Math Essentials

Alireza Samar

UTM UNIVERSITI TEKNOLOGI MALAYSIA

MLDS MACHINE LEARNING FOR DATA SCIENCE

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 transformations on data
  - Dimensionality reduction techniques
- **Optimization** theory (Maximum likelihood, Expectation maximization, Gradient descent)

MLDS

Why worry about the math?

- There are lots of easy-to-use machine learning packages out there.
  packages out there.
- After this series of workshops, 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 which we will cover in next modules.

MLDS

Why worry about the math?

These intuitions will allow you to:

- Choose the right algorithm(s) for the problem 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

MLDS

Notation

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

MLDS

Notation

- **x, y, z, u, v** — *vector* (bold, lower case)
- **A, B, X** — *matrix* (bold, upper case)
- $y = f(x)$ — function (map): assigns unique value in range of *y* to each value in domain of *x*
- dy / dx — derivative of *y* with resp g ect to single variable *x*
- $y = f(\boldsymbol{x})$ — *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 $\boldsymbol{x}$

MLDS

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 If a particular outcome happens more often than another outcome, we say it is more probable.

MLDS

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.

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

MLDS

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 | Ω | = number of possible outcomes

- Discrete space | Ω | is finite
  - Analysis involves summations ($\sum$)
- Continuous space | Ω | is infinite
  - Analysis involves integrals ($\int$)

MLDS

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 = (outcome is odd) is

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

MLDS

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 = ($O \in$ { HHT, HTH, THH }), i.e. exactly two flips are heads
  - Example: E = (O $\in$ { THT, TTT }), i.e. the first and third flips are tails
- If coin is fair, then probabilities of outcomes are equal

  p(HHH) = p(HHT) = p(HTH) = p(HTT) = p(THH) = p(THT) = p(TTH) = p(TTT) = 1 / 8
  - Example: probability of event E = (exactly two heads) is

    p(HHT) + p(HTH) + p(THH) = 3 / 8

MLDS

Example of continuous probability space

Height of a randomly chosen American male

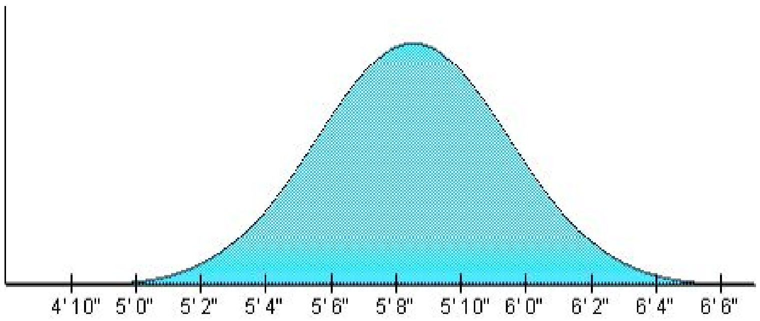
- Infinite number of possible outcomes: *O* has some has some single value in range 2 feet to 8 feet
- Infinite number of possible events
  - Example: E = (*O* | *O* < *5.5 feet*), i.e. individual chosen is less than 5.5 feet tall
- Probabilities of outcomes are not equal, and are described by a continuous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

MLDS

Example of continuous probability space

Height of a randomly chosen American male

- Probabilities of outcomes are not equal, and are described by a continuous function, $p(O)$

