $|\Omega|$ is finite
  - Analysis involves *summations* ($\sum$)
- 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, 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 = (outcome is odd) is

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

Example of discrete probability space

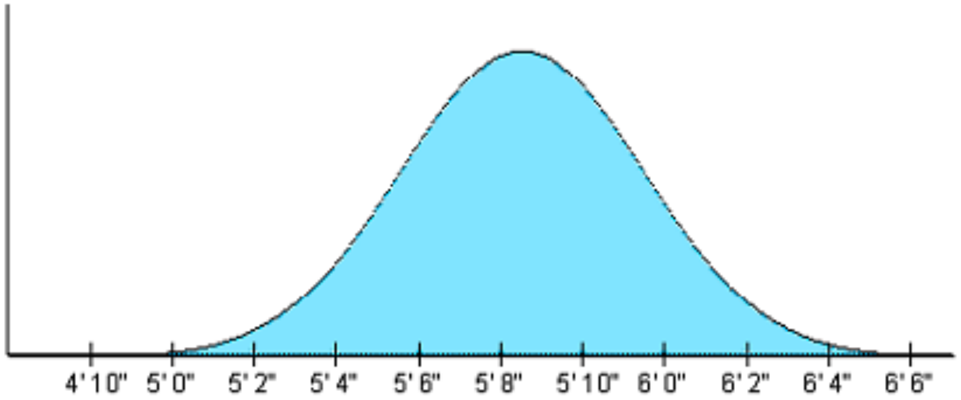
Three consecutive flips of a coin

- 8 possible outcomes: O = HHH, HHT, HTH, HTT, THH, THT, TTH, TTT
- $2^8 = 256$ 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 = (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 American male

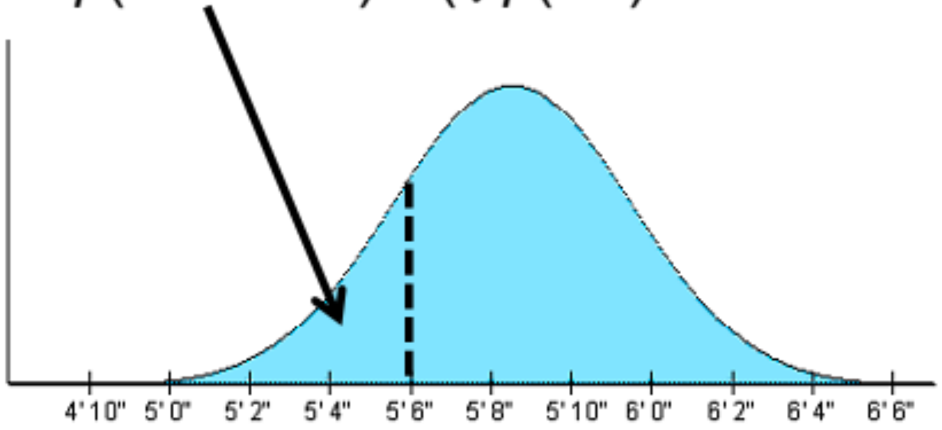
- Infinite number of possible outcomes: *O* has some single value in range 2 feet to 8 feet
- Infinite number of possible events
  - example: *E* = (*O* | *O* < 5.5 feet), i.e. individual chosen is less than 5.5 feet tall
- Probabilities of outcomes are not equal, and are described by a continuous function,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



Example of continuous probability space

Height of a randomly chosen American male

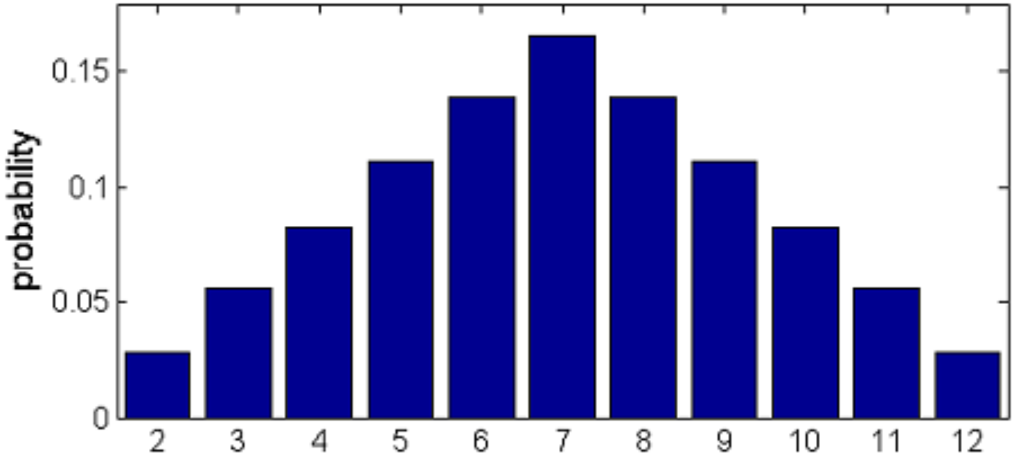
- Probabilities of outcomes *O* are not equal, and are described by a continuous function, *p*(*O*)
- *p*(*O*) is a *relative*, not an *absolute* probability
  - *p*(*O*) for any particular *O* is zero
  - ∫ *p*(*O*) from *O* = -∞ to ∞ (i.e. area under curve) is 1
  - example: *p*(*O* = 5’8”) > *p*(*O* = 6’2”)
  - example: *p*(*O* < 5’6”) = (∫ *p*(*O*) from *O* = -∞ to 5’6”) ≈ 0.25



Probability distributions

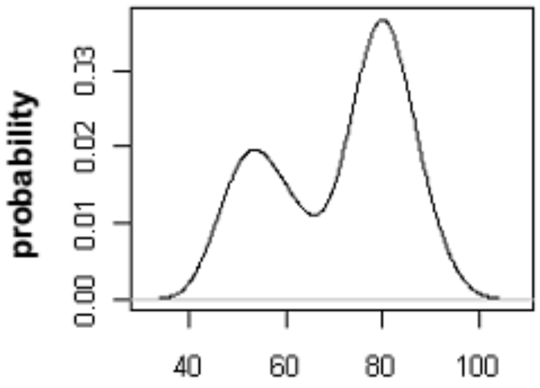
- Discrete: *probability mass function* (pmf)

  example:
  sum of two
  fair dice



- Continuous: *probability density function* (pdf)

