

Lecture 1. StatML Welcome and Maths Review

COMP90051 Statistical Machine Learning

Lecturer: Jean Honorio



THE UNIVERSITY OF
MELBOURNE

This lecture

- About COMP90051
- Review: Probability theory
- Review: Linear algebra
- Review: Sequences and limits

Subject objectives

- Develop an appreciation for the role of statistical ML, advanced foundations and applications
- Gain an understanding of a representative selection of ML techniques – *how ML works*
- Be able to design, implement and evaluate ML systems
- Become a discerning ML consumer

Subject content

- The subject will cover topics from
Foundations of statistical learning, linear models, non-linear bases, regularised linear regression, generalisation theory, kernel methods, deep neural nets, experts, multi-armed bandits, Bayesian learning, probabilistic graphical models
- Classical / in-person teaching
 - * Theory in lectures (2x per week), subject assessment;
 - * Hands-on experience in weekly workshop (starts week 2), subject assessment;
 - * *Strongly recommend you keep up (viz. quizzes)*

Subject staff / Contact

Lecturers:	<u>Jean Honorio</u> [Week 1 to Week 5] Zahra Dasht Bozorgi [Week 6 to first lecture of Week 10] Ben Rubinstein [Second lecture of Week 10 to Week 12]
Tutors:	Shijie Liu (Head Tutor), Qianjun Ding, Jinhao Li, Paul Ou, Xunye Tian, Jedwin Villanueva <i>See Canvas for latest list and contact details.</i>
Ed discussion	<i>Will open after this first lecture</i>
Contacting staff	<i>Combined staff email</i> <u>comp90051-2024s2-staff@lists.unimelb.edu.au</u>

Advanced ML: Expected Background

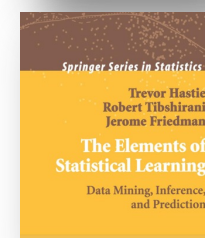
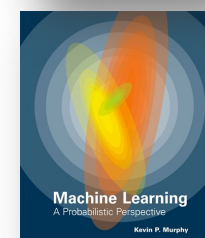
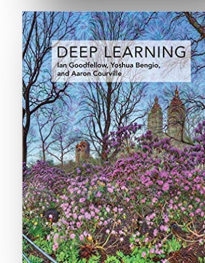
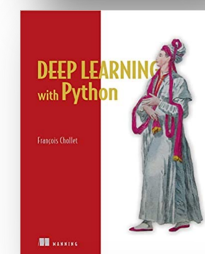
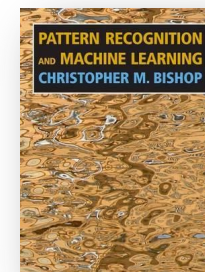
- **ML:** COMP90049, COMP20008, MAST90105
- **Algorithms & complexity:** big-O, termination; basic data structures & algorithms
- **Solid coding,** ideally experience in Python
- **Mathematics and statistics**

Maths / stats requirements

- **Probability theory**: probability calculus; discrete/continuous distributions; multivariate; exponential families; Bayes rule
- **Sequences**: sequences, limits, supremum
- **Linear algebra**: vector inner products & norms; orthonormal bases; matrix operations, inverses, eigenvectors/values
- **Calculus & optimisation**: partial derivatives; gradient descent; convexity; Lagrange multipliers

Textbooks

- We **don't have only one reference**. We prefer to pick good bits from several. We may also supplement with other readings as we go.
- All are available free online or through the library digitally. See the **Canvas subject overview** for links. Therefore, **no need to buy**.
- Primarily we refer to (good all rounder): ***Bishop (2007) Pattern Recognition and Machine Learning***
- ***Practical Deep Nets: Chollet (2017) Deep learning with Python***
- ***More deep learning detail: Goodfellow, Bengio, Courville (2016) Deep learning***
- For more on PGMs/Bayesian inference: Murphy (2012) *Machine Learning: A Probabilistic Perspective*
- For reference on frequentist ideas, SVMs, lasso, etc.:
Hastie, Tibshirani, Friedman (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*
- For reference on learning theory (though goes deeper/broader than we need)
Shalev-Shwartz and Ben-David *"Understanding Machine Learning: From Theory to Algorithms"*

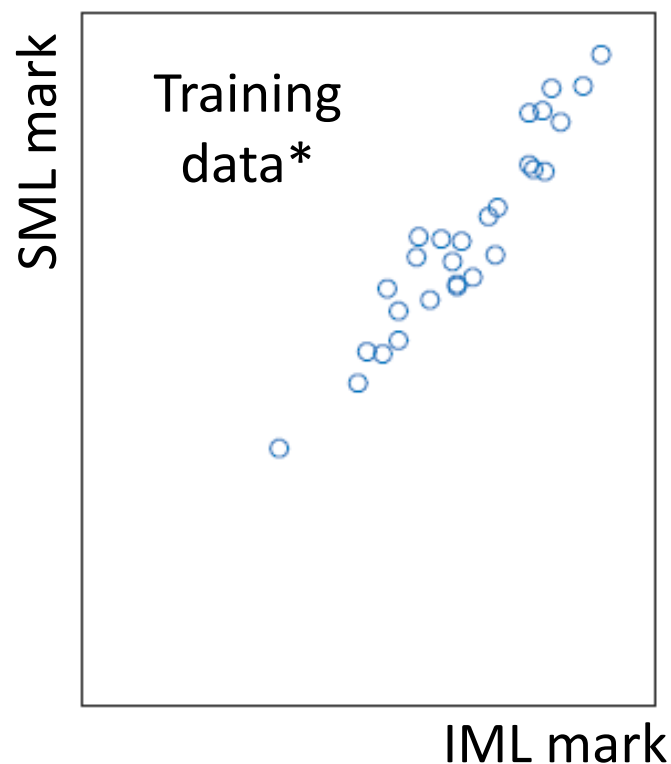


Assessment

- Assessment components
 - * Timing may differ from handbook due to non-teaching week
 - * Two assignments – w5-8 (12.5% each, individual)
 - ~1.5 weeks to complete
 - One written assignment, one programming assignment
 - * A group project – w8-12 (25%, group size 3)
 - ~4 weeks to complete
 - * Quizzes 30min fortnightly (10% - best 5 of 6)
 - Q1 to **open August 4 at 11.59pm** and **close August 5 at 11.59pm**
 - * Final Exam (40%): 2hr open book **on-campus exam** (40%)
- 50% hurdles applied to **exam**

Probability theory

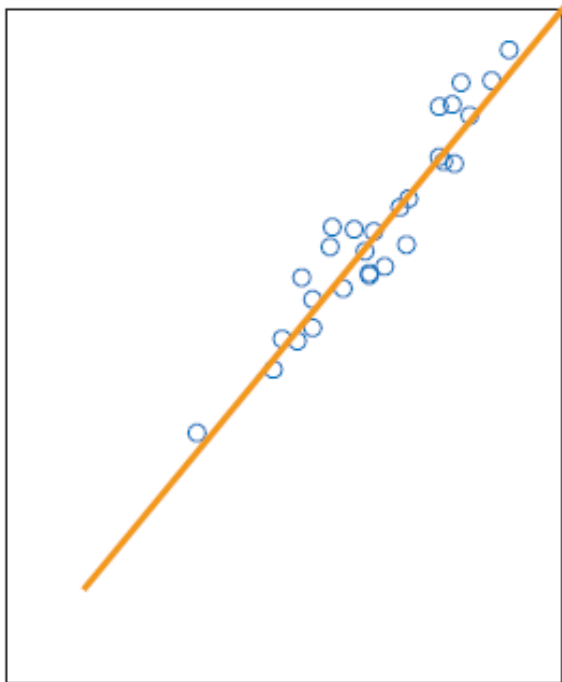
Data is noisy (almost always)



- Example:
 - * given mark for Intro ML (IML)
 - * predict mark for Stat Machine Learning (SML)