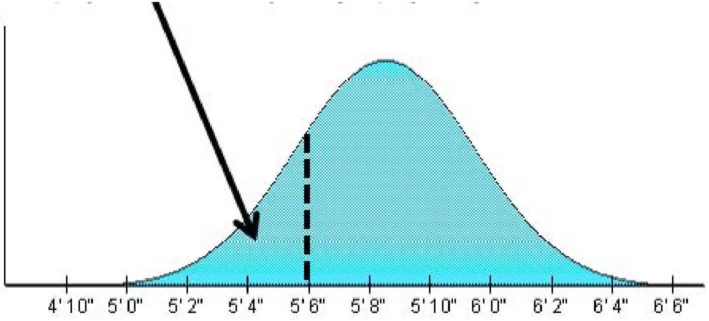


MLDS

Example of continuous probability space

Height of a randomly chosen American male

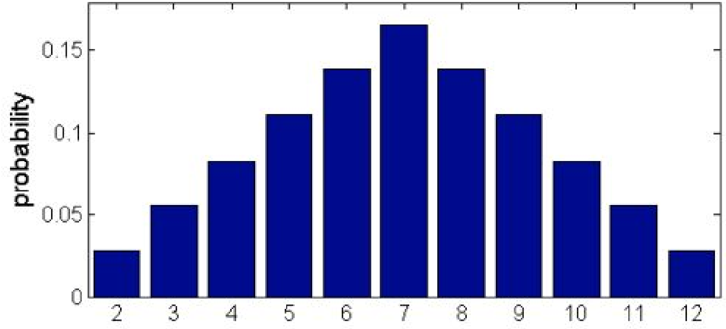
- $p(O)$ is a relative, not an *absolute* probability
  - $p(O)$ for any particular O is zero
  - $\int p(O)$ from $O = -\infty$ to ∞ (i.e. area under curve) is 1
  - Example: example: $p($ O = 5’ 8” $) > p($ O = 6’ 2” $)$
  - Example: $p($ O < 5’ 6” $) = (\int p($ O $)$ from O = - ∞ to 5’ 6” $) \approx 0.25$



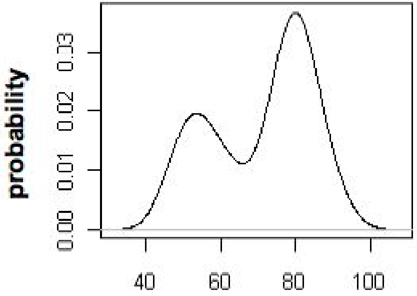
MLDS

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)



MLDS

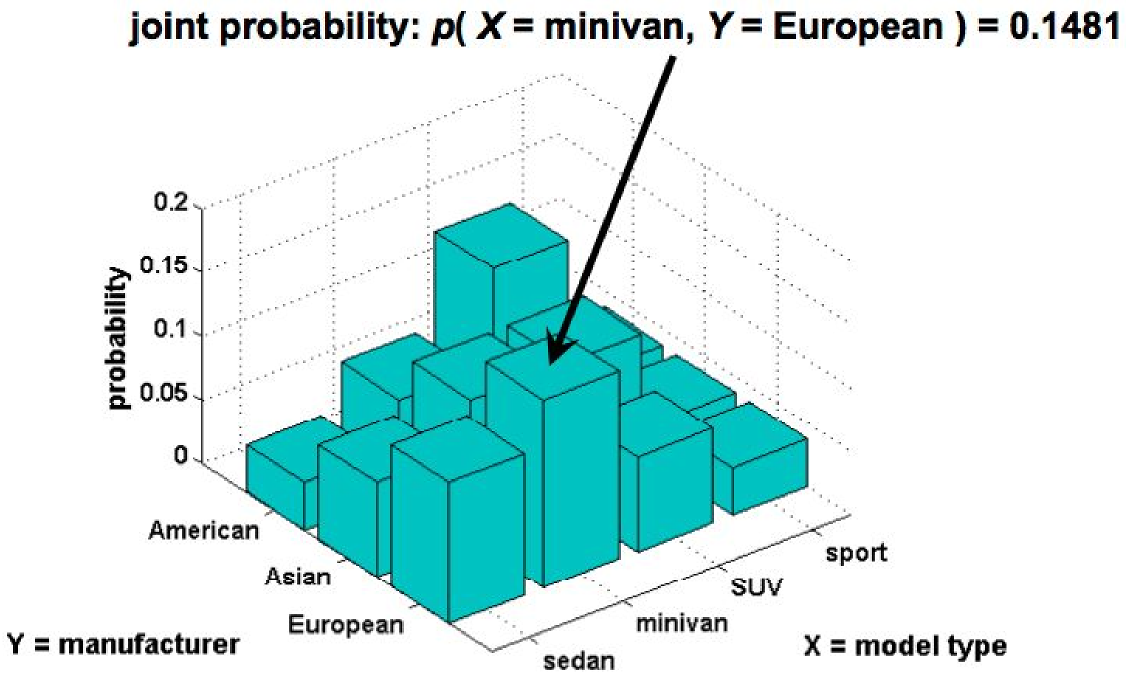
Multivariate probability distributions

- 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:

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

MLDS

Example of multivariate distribution



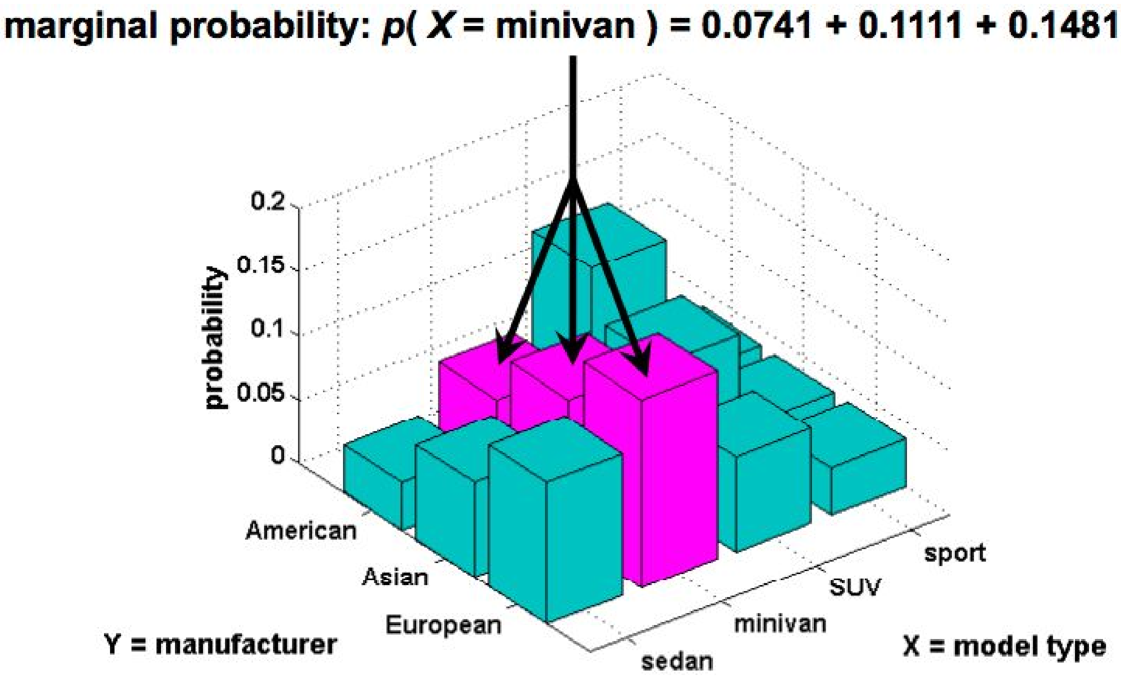
MLDS

Multivariate probability distributions

- *Marginal* probability
  - Probability distribution of a single variable in a joint distribution
  - Example: two random variables X and Y:
    $p(\ X = \mathrm{x}\ ) = \sum_{b=\text{all values of } Y} p(\ X = \mathrm{x},\ Y = \mathrm{b}\ )$
- *Conditional* probability
  - Probability distribution of one variable *given* that another variable takes a certain value
  - Example: two random variables X and Y:
    $p(\ X = \mathrm{x} \mid Y = \mathrm{y}\ ) = p(\ X = \mathrm{x},\ Y = \mathrm{y}\ )\ /\ p(\ Y = \mathrm{y}\ )$

