204456: Machine Learning

Ch02 - Maths refresher

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

Based on materials for CSS490 by Prof. Jeff Howbert

Objective

- To look at essential maths concepts for this course
  - Linear algebra
  - Statistics
  - Optimisation

Area of maths essential to ML

- Linear algebra
  - A study of vector/matrix
  - Data in ML is represented in vector/matrix form
- Statistics
  - Some say 'Machine learning is part of both statistics and CS'
  - Probability, statistical inference, validation
- Optimisation theory
  - The 'learning' part in machine learning
  - Rely hugely on calculus

Why worry about the maths ?

- You will know how to apply ML packages after this course
- However to get really useful results, you need
- to have good mathmatical intuition of ML principles
- to understand the working of those algorithms so that
  - know how to choose the right algorithm
  - know how to set hyper-parameters
  - troubleshoot poor results

Notations

$a \in A$	set membership: a is a member of set A
$\|B\|$	cardinality: number of items in set B
$\|\|\mathbf{v}\|\|$	norm: length of vector v
$\sum$	summation
$\int$	integral
$\mathcal{R}$	the set of real number
$\mathcal{R}^d$	real number space of dimension d

Notations

$\mathbf{x}, \mathbf{u}, \mathbf{v}$	vector (bold, lower case)
$\mathbf{X}, \mathbf{B}$	matrix (bold, upper case)
$y = f(x)$	function: assign unique value in set Y to each value in set X
$\frac{dy}{dx}$	derivative of y with respect to single variable x
$y = f(\mathbf{x})$	function in d-space
$\frac{\partial y}{\partial \mathbf{x}_i}$	partial derivative of y with respect to element i of $\mathbf{x}$

Linear algebra

Applications

- Operations on or between vectors and matrices
- Dimensionality reduction
- Linear regression
- Support Vector Machine

Why vector and matrices ?

- Most common form of data organisation for ML is 2D arry
  - rows represent samples (datapoints)
  - columns represent attributes (features)
- Natural to think of each sample as a vector of attributes and whole array as a matrix

Data 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

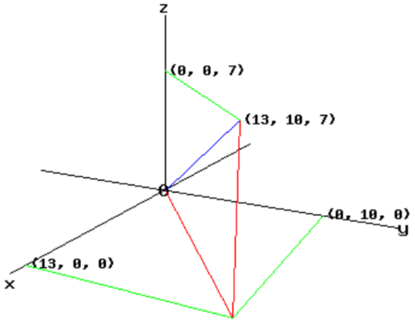
matrix

Vectors

- Definition: an d-tuple of values (usually real numbers)
  - d referred to as the dimension of the vector
  - d can be any positive integer, from 1 to infinity
- Can be written in column form (conventional) or row form
  - vector elements indexed by superscript
- $\mathbf{x}_i = \begin{bmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^d \end{bmatrix} \qquad \mathbf{x}_i^T = [x_i^1, x_i^2, \ldots, x_i^d]$

Vectors

can think of a vector as a point in space



Vector arithmetic

- Addition of two vectors
  - add corresponding elements
  - $\mathbf{z} = \mathbf{x} + \mathbf{y} = (x^1 + y^1, \cdots, x^d + y^d)^T$
  - result is a vector
- Scalar multiplication
  - multiply each element by scalar
  - $\mathbf{y} = a\mathbf{x} = (ax^1, \cdots, ax^d)^T$
  - result is a vector

Vector arithmetic

- Dot product of two vectors
  - multiply corresponding elements, then add products
  - $a = \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{d} x^i y^i$
  - result is <u>scalar</u>

Matrices

- Definition: an $n \times d$ two-dimensional array of values (usually real numbers)
  - n rows, d columns
- Matrix referenced by two-element subscript
  - first element in subscript is row
  - second element is column
  - $\mathbf{A}_{13}$ or a_{13} is element in the first row, third column of $\mathbf{A}$

Matrices

- A vector can be regarded as special case of a matrix, where one of matrix dimensions = 1
- Matrix transpose (denoted A^T)
  - swap columns and rows
  - $n \times d$ matrix becomes $d \times n$ matrix
  - $\mathbf{A} = \begin{bmatrix} 2 & 7 & 3 \\ 1 & -2 & 5 \end{bmatrix}$
  - $\mathbf{A}^T = \begin{bmatrix} 2 & 1 \\ 7 & -2 \\ 3 & 5 \end{bmatrix}$

Matrix arithmetic

- Addition of two matrices
  - $\mathbf{C} = \mathbf{A} + \mathbf{B}$
  - $c_{ij} = a_{ij} + b_{ij}$
  - result is a matrix of same size
- Scalar multiplication of matrix
  - $\mathbf{B} = a \cdot \mathbf{C}$
  - $b_{ij} = a \cdot c_{ij}$
  - result is a matrix of same size

Matrix multiplication

- TO THE BOARD
- Multiplication is associative: $\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}) = (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C}$
- Not commutative: $\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}$
- Transposition rule: $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$

Matrix multiplication

- 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.