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



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
  - $p(X = 0) = p(X = 3) = 1 / 8$ $\quad$ $p(X = 1) = p(X = 2) = 3 / 8$
  - Size of space (number of "outcomes") reduced from 8 to 4
- Example: X = average height of five randomly chosen American men
  - Size of space unchanged (X can range from 2 feet to 8 feet), but pdf of X different than for single man

Expected value

Given:

- A discrete random variable X, with possible values $x = x_1, x_2, \ldots x_n$
- 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 $f(x_i)$:

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

Example of expected value

- Process: game where one card is drawn from the deck
  - If face card, dealer pays you \$10
  - If not a face card, you pay dealer \$4
- Random variable X = { face card, not face card }
  - p(face card) = 3/13
  - p(not face card) = 10/13
- Function $f(X)$ is payout to you
  - f(face card) = 10
  - f(not face card) = -4
- *Expected value* of payout is:

  $$E(f) = \sum_i p(x_i) \cdot f(x_i) = 3/13 \cdot 10 + 10/13 \cdot -4 = -0.77$$

Common forms of expected value (1)

- Mean (μ)

  $f(x_i) = x_i \Rightarrow \quad \mu = E(f) = \sum_i p(x_i) \cdot x_i$

  - Average value of $X = x_i$, taking into account probability of the various x_i
  - Most common measure of "center" of a distribution

- Compare to formula for mean of an actual sample

  $$\mu = \frac{1}{N}\sum_{i=1}^{n} x_i$$

Common forms of expected value (2)

- Variance (σ^2)

  $f(x_i) = (x_i - \mu) \quad \Rightarrow \qquad \sigma^2 = \sum_i p(x_i) \cdot (x_i - \mu)^2$

  – 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*

- Compare to formula for variance of an actual sample

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

Common forms of expected value (3)

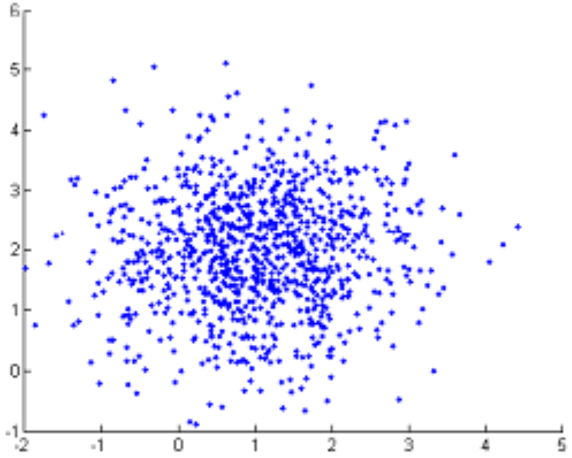
- Covariance

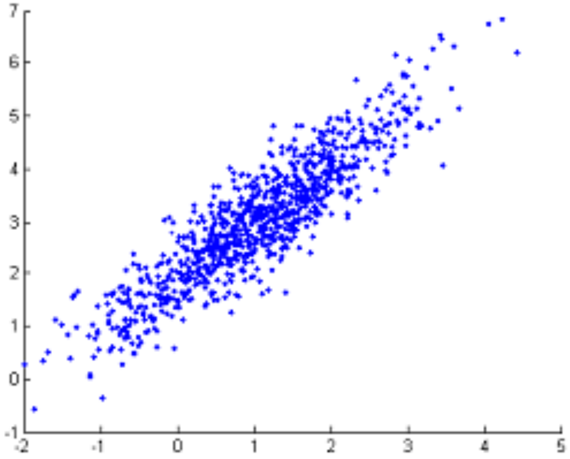
  $f(x_i) = (x_i - \mu_x), \quad g(y_i) = (y_i - \mu_y) \quad \Rightarrow$

  $\mathrm{cov}(x, y) = \sum_i p(x_i, y_i) \cdot (x_i - \mu_x) \cdot (y_i - \mu_y)$

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

no covariance





high (positive) covariance

- Compare to formula for covariance of actual samples

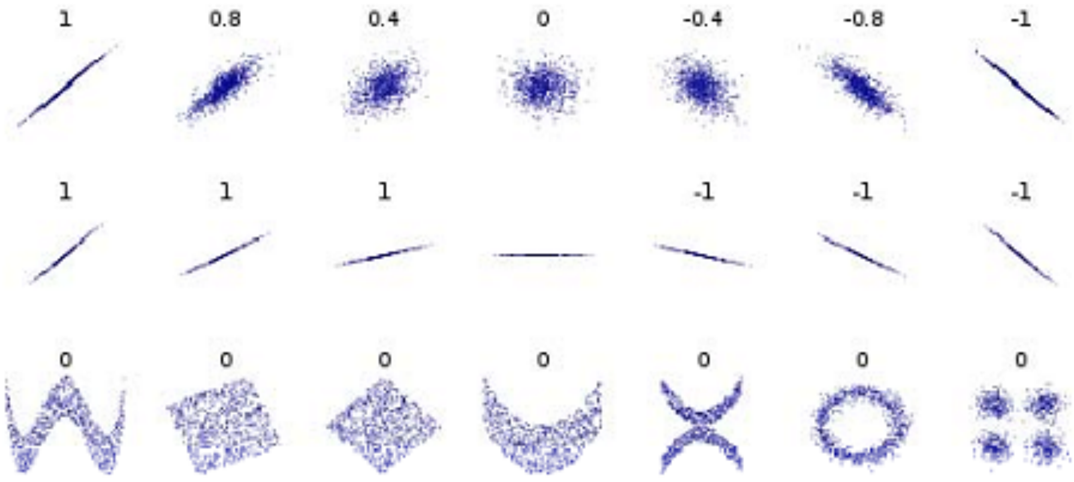
$$\mathrm{cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)$$