MLDS

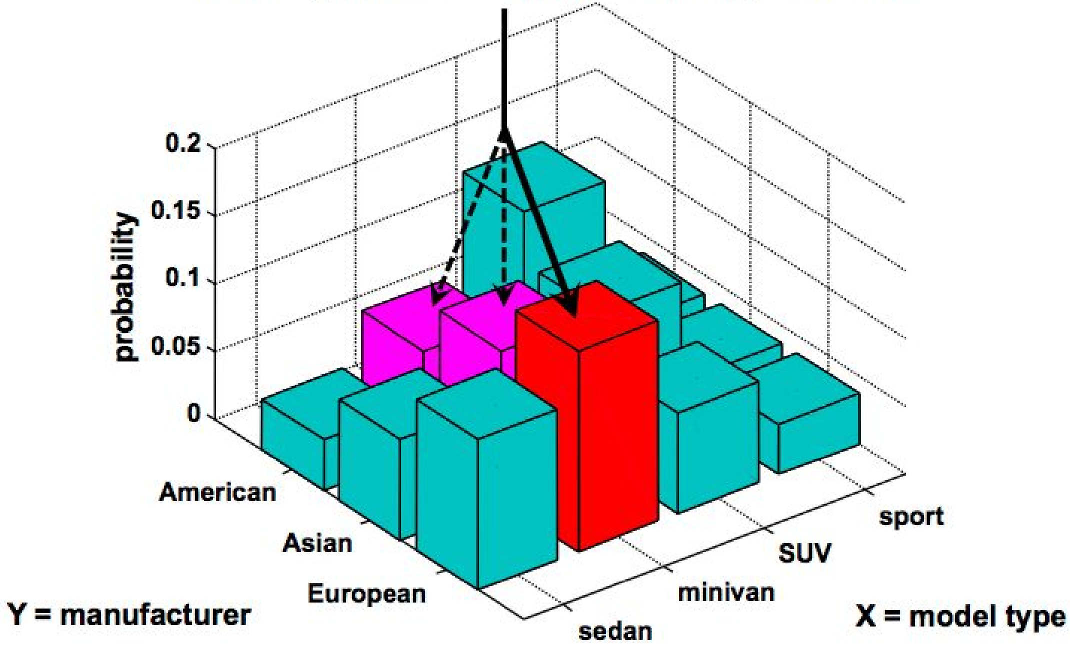
Example of marginal probability



MLDS

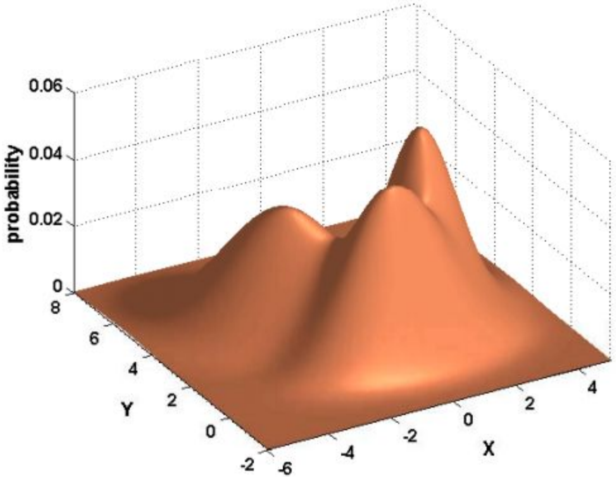
Example of conditional probability

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

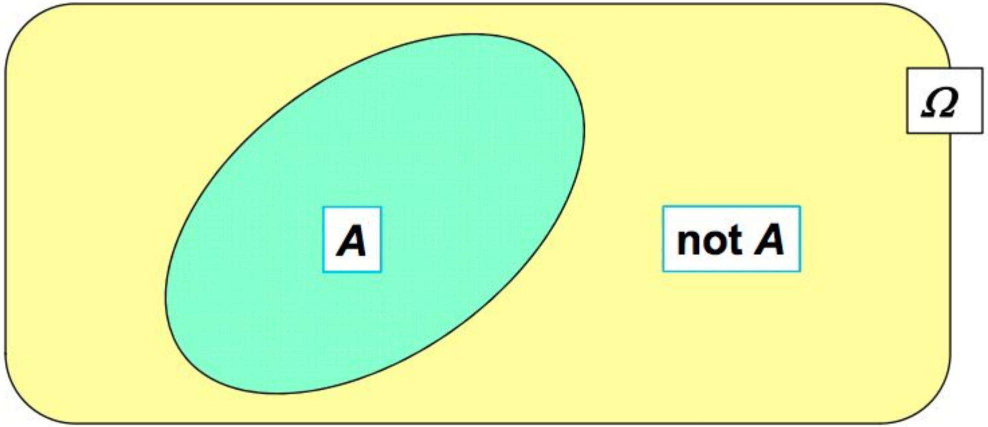


MLDS

Complement rule

Given: event A, which can occur or not

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



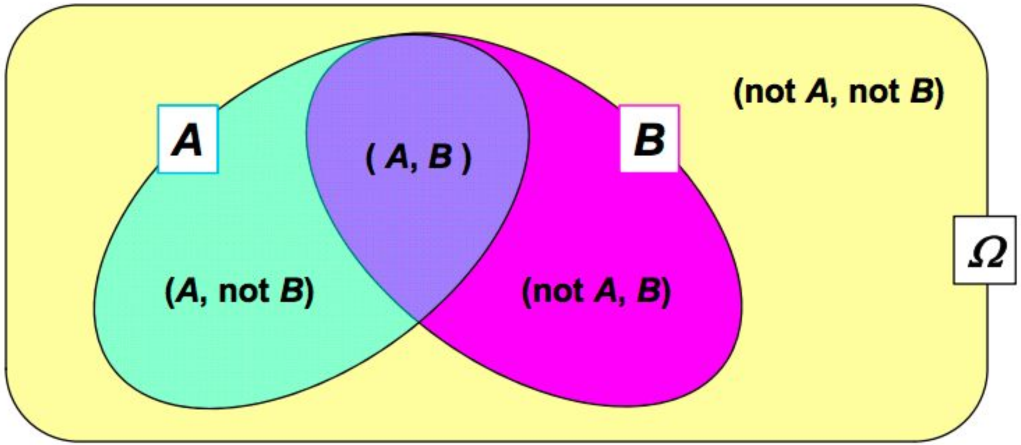