- Example: **A** 3×5, **B** 5×5, **C** 3×1

  Right: $\mathbf{A} \cdot \mathbf{B} \cdot \mathbf{A}^T$

  Wrong: $\mathbf{C} \cdot \mathbf{A} \cdot \mathbf{B}$

Statistics

Concept of probability

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 outcome happens more often than another outcome we say it is more probable.

Probability spaces

- A probability space is a random process or experiment with three components:
  - Ω, the set of possible outcomes
    - number of possible outcomes $= |\Omega| = N$
  - F, the set of possible events E
    - an event comprises 0 to N outcomes
    - think of as a dichotomy of 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

Axioms of probability

1. Non-negativity
   - $p(E) \geq 0$ for all $E \in F$
2. All possible outcomes $p(\Omega) = 1$
3. Additivity of disjoint events: for all events $E, E' \in F$ where $E \cap E' = \emptyset$, $p(E \cup E') = p(E) + p(E')$

Types of probability spaces

- Discrete space $|\Omega|$ is finite
- Continuous space $|\Omega|$ is infinite

Example of discrete probability space

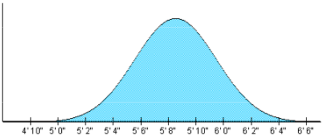
Single roll of a six-sided die (singular of dice)

- 6 possible outcomes: $O = \{1, 2, 3, 4, 5, 6\}$
- $2^6 = 64$ possible events
  - $E = \{O \in \{1, 3, 5\}\}$ outcome is odd
- If die if fair, $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = 1/6$

Example of continuous probability space

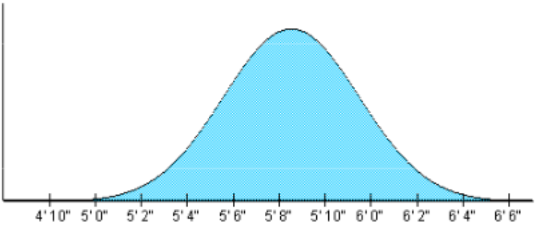
Height of randomly chosen Thai male

- Infinite number of outcomes
- Infinite number of events
  - $E = \{O | O < 160\}$ individual chosen is smaller than 160 cm.
- Probabilities of outcomes are not equal, and are described by a continuous function, $p(O)$



Example of continuous probability space

Height of randomly chosen Thai male



- $p(O)$ is relative not absolute
- $p(O = 175) = 0$
- but we can still make comparison $p(O = 170) > p(O = 180)$?

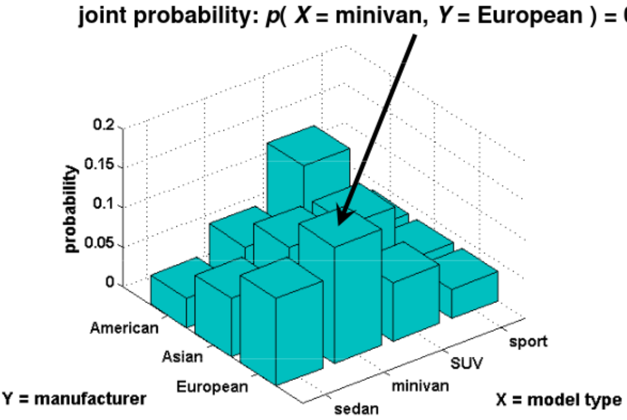
Random variables

- A random variable X is a function that associates a number (label) x with each outcome O of a process
- Basically a way to redefine (usually simplify) a probability space to a new probability space
- Example $X =$ number of heads in three coin flips
  - possible values of X are 0,1,2,3
- Example $X =$ region of car manufacturer
  - Original outcomes could be a set of countries
  - possible values of X are 1=European, 2=Asia, 3=America

Multivariate probability distribution

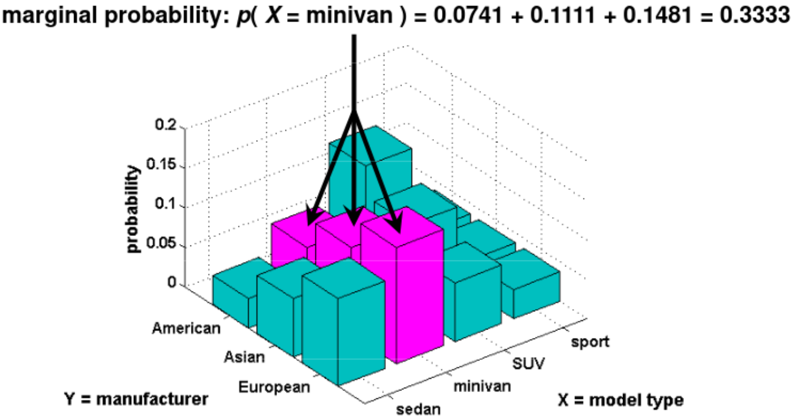
- Scenario
  - Several random processes occur
  - Want to know probabilities for each possible combination of outcomes.
- Can describe as joint probability of random variables
  - two processes whose outcomes are represented by random variables X and Y, Probability that process X has outcome x and process Y has outcome y is denoted as: $p(X = x, Y = y)$