Correlation

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

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

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



Linear dependence with noise

Linear dependence without noise

Various nonlinear dependencies

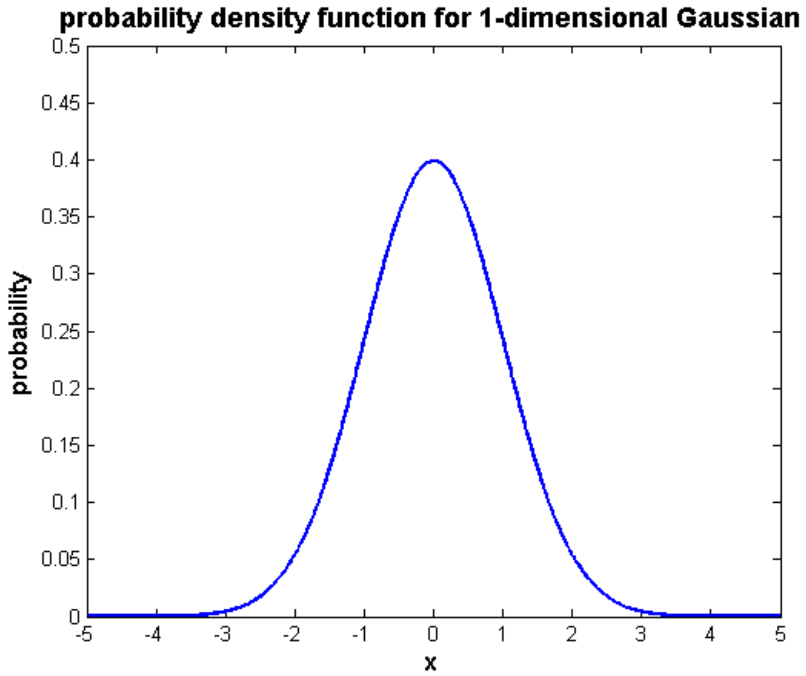
Gaussian distribution

- Most commonly used continuous probability distribution
- Also known as the normal distribution
- Two parameters define a Gaussian:
  - Mean μ location of center
  - Variance σ^2 width of curve

Gaussian distribution

In one dimension

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Gaussian distribution

In one dimension

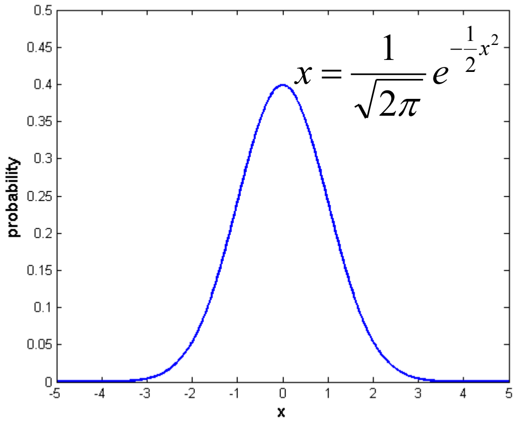
Causes pdf to decrease as distance from center increases