areas represent relative probabilities

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

MLDS

Example of product rule

Probability that a man has white hair (event *A*) **and** is over 65 (event *B*)

$$p(B) = 0.18$$

$$p(A \mid B) = 0.78$$

$$\begin{aligned} p(A, B) &= p(A \mid B) \cdot p(B) = \\ &0.78 \cdot 0.18 = \\ &0.14 \end{aligned}$$

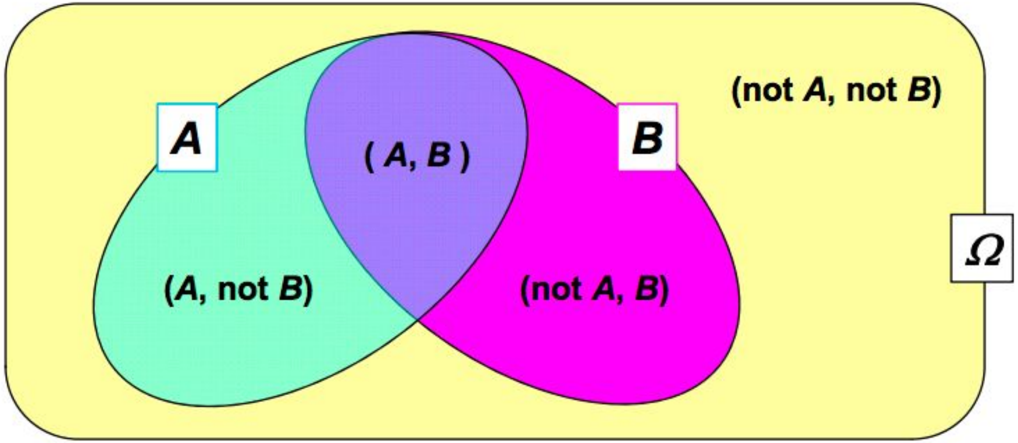
MLDS

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)