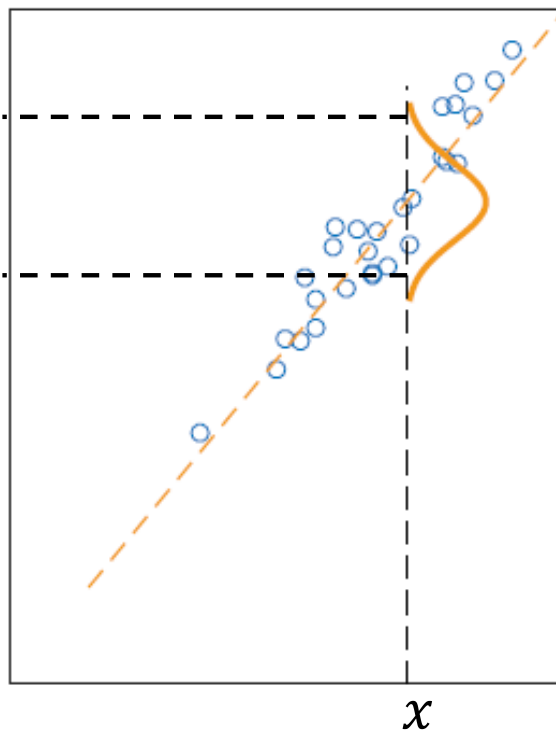
* synthetic data :)

Types of models



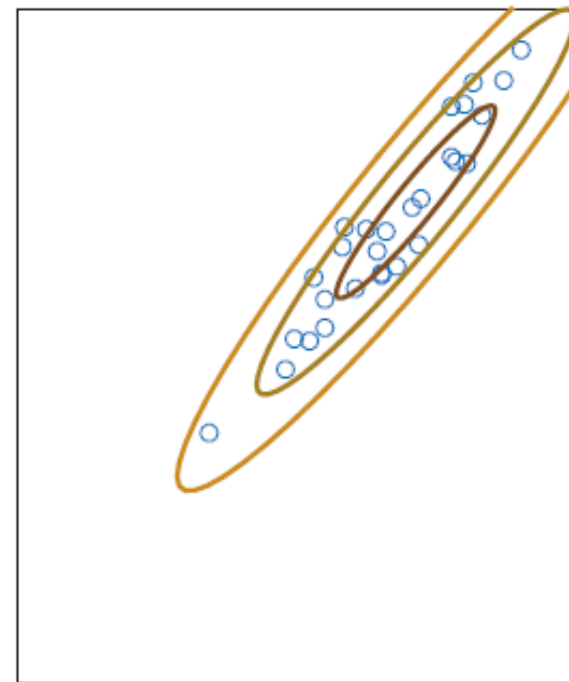
$$\hat{y} = f(x)$$

IntroML mark was 95,
SML mark is predicted
to be 95



$$P(y|x)$$

IntroML mark was 95,
SML mark is likely to
be in (92, 97)



$$P(x, y)$$

probability of having
($IML = x, SML = y$)

Basics of probability theory



- A probability space:
 - * Set Ω of possible outcomes
 - * Set F of events (subsets of outcomes)
 - * Probability measure $P: F \rightarrow \mathbb{R}$
- Example: a die roll
 - * $\{1, 2, 3, 4, 5, 6\}$
 - * $\{ \varnothing, \{1\}, \dots, \{6\}, \{1,2\}, \dots, \{5,6\}, \dots, \{1,2,3,4,5,6\} \}$
 - * $P(\varnothing)=0$, $P(\{1\})=1/6$, $P(\{1,2\})=1/3$, ...

Axioms of probability

F contains all of: Ω ; all complements $\Omega \setminus f$, $f \in F$; the union of any countable set of events in F .

1. $P(f) \geq 0$ for every event $f \in F$.
2. $P(\cup_f f) = \sum_f P(f)$ for all countable sets of pairwise disjoint events.
3. $P(\Omega) = 1$

Random variables (r.v.'s)



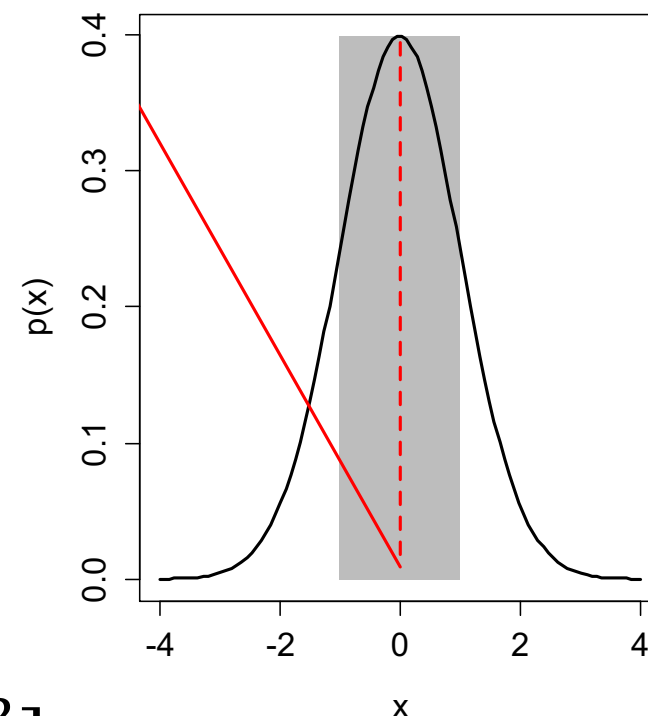
- A random variable X is a numeric function of outcome $X(\omega) \in \mathbf{R}$
- $P(X \in A)$ denotes the probability of the outcome being such that X falls in the range A
- Example: X winnings on \$5 bet on even die roll
 - * X maps 1,3,5 to -5
 - X maps 2,4,6 to 5
 - * $P(X=5) = P(X=-5) = \frac{1}{2}$

Discrete vs. continuous distributions

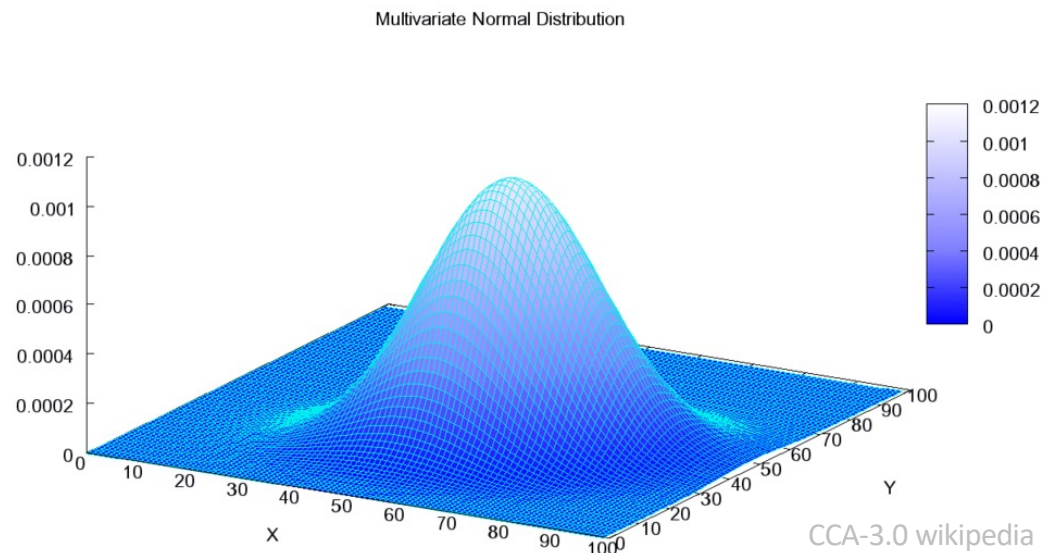
- Discrete distributions
 - * Govern r.v. taking discrete values
 - * Described by **probability mass function** $p(x)$ which is $P(X=x)$
 - * $P(X \leq x) = \sum_{a=-\infty}^x p(a)$
 - * **Examples:** Bernoulli, Binomial, Multinomial, Poisson
- Continuous distributions
 - * Govern real-valued r.v.
 - * Cannot talk about PMF but rather **probability density function** $p(x)$
 - * $P(X \leq x) = \int_{-\infty}^x p(a) da$
 - * **Examples:** Uniform, Normal, Laplace, Gamma, Beta, Dirichlet

Expectation

- Expectation $E[X]$ is the r.v. X 's “average” value
 - * Discrete: $E[X] = \sum_x x P(X = x)$
 - * Continuous: $E[X] = \int_x x p(x) dx$
- Properties
 - * Linear: $E[aX + b] = aE[X] + b$
 $E[X + Y] = E[X] + E[Y]$
 - * Monotone: $X \geq Y \Rightarrow E[X] \geq E[Y]$
- Variance: $Var(X) = E[(X - E[X])^2]$