Example of multivariate distribution



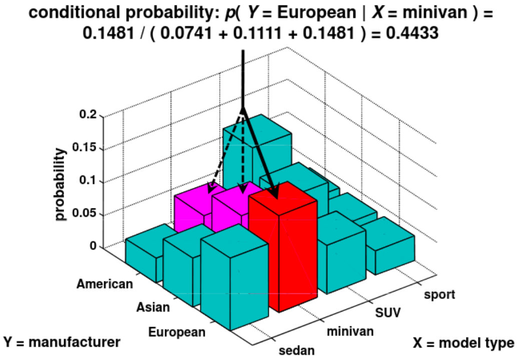
Multivariate probability distribution

- Marginal probability
  - Probability distribution of a single variable in a joint distribution
  - $p(X = x) = \sum_y p(X = x, Y = y)$



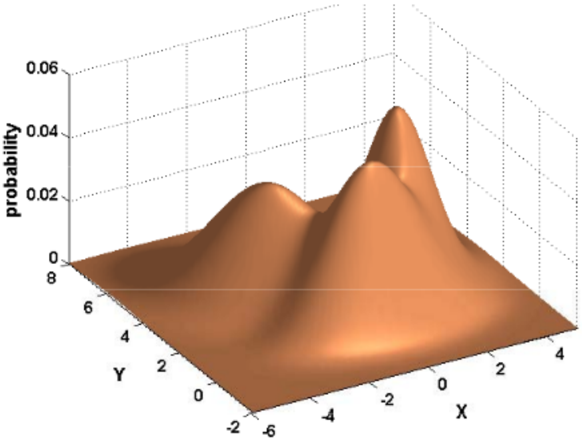
Multivariate probability distribution

- Conditional probability
  - Probability distribution of one variable given that another variable takes a certain value
  - $p(X = x | Y = y) = p(X = x, Y = y)/p(Y = y)$



Continuous probability distribution

- Same concepts of joint, marginal, and conditional probabilities apply (except use integrals)



Expected value

Given

- A discrete random variable X, with possible values $x = x_1, x_2, \dots, x_n$
- Probability $p(X = x_i)$
- A function $y_i = f(x_i)$ defined on X

Expected value is the probability-weighted "average" of $f(x_i)$

$$E(f) = \sum_i p(x_i) \cdot f(x_i) \tag{1}$$

Calculus

Derivative

A derivative of function at x_0 is the rate of change of function values as input changes near x_0

$$\frac{dy}{dx} = \frac{df(x_o)}{dx} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

Partial Derivative

A partial derivative of multivariate function at $\mathbf{x}_0$ is the rate of change of function values as the i-th component of the changes near $\mathbf{x}_0$

$$\frac{\partial y}{\partial x_i} = \frac{\partial f(\mathbf{x}_0)}{dx_i} = \lim_{h_i \to 0} \frac{f(\mathbf{x}_0 + h_i) - f(\mathbf{x}_0)}{h_i}$$

h_i is infinitesimal for component i

Common derivatives

- $\frac{daf(x)}{dx} = a\frac{df(x)}{dx}$
- $\frac{dx^k}{dx} = kx^{k-1}$
- $\frac{df(x)+g(x)}{dx} = \frac{df(x)}{dx} + \frac{dg(x)}{dx}$
- $\frac{df(x)g(x)}{dx} = f(x)\frac{dg(x)}{dx} + g(x)\frac{df(x)}{dx}$
- $\frac{de^x}{dx} = e^x$
- $\frac{d\log(x)}{dx} = \frac{1}{x}$

Gradient

A vector of partial derivatives

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{dx_1} \\ \frac{\partial f(\mathbf{x})}{dx_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{dx_d} \end{bmatrix}$$

Hessian

A matrix of second partial derivatives

$$\mathbf{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_D} \\ \frac{\partial^2 f}{\partial w_1 \partial w_2} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_D} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 f}{\partial w_1 \partial w_D} & \frac{\partial^2 f}{\partial w_2 \partial w_D} & \cdots & \frac{\partial^2 f}{\partial w_D^2} \end{bmatrix}$$

References

- Math for Machine learning by Hal Daume III http://users.umiacs.umd.edu/~hal/courses/2013S_ML/math4ml.pdf
- Machine learning math essentials by Jeff Howbert http://courses.washington.edu/css490/2012.Winter/lecture_slides/02_math_essentials.pdf