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

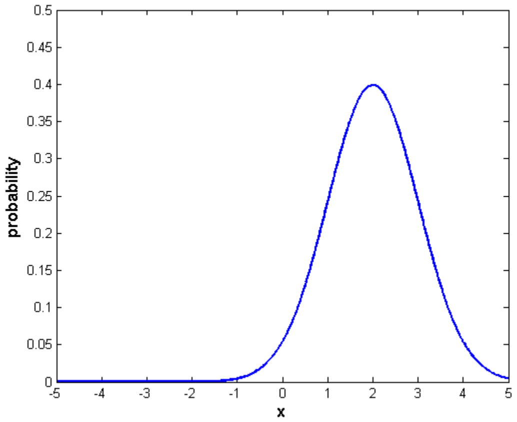
Controls width of curve

Normalizing constant: insures that distribution integrates to 1

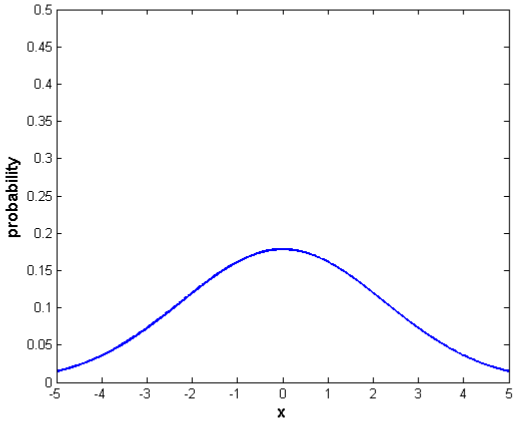
Gaussian distribution



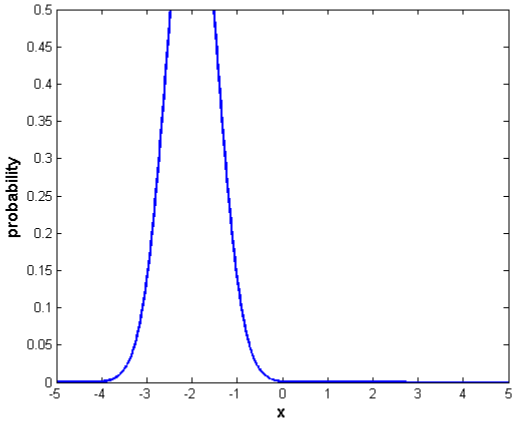
$\mu = 0 \quad \sigma^2 = 1$



$\mu = 2 \quad \sigma^2 = 1$



$\mu = 0 \quad \sigma^2 = 5$



$\mu = -2 \quad \sigma^2 = 0.3$

Multivariate Gaussian distribution

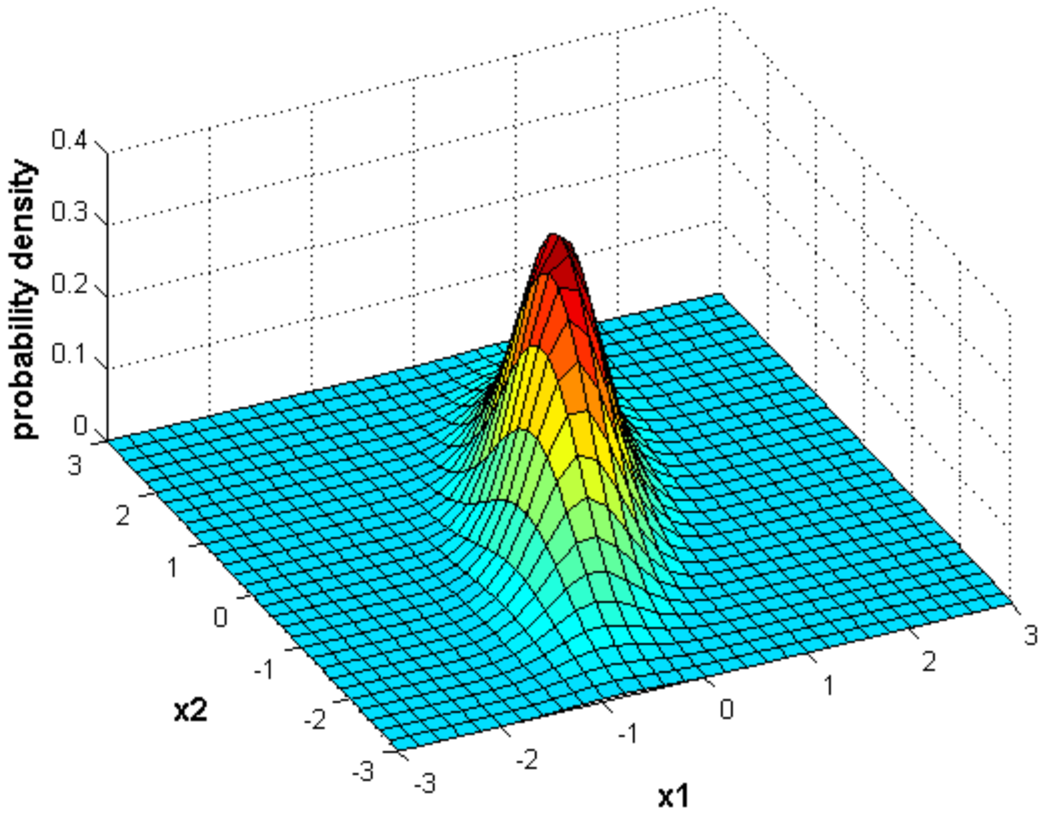
In *d* dimensions

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- x and μ now *d*-dimensional vectors
  - μ gives center of distribution in *d*-dimensional space
- σ^2 replaced by Σ, the *d* x *d* covariance matrix
  - Σ contains pairwise covariances of every pair of features
  - Diagonal elements of Σ are variances σ^2 of individual features
  - Σ describes distribution's shape and spread