Multivariate distributions

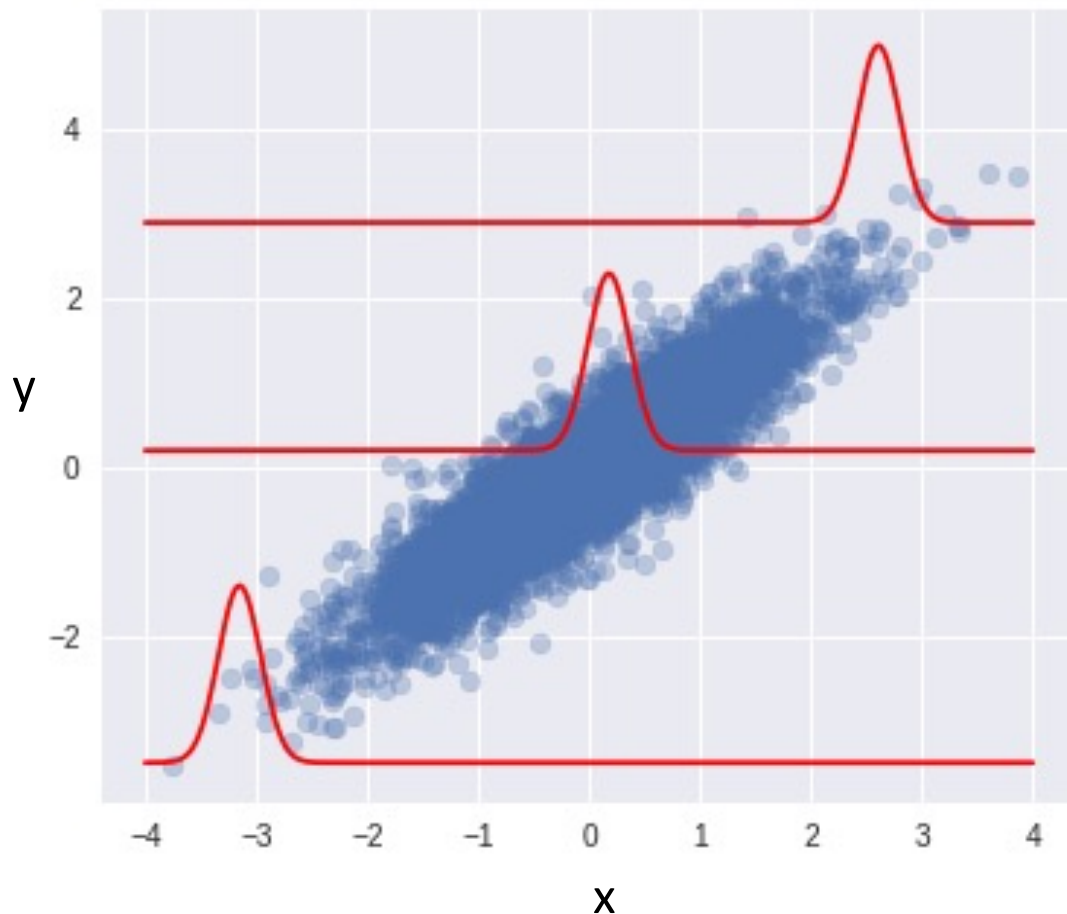


- Specify joint distribution over multiple variables
- Probabilities are computed as in univariate case, we now just have repeated summations or repeated integrals
- Discrete: $P((X, Y) \in A) = \sum_{(x,y) \in A} p(x, y)$
- Continuous: $P((X, Y) \in A) = \int_A p(x, y) dx dy$

Independence and conditioning

- X, Y are **independent** if
 - * $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$
 - * Similarly for densities:
 $p_{X,Y}(x, y) = p_X(x)p_Y(y)$
 - * **Intuitively**: knowing value of Y reveals nothing about X
 - * **Algebraically**: the joint on X, Y factorises!
- **Conditional probability**
 - * $P(A|B) = \frac{P(A \cap B)}{P(B)}$
 - * Similarly for densities
 $p(y|x) = \frac{p(x,y)}{p(x)}$
 - * **Intuitively**: probability event A will occur given we know event B has occurred
 - * **The Product Rule**:
 $P(A \cap B) = P(A|B) P(B)$
 - * X, Y independent equiv to
 $P(Y = y|X = x) = P(Y = y)$

Independence and conditioning



- Not independent: knowing the value of y gives you insight into the value that x will take

Inverting conditioning: Bayes' Theorem

- In terms of events A, B

- * $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

- * $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$



Bayes

- Simple rule that lets us swap conditioning order

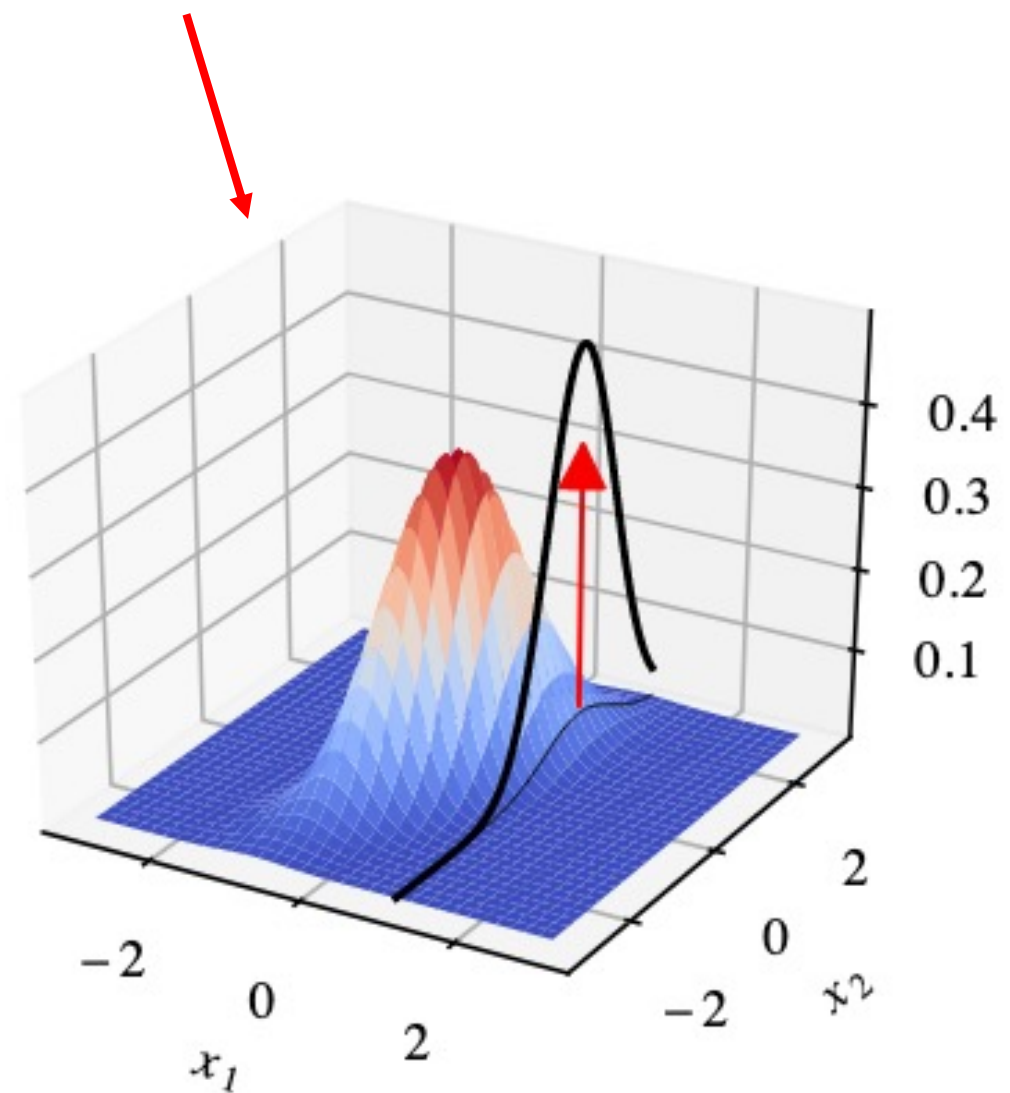
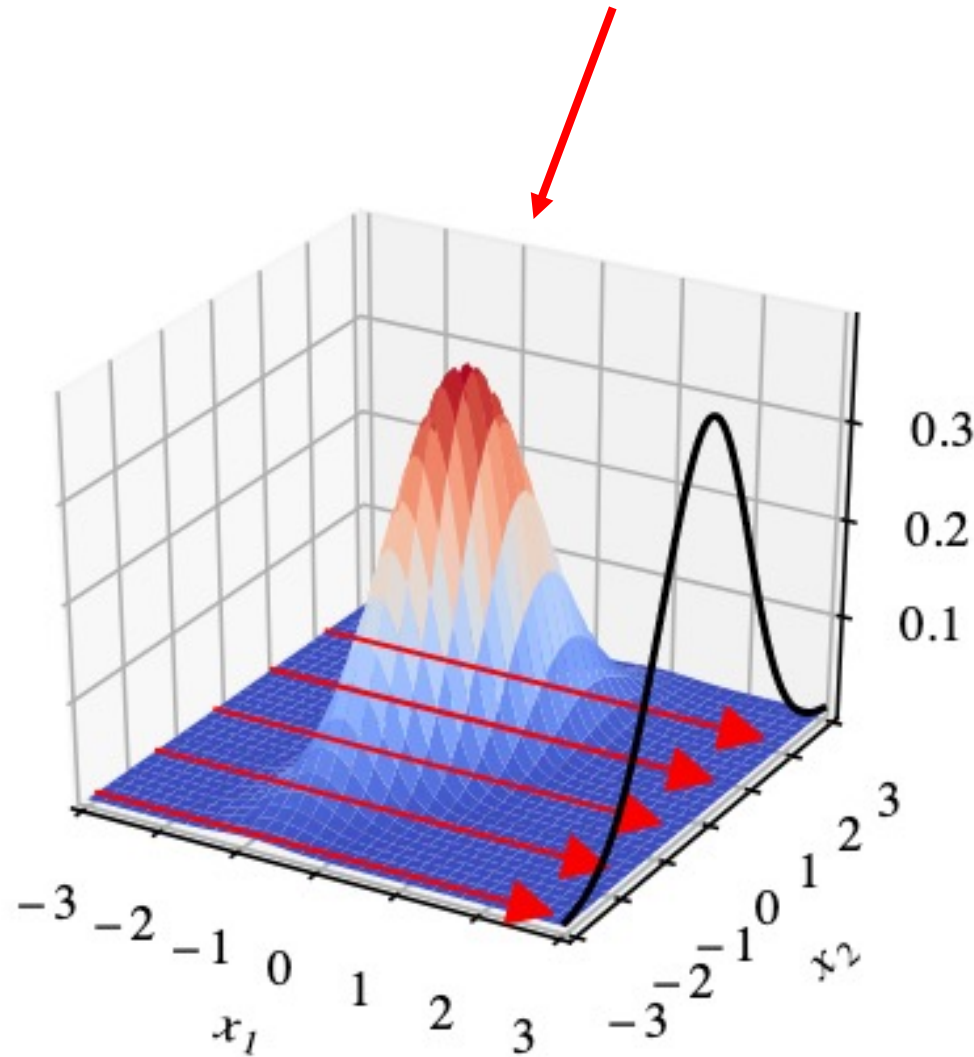
Inverting conditioning: Bayes' Theorem