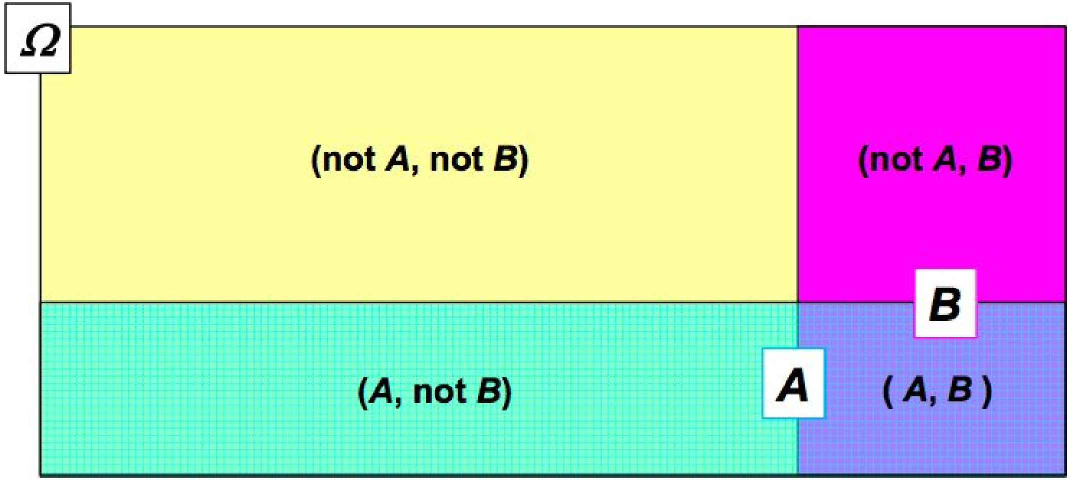
areas represent relative probabilities

MLDS

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

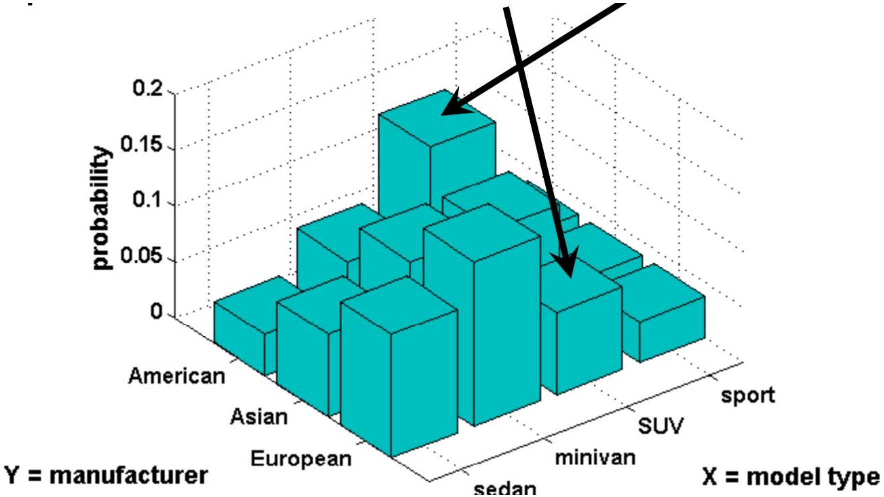
MLDS

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 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 / 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.