Multivariate Gaussian distribution

In two dimensions

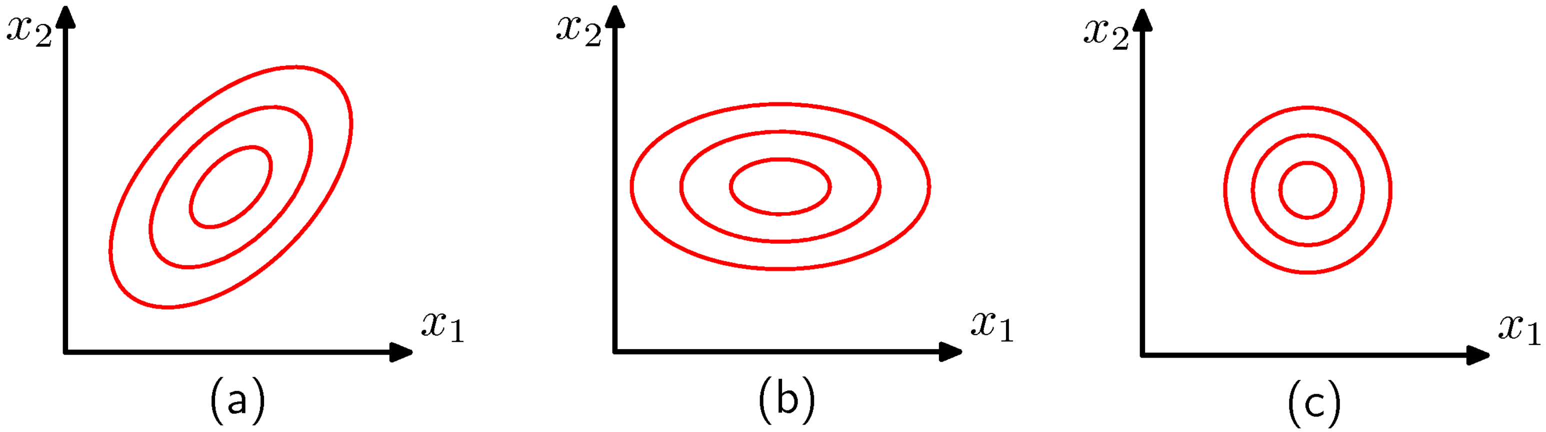


$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$

Multivariate Gaussian distribution

In two dimensions

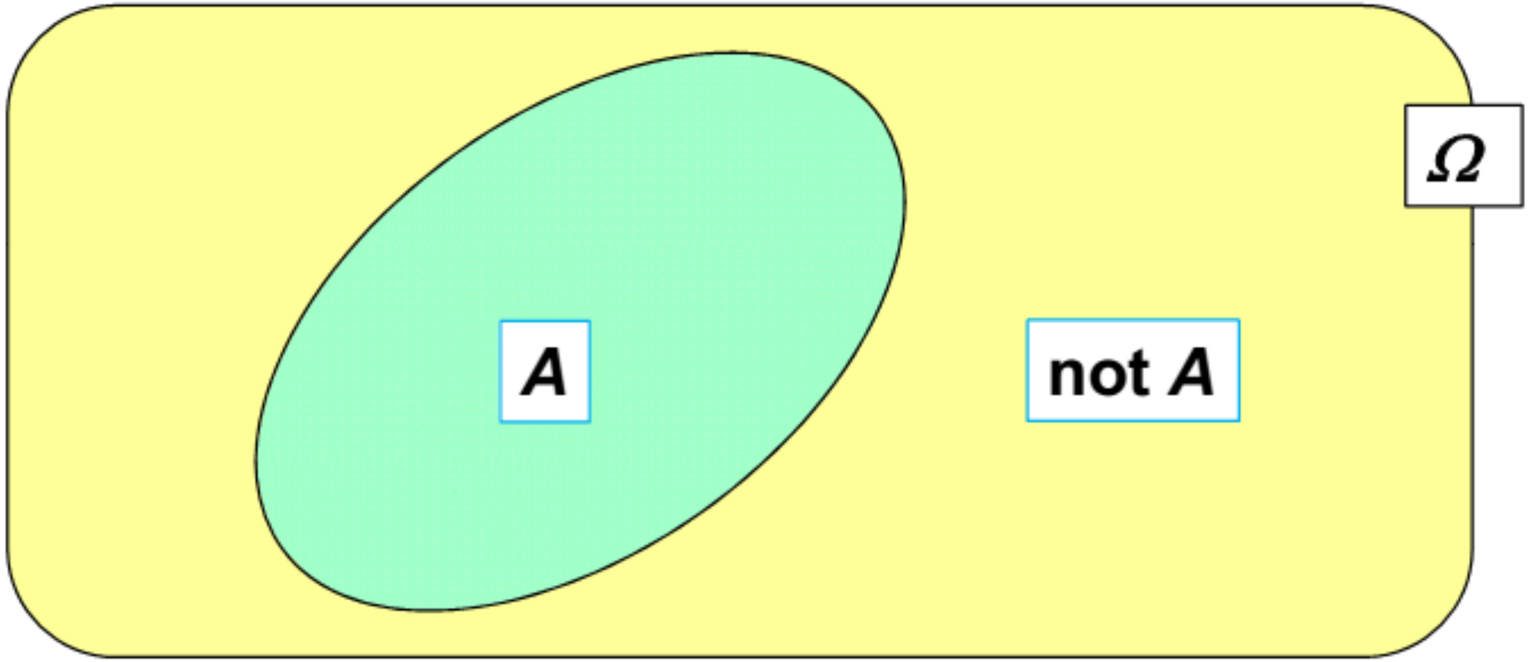


$$\mathbf{\Sigma} = \begin{bmatrix} 2 & 0.6 \\ 0.6 & 2 \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Complement rule

Given: event A, which can occur or not

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



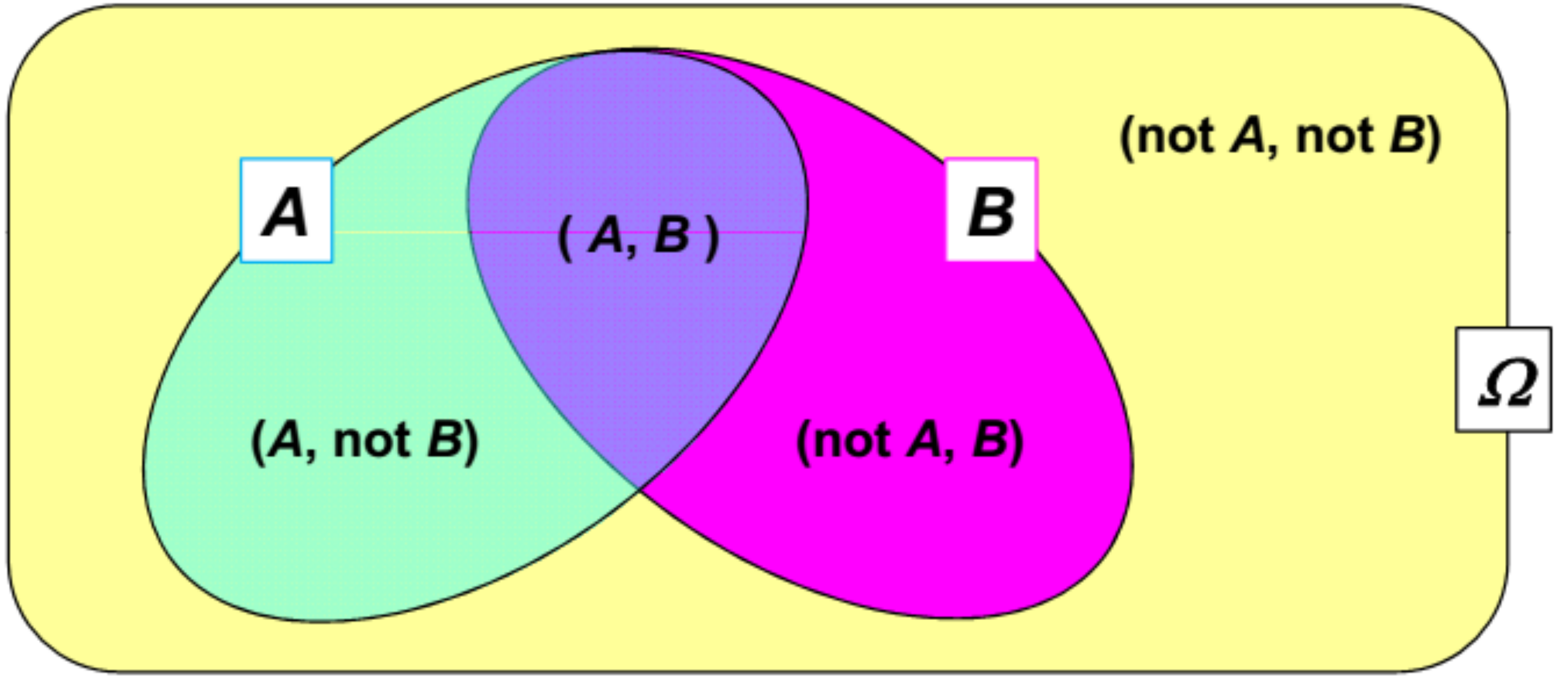
areas represent relative probabilities

Product rule

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

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

(same expression given previously to define conditional probability)