Bayes

- In terms of events A, B
 - * $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
 - * $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Simple rule that lets us swap conditioning order
- Probabilistic and Bayesian inference make heavy use
 - * **Marginals**: probabilities of individual variables
 - * **The Sum Rule** of marginalisation: summing away all but r.v.'s of interest. $P(A) = \sum_b P(A, B = b)$

Marginalisation vs conditioning



Mini Summary

- Probability spaces, axioms of probability
- Discrete vs continuous; Univariate vs multivariate
- Expectation, Variance
- Independence and conditioning
- Bayes rule and marginalisation

Next: Linear algebra primer/review

Vectors

Link between geometric and algebraic
interpretation of ML methods

What are vectors?

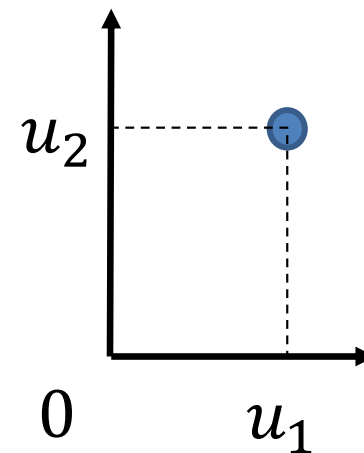
Suppose $\mathbf{u} = [u_1, u_2]'$. What does \mathbf{u} really represent?



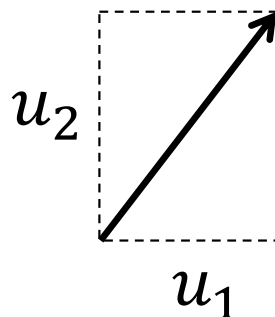
Ordered set of numbers $\{u_1, u_2\}$



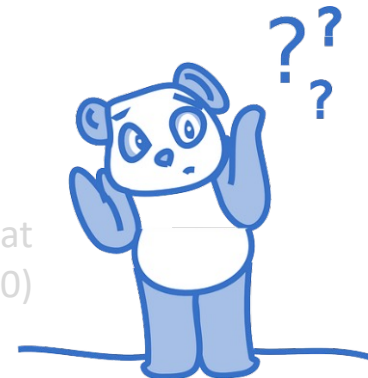
Cartesian coordinates of a point



A direction



art: OpenClipartVectors at
pixabay.com (CC0)



Dot product: Algebraic definition

- Given two m -dimensional vectors \mathbf{u} and \mathbf{v} , their dot product is $\mathbf{u} \cdot \mathbf{v} \equiv \mathbf{u}'\mathbf{v} \equiv \sum_{i=1}^m u_i v_i$

* E.g., weighted sum of terms is a dot product $\mathbf{x}'\mathbf{w}$

- If k is a scalar, \mathbf{a} , \mathbf{b} , \mathbf{c} are vectors then

$$(k\mathbf{a})'\mathbf{b} = k(\mathbf{a}'\mathbf{b}) = \mathbf{a}'(k\mathbf{b})$$

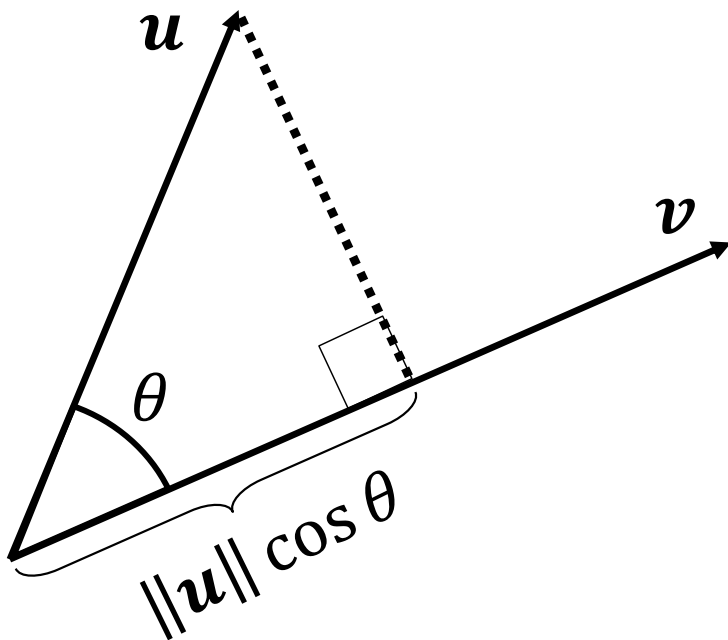
$$\mathbf{a}'(\mathbf{b} + \mathbf{c}) = \mathbf{a}'\mathbf{b} + \mathbf{a}'\mathbf{c}$$

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 3 & 4 \end{pmatrix}$$

$$2 \times 1 \quad 1 \times 2 \quad \text{ⓔ}$$

Dot product: Geometric definition

- Given two m -dimensional Euclidean vectors \mathbf{u} and \mathbf{v} , their dot product is $\mathbf{u} \cdot \mathbf{v} \equiv \mathbf{u}'\mathbf{v} \equiv \|\mathbf{u}\|\|\mathbf{v}\| \cos \theta$
 - * $\|\mathbf{u}\|, \|\mathbf{v}\|$ are L_2 norms for \mathbf{u}, \mathbf{v} also written as $\|\mathbf{u}\|_2$
 - * θ is the angle between the vectors