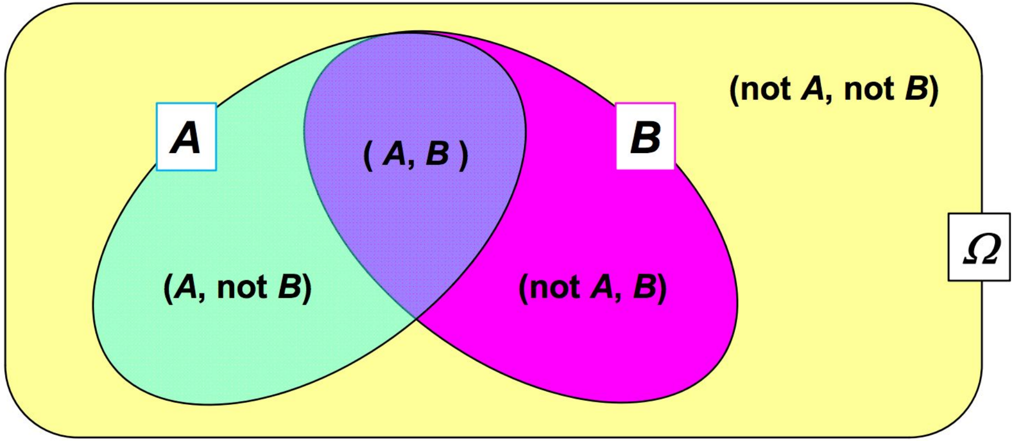
MLDS

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

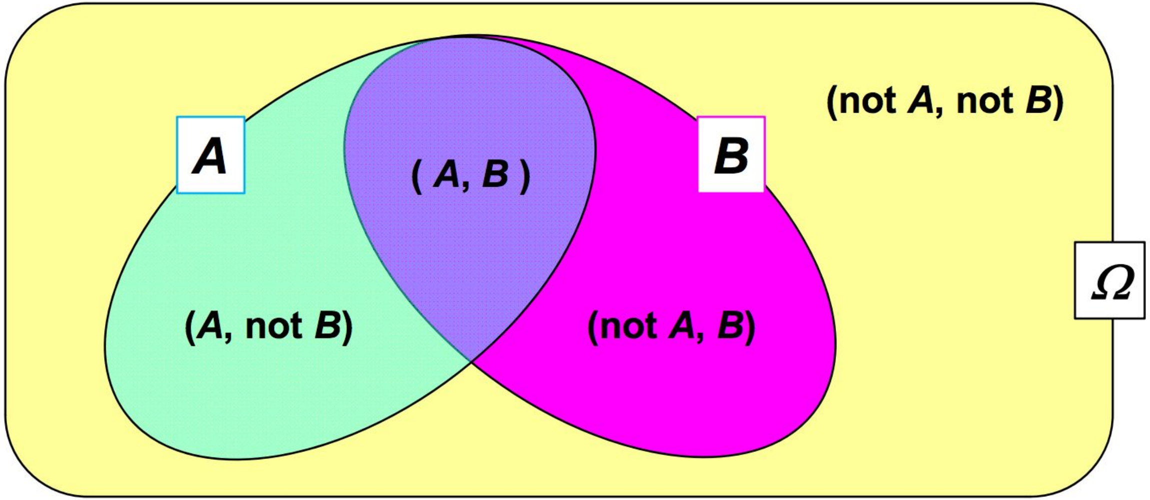


MLDS

Bayes rule

$$\text{posterior probability} \propto \text{likelihood} \times \text{prior probability}$$

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



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.

MLDS

Bayes rule

- 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.]

MLDS

Bayes rule

- 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 by Bayes rule:

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

$$p(A) = p(A \mid B) \cdot p(B) + p(A \mid \text{not } B) \cdot p(\text{not } B) = (0.9)(0.014) + (0.1)(0.986) = 0.111$$

$$p(B \mid A) = (0.9)(0.0137) \,/\, 0.111 = 0.111$$

- 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.

MLDS

Probabilities: when to add, when to multiply

- **ADD**: When you want to allow for occurrence of any of several possible outcomes of a 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.

MLDS

Linear algebra applications

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

* Applications 1 to 4 are directly relevant to this course.

MLDS

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 Column form is conventional
  - Vector elements referenced by subscript

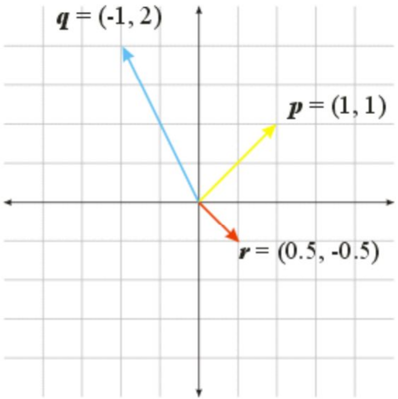
$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \mathbf{x}^{\mathrm{T}} = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$$

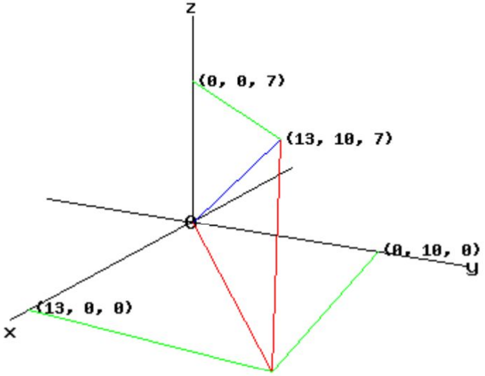
$^{\mathrm{T}}$ means "transpose"

MLDS

Vectors

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





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**

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

MLDS

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:

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

MLDS

Thanks!

Machine Learning for Data Science Interest Group
Advanced Informatics School
Universiti Teknologi Malaysia

@utmmlds
ais.utm.my/mlds

February 2017

MLDS