areas represent relative probabilities

Example of product rule

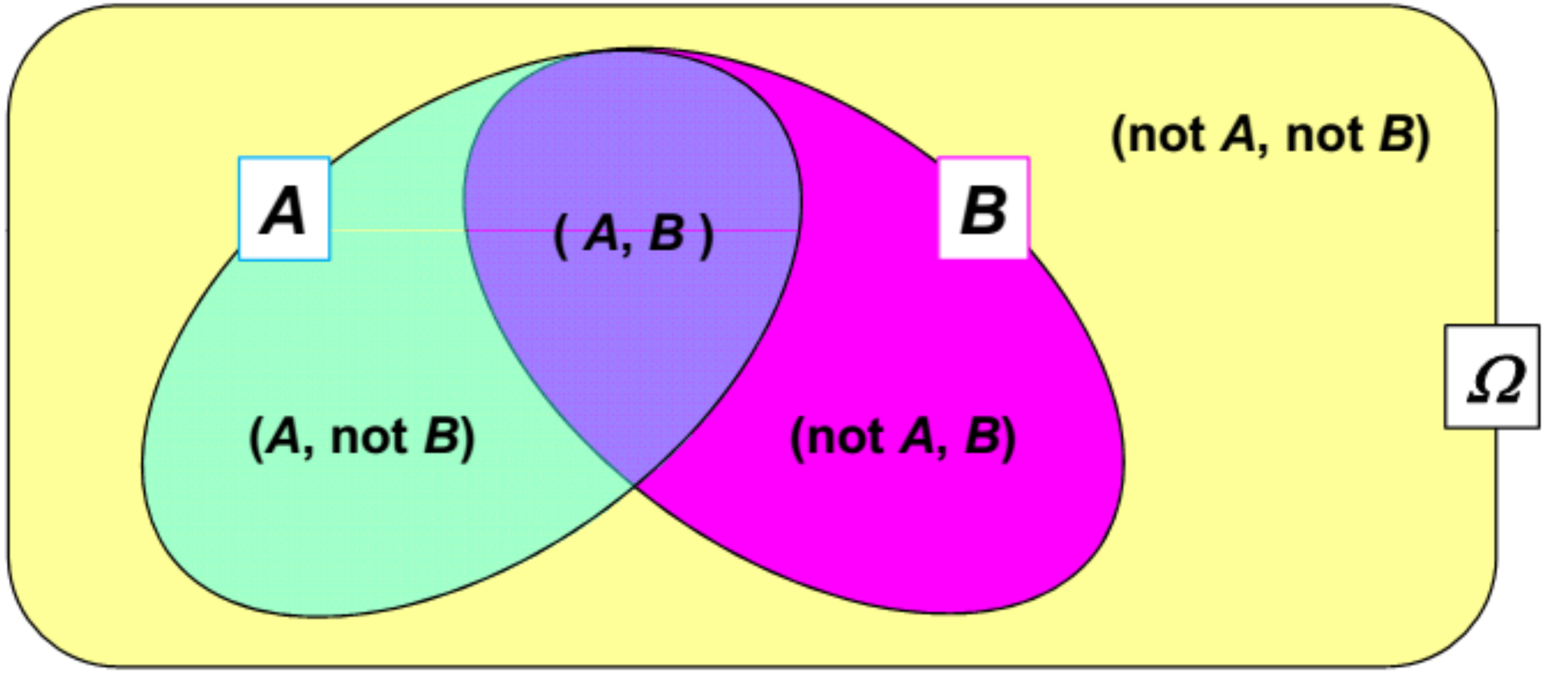
- Probability that a man has white hair (event A) and is over 65 (event B)
  - $p(B) = 0.18$
  - $p(A \mid B) = 0.78$
  - $p(A, B) = p(A \mid B) \cdot p(B) =$
    $0.78 \cdot 0.18 =$
    0.14

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)