The *scalar projection* of \mathbf{u} onto \mathbf{v} is given by

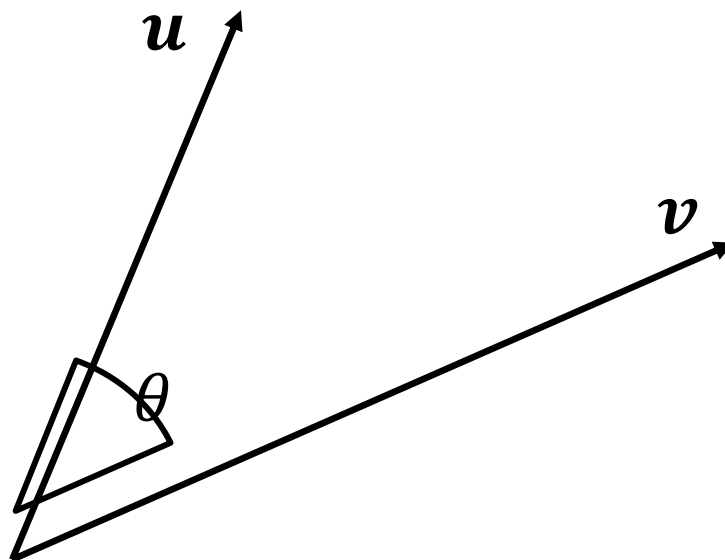
$$u_v = \|\mathbf{u}\| \cos \theta$$

Thus dot product is

$$\mathbf{u}'\mathbf{v} = u_v \|\mathbf{v}\| = v_u \|\mathbf{u}\|$$

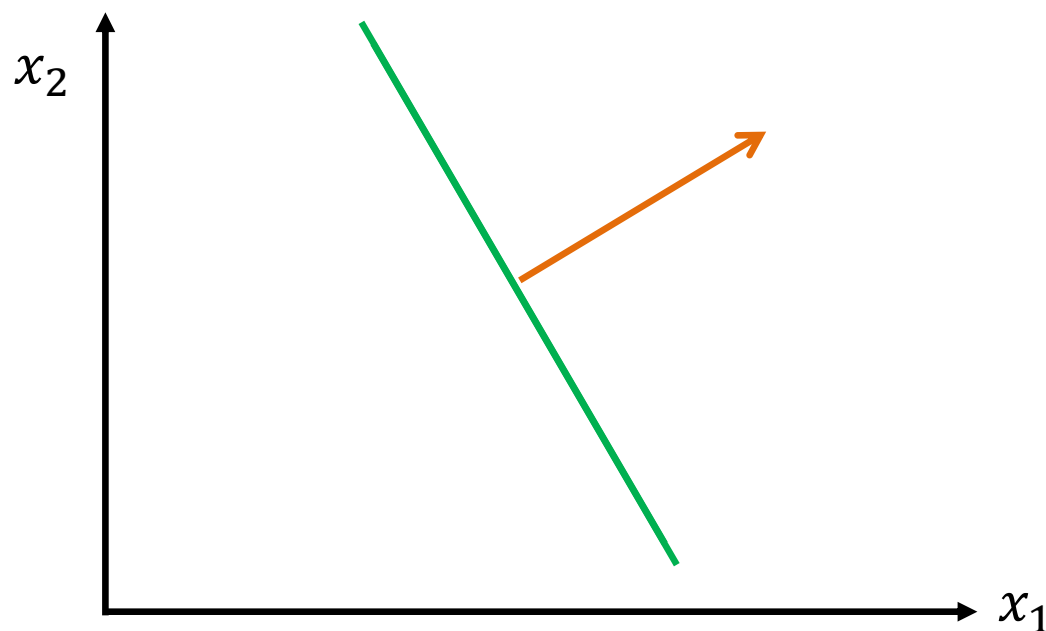
Geometric properties of the dot product

- If the two vectors are orthogonal then $\mathbf{u}'\mathbf{v} = 0$
- If the two vectors are parallel then $\mathbf{u}'\mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\|$, if they are anti-parallel then $\mathbf{u}'\mathbf{v} = -\|\mathbf{u}\|\|\mathbf{v}\|$
- $\mathbf{u}'\mathbf{u} = \|\mathbf{u}\|^2$, so $\|\mathbf{u}\| = \sqrt{\mathbf{u}'\mathbf{u}} = \sqrt{u_1^2 + \dots + u_m^2}$ defines the Euclidean vector length



Hyperplanes and normal vectors

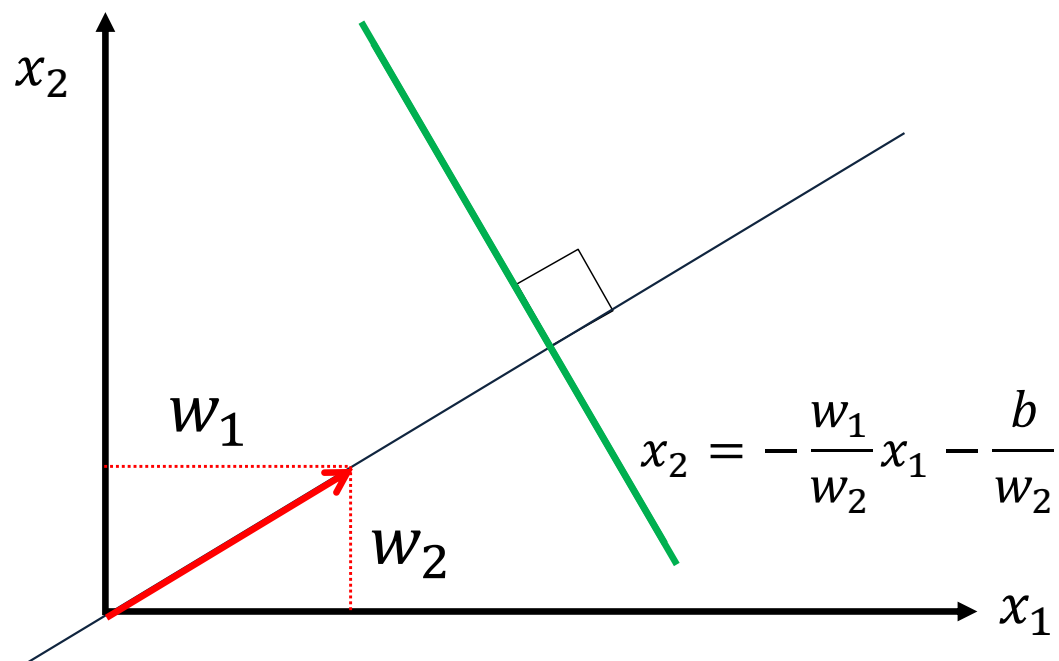
- A hyperplane defined by parameters \mathbf{w} and b is a set of points \mathbf{x} that satisfy $\mathbf{x}'\mathbf{w} + b = 0$
- In 2D, a hyperplane is a line: a line is a set of points that satisfy $w_1x_1 + w_2x_2 + b = 0$



A normal vector for a hyperplane is a vector perpendicular to that hyperplane

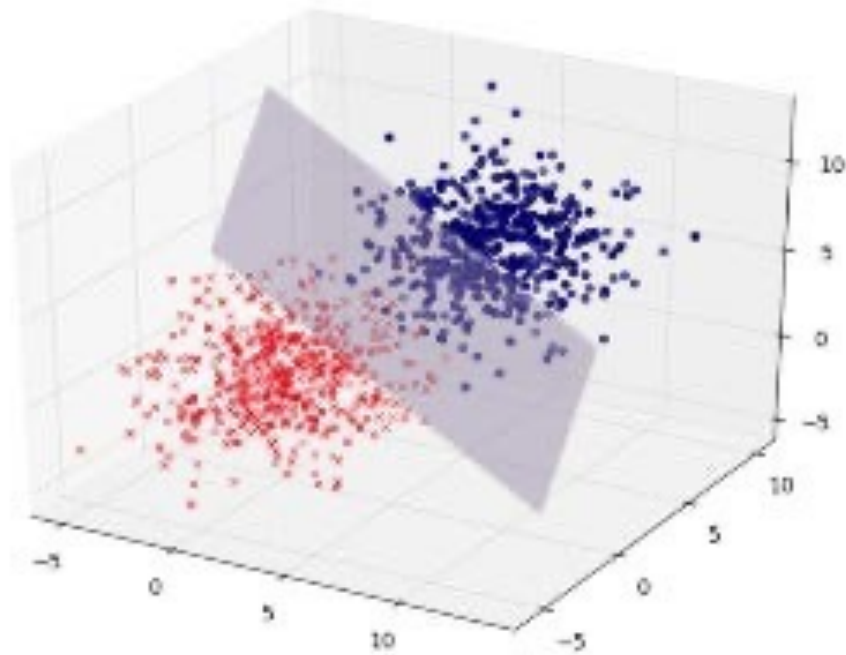
Hyperplanes and normal vectors

- Consider a line defined by w_1 , w_2 and b
- Vector $\mathbf{w} = [w_1, w_2]'$ is a normal vector

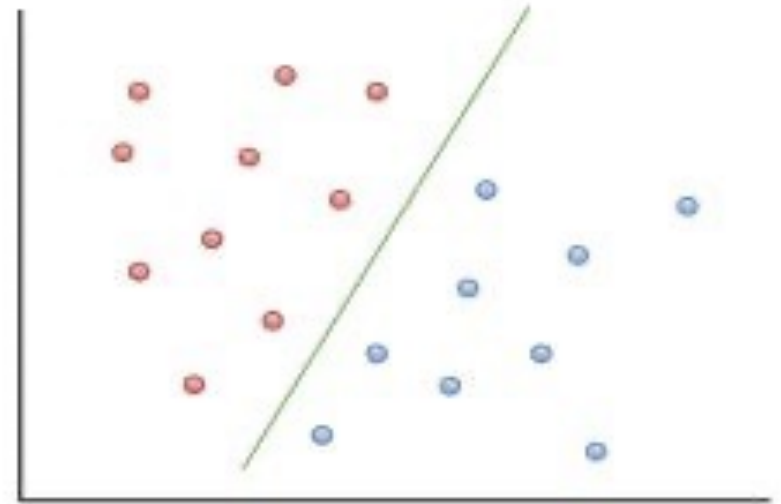


Hyperplane vs line

Hyperplane



Line



L_1 and L_2 norms

- Throughout the subject we will often encounter **norms** that are functions $\mathbb{R}^n \rightarrow \mathbb{R}$ of a particular form
 - * Intuitively, norms measure lengths of vectors in some sense
 - * Often component of objectives or stopping criteria in optimisation-for-ML

- More specifically, we will often use the L_2 norm (*aka* **Euclidean distance**)

$$\|\mathbf{a}\| = \|\mathbf{a}\|_2 \equiv \sqrt{a_1^2 + \cdots + a_n^2}$$

- And also the L_1 norm (*aka* absolute norm or **Manhattan distance**)

$$\|\mathbf{a}\|_1 \equiv |a_1| + \cdots + |a_n|$$

Vector Spaces and Bases

Useful in interpreting matrices and some algorithms like PCA

Linear combinations, Independence

- A **linear combination** of vectors $v_1, \dots, v_k \in V$ some vector space, is a new vector $\sum_{i=1}^k a_i v_i$ for some scalars a_1, \dots, a_k
- A set of vectors $\{v_1, \dots, v_k\} \subseteq V$ is called **linearly dependent** if one element v_j can be written as a linear combination of the other elements
- A set that isn't linearly dependent is **linearly independent**
- For formal definition of **vector spaces**:
https://en.wikipedia.org/wiki/Vector_space#Definition

Spans, Bases

- The **span** of vectors $v_1, \dots, v_k \in V$ is the set of all obtainable linear combinations (ranging over all scalar coefficients) of the vectors
- A set of vectors $B = \{v_1, \dots, v_k\} \subseteq V$ is called a **basis** for a vector subspace $V' \subseteq V$ if
 1. The set B is linearly independent; and
 2. Every $v \in V'$ is a linear combination of the set B .
- An **orthonormal basis** is a basis in which:
 1. Each pair of basis vectors are orthogonal (zero dot prod); and
 2. Each vector has norm equal to 1.

Matrices

Some useful facts for ML

Basic matrices

- A rectangular array, often denoted by upper-case, with two indices first for row, second for column
- **Square matrix** has equal dimensions (numbers of rows and columns)
- **Matrix transpose** \mathbf{A}' or \mathbf{A}^T of m by n matrix \mathbf{A} is an n by m matrix with entries $A'_{ij}=A_{ji}$
- A square matrix \mathbf{A} with $\mathbf{A}=\mathbf{A}'$ is called **symmetric**
- The (square) **identity matrix** \mathbf{I} has 1 on the diagonal, 0 off-diagonal
- **Matrix inverse** \mathbf{A}^{-1} of square matrix \mathbf{A} (if it exists) satisfies $\mathbf{A}^{-1}\mathbf{A}=\mathbf{I}$
- See more: [https://en.wikipedia.org/wiki/Matrix_\(mathematics\)](https://en.wikipedia.org/wiki/Matrix_(mathematics))
 - * Including matrix-matrix and matrix-vector products

Matrix eigenspectrum

- Scalar, vector pair (λ, \mathbf{v}) are called an **eigenvalue-eigenvector** pair of a **square matrix** \mathbf{A} if $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$
 - * Intuition: matrix \mathbf{A} doesn't rotate \mathbf{v} it just **stretches** it
 - * Intuition: the eigenvalue represents stretching factor
- In general eigenvalues may be zero, negative or even complex (imaginary) – we'll only encounter reals

Spectra of common matrices

- Eigenvalues of **symmetric matrices** are always real (no imaginary component)
- A matrix with **linear dependent** columns has some zero eigenvalues (called rank deficient) \rightarrow no matrix inverse exists

Positive (semi)definite matrices

- A symmetric square matrix \mathbf{A} is called positive semidefinite if for all vectors \mathbf{v} we have $\mathbf{v}'\mathbf{A}\mathbf{v} \geq 0$.
 - * Then \mathbf{A} has non-negative eigenvalues
 - * For example, any $\mathbf{A} = \mathbf{X}'\mathbf{X}$ since: $\mathbf{v}'\mathbf{X}'\mathbf{X}\mathbf{v} = \|\mathbf{X}\mathbf{v}\|^2 \geq 0$
- Further if $\mathbf{v}'\mathbf{A}\mathbf{v} > 0$ holds as a strict inequality then \mathbf{A} is called **positive definite**
 - * Then \mathbf{A} has (strictly) positive eigenvalues

Mini Summary

- Vectors: Vector spaces, dot products, independence, hyperplanes
- Matrices: Eigenvalues, positive semidefinite matrices

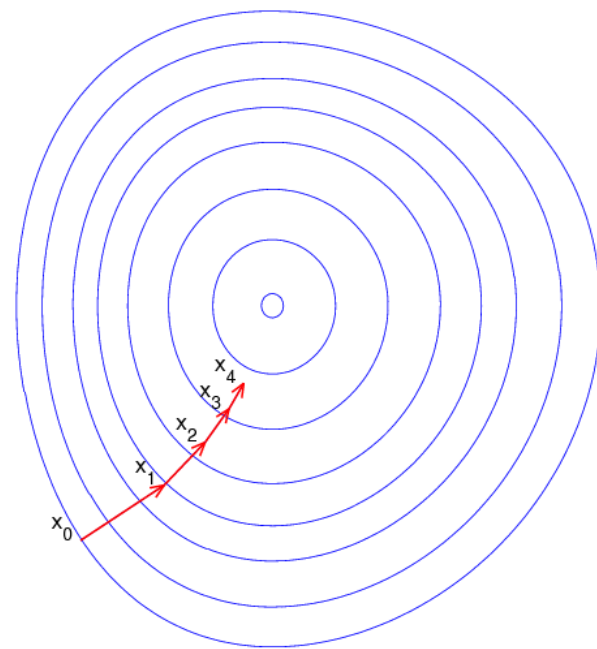
Next: Sequences and limits review/primer

Sequences and Limits

Sequences arise whenever we have iterations (e.g. training loops, growing data sample size). Limits tell us about where sequences tend towards.