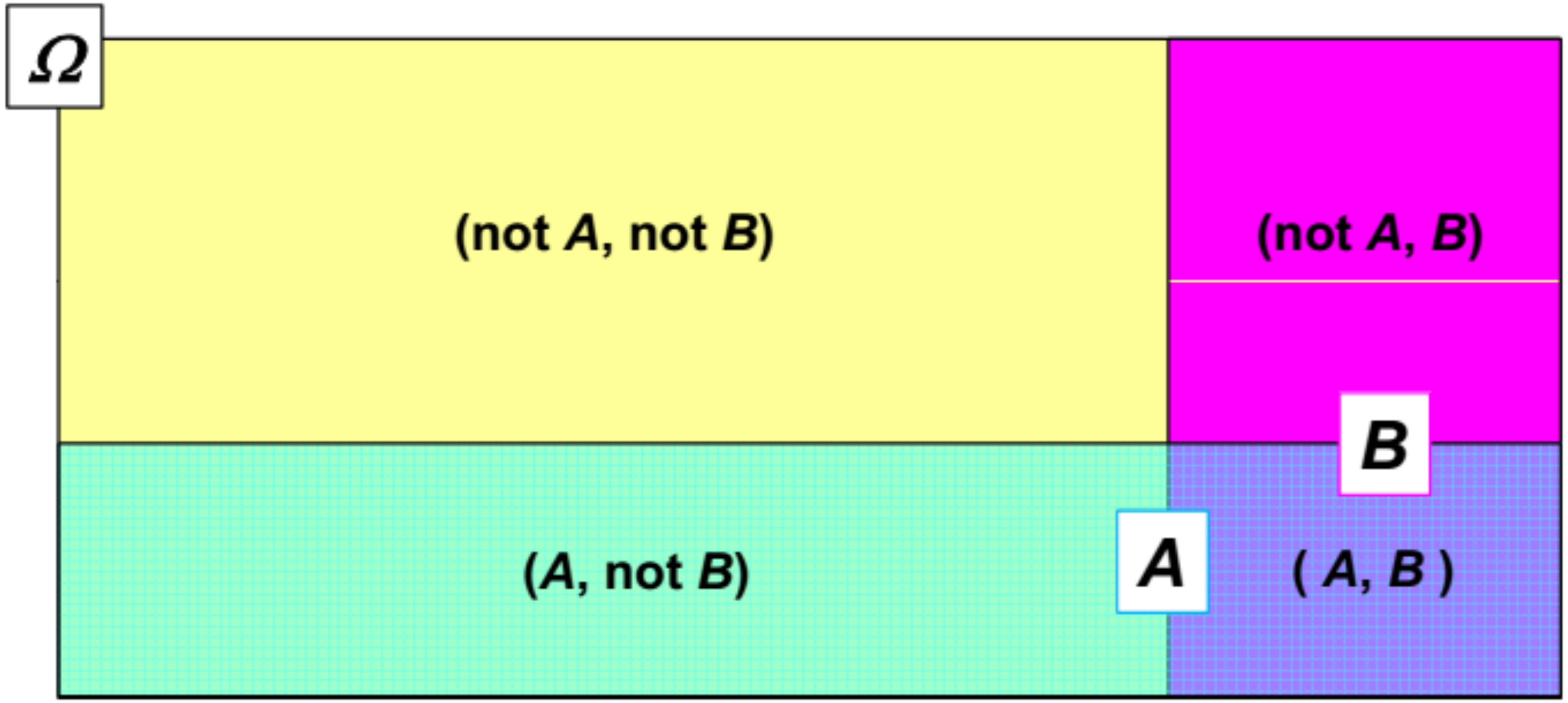


areas represent relative probabilities

Independence

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

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



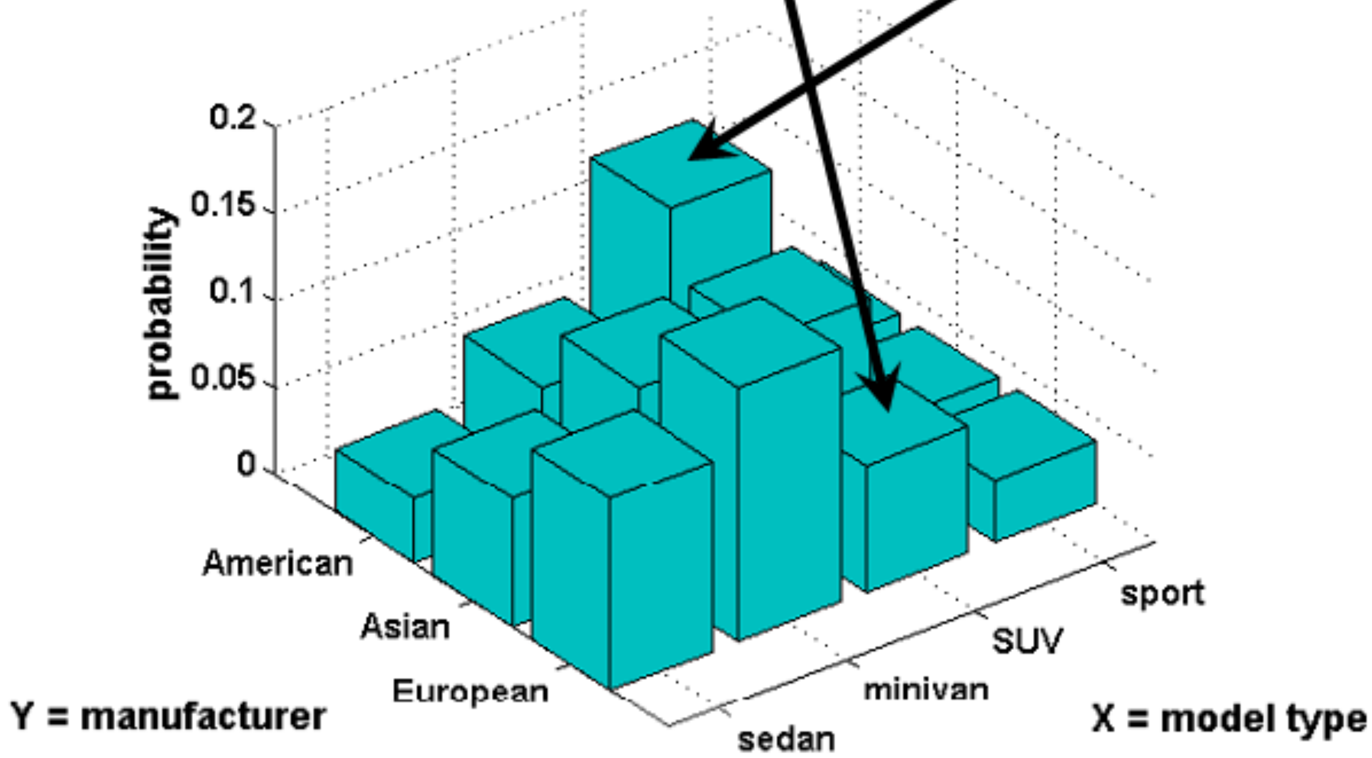
areas represent relative probabilities

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*

Examples 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.