Infinite Sequences

- Written like x_1, x_2, \dots or $\{x_i\}_{i \in \mathbb{N}}$
- Formally: a function from the positive (from 1) or non-negative (from 0) integers
- Index set: subscript set e.g. \mathbb{N}
- Sequences allow us to reason about test error when training data grows indefinitely, or training error (or a stopping criterion) when training runs arbitrarily long



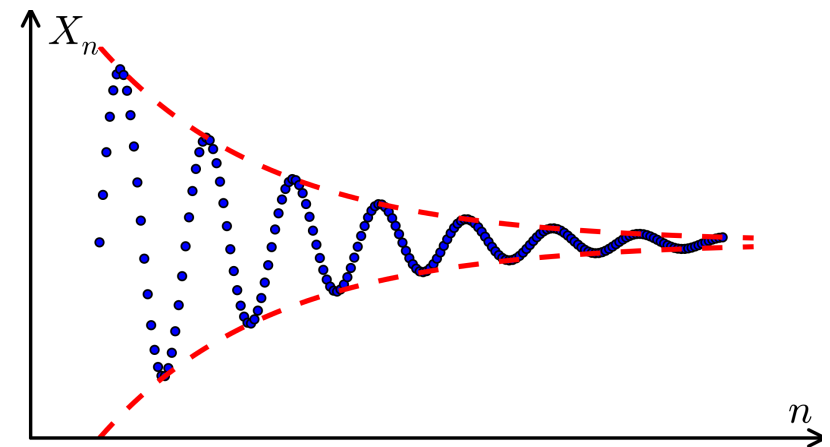
Wikipedia public domain

Limits and Convergence

- A sequence $\{x_i\}_{i \in \mathbb{N}}$ **converges** if its elements become and remain arbitrarily close to a fixed **limit** point L .
- Formally: $x_i \rightarrow L$ if, for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$ we have $\|x_n - L\| < \varepsilon$

Notes:

- Epsilon ε represents distance of sequence to limit point
- Distance can be arbitrarily small
- Definition says we eventually get that close (at some finite N) and we stay *at least* that close for ever more



Wikipedia public domain

Supremum

Generalising the maximum: When a sequence never quite peaks.

When does the Maximum Exist?

- Can you always take a **max of a set**?
- Finite sets: what's the max of $\{1, 7, 3, 2, 9\}$?
- Closed, bounded intervals: what's the max of $[0,1]$?
- Open, bounded intervals: what's the max of $[0,1)$?
- Open, unbounded intervals: what's the max of $[0,\infty)$?

What about “Least Upper Bound”?

- Can you always take a **least-upper-bound of a set**? (much more often!)
- Finite sets: what's the max of $\{1, 7, 3, 2, 9\}$?

max=9 LUB=9

- Closed, bounded intervals: what's the max of $[0,1]$?

max=1 LUB=1

- Open, bounded intervals: what's the max of $[0,1)$?

max=N/A LUB=1

- Open, unbounded intervals: what's the max of $[0,\infty)$?

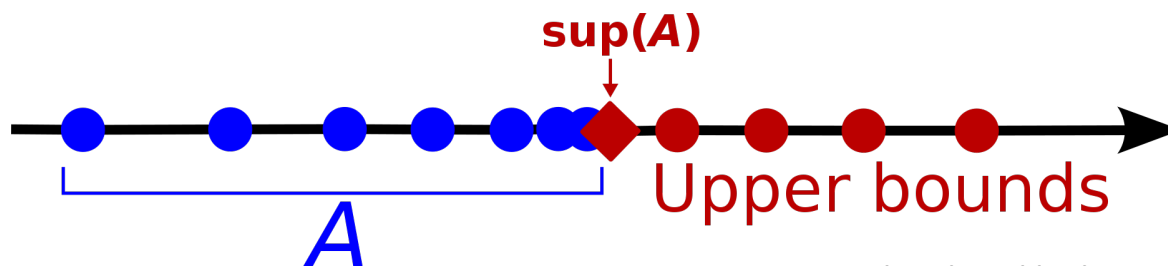
max=N/A LUB= ∞

The Supremum

- Consider any subset S of the reals
- Upper bound** $u \in \mathbb{R}^+$ of set S has: $u \geq x$ for all $x \in S$
- If u is no bigger than any other upper bound of S then it's called a least upper bound or **supremum** of S , written as $\sup(S)$ and pronounced “soup”:
 - * $z \geq u$ for all upper bounds $z \in \mathbb{R}^+$ of S
- When we don't know, or can't guarantee, that a set or sequence has a max, it is better to use its sup



FreeSVG public domain



Wikipedia public domain

Infimum

- The greatest lower bound or **infimum** is generalisation of the minimum
- Written $\inf(S)$ pronounced “inf”
- Useful if we’re minimising training error but don’t know if the minimum is ever attained.

Mini Summary

- Sequences
- Limits of sequences
- Supremum is the new maximum

Next time: L02 Statistical schools

Jupyter notebooks setup and launch (at home)

Homework 0 (no credit) will be released soon

Lecture 2. Statistical Schools of Thought

COMP90051 Statistical Machine Learning

Lecturer: Jean Honorio



THE UNIVERSITY OF
MELBOURNE

This lecture

How do learning algorithms come about?

- Frequentist statistics
- Statistical decision theory
- Bayesian statistics

Types of probabilistic models

- Parametric vs. Non-parametric
- Generative vs. Discriminative

Frequentist Statistics

Wherein unknown model parameters are treated as having fixed but unknown values.

Frequentist statistics

- Abstract problem

- * Given: X_1, X_2, \dots, X_n drawn i.i.d. from some distribution
- * Want to: identify unknown distribution, or a property of it

Independent and
identically distributed

- Parametric approach (“**parameter estimation**”)

- * Class of **models** $\{p_\theta(x) : \theta \in \Theta\}$ indexed by **parameters** Θ (could be a real number, or vector, or)
- * **Point estimate** $\hat{\theta}(X_1, \dots, X_n)$ a function (or **statistic**) of data

Hat means estimate
or estimator

- Examples

- * Given n coin flips, determine probability of landing heads
- * Learning a classifier

Estimator Bias

Frequentists seek good behaviour, in ideal conditions

- **Bias:** $B_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta$

Example: for $i=1\dots 40$

- $X_{i,1}, \dots, X_{i,20} \sim p_{\theta} = \text{Normal}(\theta = 0, \sigma^2 = 1)$

Subscript θ means data
really comes from p_{θ}

Estimator Bias

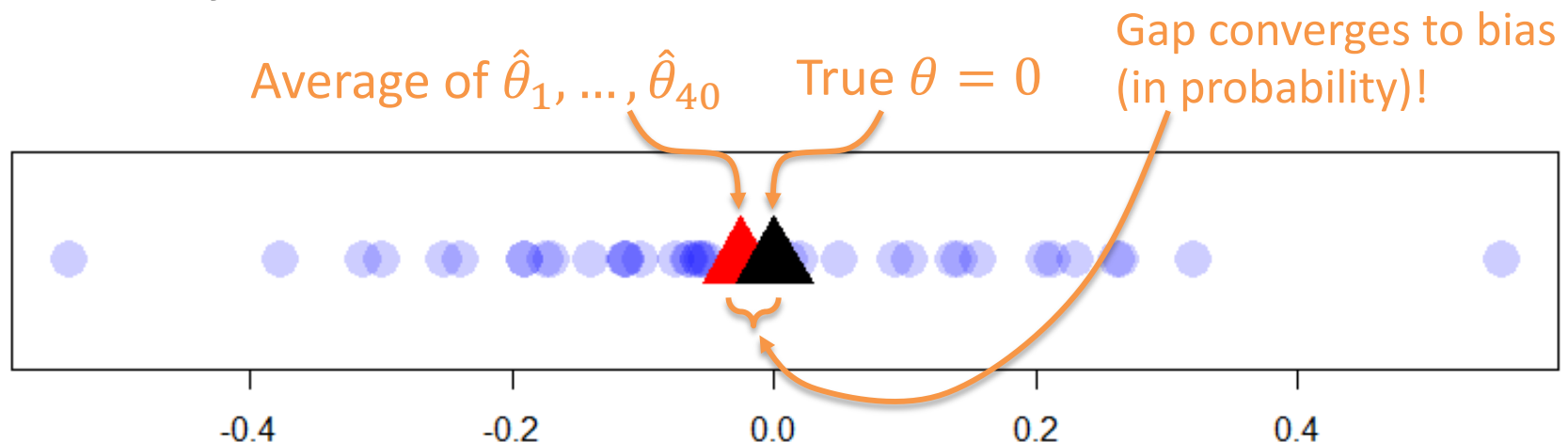
Frequentists seek good behaviour, in ideal conditions

- **Bias:** $B_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta$

Subscript θ means data really comes from p_{θ}

Example: for $i=1\dots 40$

- $X_{i,1}, \dots, X_{i,20} \sim p_{\theta} = \text{Normal}(\theta = 0, \sigma^2 = 1)$
- $\hat{\theta}_i = \frac{1}{20} \sum_{j=1}^{20} X_{i,j}$ the sample mean, plot as ●



Estimator Variance

Frequentists seek good behaviour, in ideal conditions

- **Variance:** $\text{Var}_\theta(\hat{\theta}) = \text{E}_\theta[(\hat{\theta} - \text{E}_\theta[\hat{\theta}])^2]$