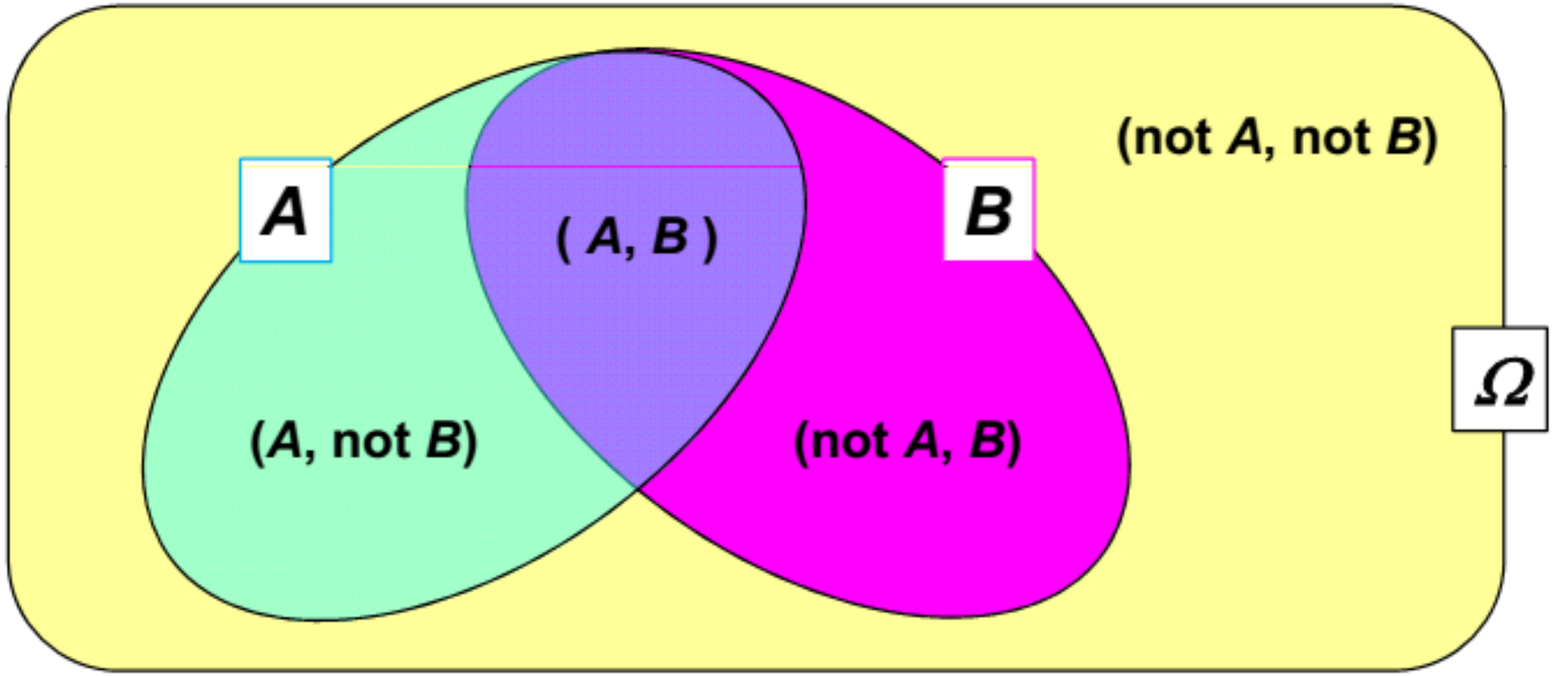


Bayes rule

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

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

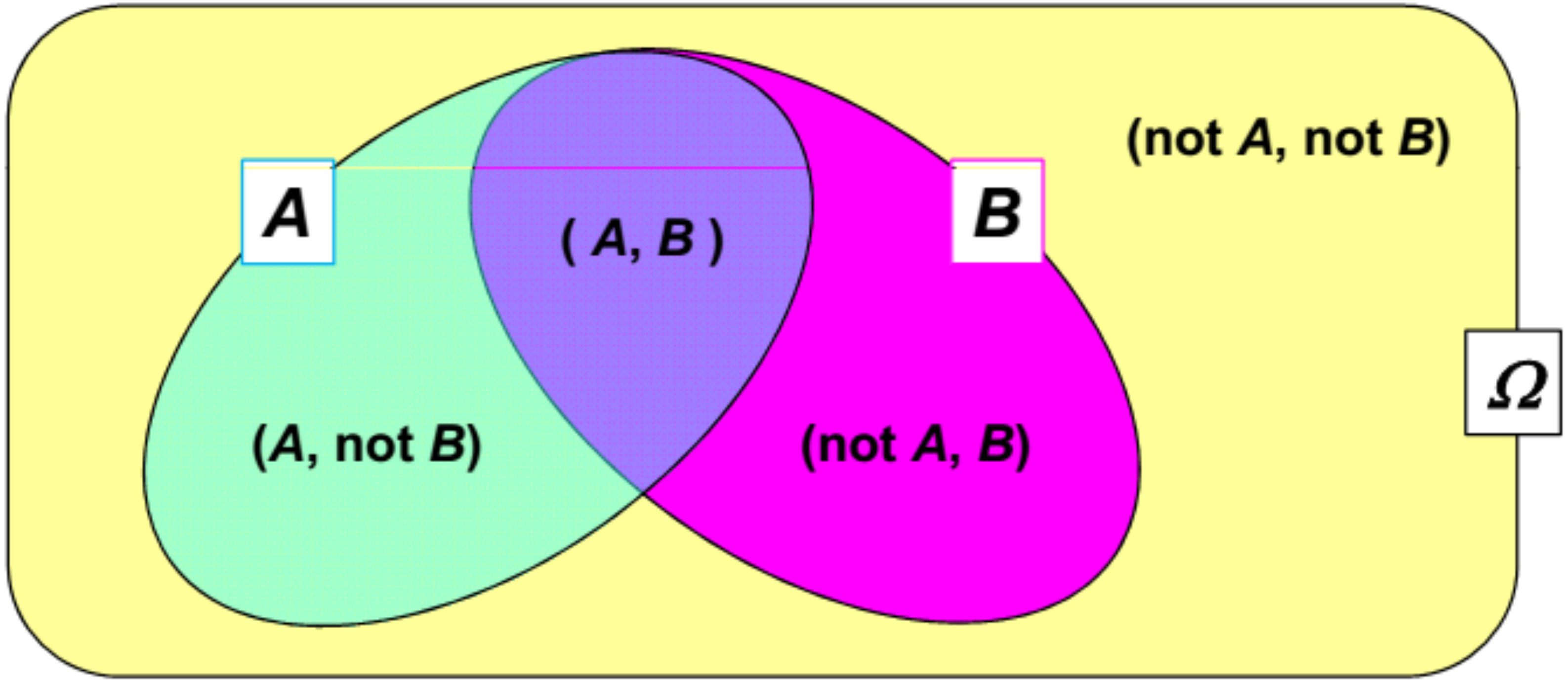
where $p(A) = p(A, B) + p(A, \text{not } B)$



Bayes rule

posterior probability $\propto$ likelihood $\times$ prior probability

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



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.]
  - $p(\text{not } B) = 360 / 365 = 0.9863$
  - $p(A \mid B) = 0.9$ [When it rains, the weatherman has forecast rain 90% of the time.]
  - $p(A \mid \text{not } B) = 0.1$ [When it does not rain, the weatherman has forecast rain 10% of the time.]

Example of Bayes rule (2)

- We want to know $p(B \mid A)$, 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. $p(B \mid A) = p(A \mid B) \cdot p(B) / p(A)$
2. $p(A) = p(A \mid B) \cdot p(B) + p(A \mid \text{not } B) \cdot p(\text{not } B) =$
   $(0.9)(0.014) + (0.1)(0.986) = 0.111$
3. $p(B \mid A) = (0.9)(0.0137) / 0.111 = 0.111$

- 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.

Advantages and disadvantages of Bayes

Advantages

- Very good where the features and result have explicit frequency dependencies
- Easy to use expert knowledge (apriority information)
- Allows to describe the phenomenon most accurately
- The result easily interpreted

Disadvantages

- The Bayes method gives a poor generalization, especially on high-level signs.
- Requires expert work (preparing data is more difficult)
- The distribution may be different from normal and for this you need to conduct a separate study.
- Naive Bayes assumes that the signs are independent of each other, which may not be the case.
- Poor performance with a small amount of data
- Poor performance with high dimensional data sets

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*.