$\hat{\theta}$ still function of data

Estimator Variance

Frequentists seek good behaviour, in ideal conditions

- **Variance:** $\text{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2]$

$\hat{\theta}$ still function of data

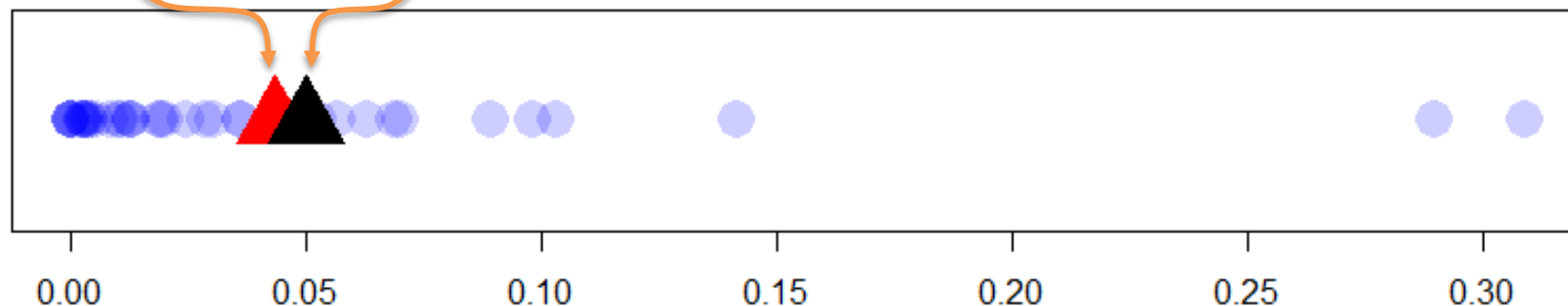
Example cont.

- Plot each $(\hat{\theta}_i - \mathbb{E}_\theta[\hat{\theta}_i])^2 = \hat{\theta}_i^2$ as ●

Average of $\hat{\theta}_1^2, \dots, \hat{\theta}_{40}^2$

True $\text{Var}_\theta(\hat{\theta}) = \frac{\sigma^2}{20} = 0.05$

Once again, average converges to true (in probability)!



Asymptotically Well Behaved

For our example estimator (sample mean), we could calculate its exact bias (zero) and variance (σ^2). Usually can't guarantee low bias/variance exactly 😞

Asymptotic properties often hold! 😊

- **Consistency**: $\hat{\theta}(X_1, \dots, X_n) \rightarrow \theta$ in probability
- **Asymptotic efficiency**: $\text{Var}_{\theta} \left(\hat{\theta}(X_1, \dots, X_n) \right)$ converges to the smallest possible variance of any estimator of θ

Bias closer and
closer to zero

Variance closer &
closer to optimal

Maximum-Likelihood Estimation

- A **general principle** for designing estimators
- Involves **optimisation**
- $\hat{\theta}(x_1, \dots, x_n) \in \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n p_{\theta}(x_i)$
- *“The best estimate is one under which observed data is most likely”*



Fischer

Example I: Bernoulli

- Know data comes from Bernoulli distribution with unknown parameter (e.g., biased coin); find mean
- MLE for mean

$$* p_{\theta}(x) = \begin{cases} \theta, & \text{if } x = 1 \\ 1 - \theta, & \text{if } x = 0 \end{cases} = \theta^x (1 - \theta)^{1-x}$$

(note: $p_{\theta}(x) = 0$ for all other x)

$$* \text{ Maximising likelihood yields } \hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$$

Example I: Bernoulli

1. Write as joint distribution over all X

$$\prod_{i=1}^n p_{\theta}(x_i) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}$$

2. Take logarithms

$$\begin{aligned} \log \prod_{i=1}^n p_{\theta}(x_i) &= \sum_{i=1}^n x_i \log \theta + (1 - x_i) \log(1 - \theta) \\ &= L(\theta) \end{aligned}$$

3. Find derivative. Set $\sum_{i=1}^n x_i = \bar{X}$

$$\frac{d}{d\theta} L(\theta) = \frac{\bar{X}}{\theta} - \frac{n - \bar{X}}{1 - \theta}$$

4. Set equal to zero and solve for $\hat{\theta}$:

$$\hat{\theta} = \frac{\bar{X}}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Example II: Normal

- Know data comes from Normal distribution with variance 1 but unknown mean; find mean

- MLE for mean

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

- * Maximising likelihood yields $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$

- Exercise: derive MLE for *variance* σ^2 based on

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

MLE 'algorithm'

1. Given data X_1, \dots, X_n **define** probability distribution, p_θ , assumed to have **generated the data**
2. Express likelihood of data, $\prod_{i=1}^n p_\theta(X_i)$
(usually its **logarithm... why?**)
3. Optimise to find *best* (most likely) parameters $\hat{\theta}$
 1. take partial derivatives of log likelihood wrt θ
 2. set to 0 and solve
(failing that, use **gradient descent**)

Mini Summary

- Frequentist school of thought
- Point estimates
- Quality: bias, variance, consistency, asymptotic efficiency
- Maximum-likelihood estimation (MLE)

Next: Statistical Decision Theory

Statistical Decision Theory

Branch within statistics, optimisation, economics, control, emphasising utility maximisation.

Decision theory



Wald

- Act to maximise utility - connected to economics and operations research
- **Decision rule** $\delta(\mathbf{x}) \in A$ an action space
 - * E.g. Point estimate $\hat{\theta}(x_1, \dots, x_n)$
 - * E.g. Out-of-sample prediction $\hat{Y}_{n+1} | X_1, Y_1, \dots, X_n, Y_n, X_{n+1}$
- **Loss function** l : economic cost, error metric
 - * E.g. square loss of estimate $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
 - * E.g. 0-1 loss of classifier predictions $l(y, \hat{y}) = 1[y \neq \hat{y}]$

Risk & Empirical Risk Minimisation (ERM)

- In decision theory, really care about *expected* loss

- **Risk** $R_\theta[\delta] = \mathbb{E}_{X \sim \theta}[l(\delta(X), \theta)]$ or

$$R_\theta[\delta] = \mathbb{E}_{(X,Y) \sim \theta}[l(\delta(X), Y)]$$

* E.g. true test error, aka generalization error

- Want: Choose δ to minimise $R_\theta[\delta]$

- Can't directly! Why?

- **ERM**: Use training set \mathbf{X} to approximate R_θ

* Minimise **empirical risk** $\hat{R}_\theta[\delta] = \frac{1}{n} \sum_{i=1}^n l(\delta(X_i), \theta)$ or

$$\hat{R}_\theta[\delta] = \frac{1}{n} \sum_{i=1}^n l(\delta(X_i), Y_i)$$

Decision theory vs. Bias-variance

We've already seen

- Bias: $B_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta$
- Variance: $\text{Var}_{\theta}(\hat{\theta}) = E_{\theta}[(\hat{\theta} - E_{\theta}[\hat{\theta}])^2]$

But are they equally important? How related?

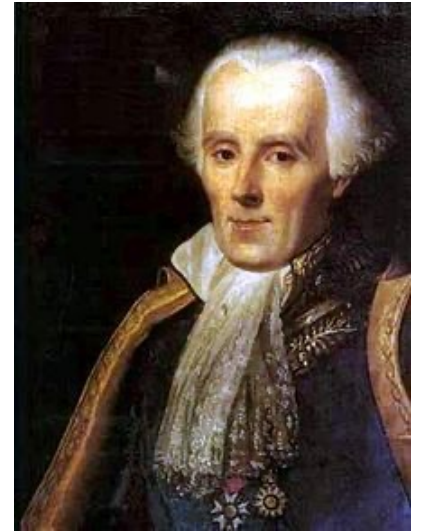
- **Bias-variance decomposition** of square-loss risk

$$E_{\theta}[(\theta - \hat{\theta})^2] = [B(\hat{\theta})]^2 + \text{Var}_{\theta}(\hat{\theta})$$

Bayesian Statistics

Wherein unknown model parameters have associated distributions reflecting prior belief.

Bayesian statistics



Laplace

- Probabilities correspond to **beliefs**
- Parameters
 - * Modeled as r.v.'s having distributions
 - * Prior belief in θ encoded by **prior distribution** $P(\theta)$
 - Parameters are modeled like r.v.'s (even if not really random)
 - Thus: data likelihood $P_{\theta}(X)$ written as conditional $P(X|\theta)$
 - * Rather than point estimate $\hat{\theta}$, Bayesians update belief $P(\theta)$ with observed data to $P(\theta|X)$ the **posterior distribution**

Tools of probabilistic inference

- Bayesian probabilistic inference

- * Start with prior $P(\theta)$ and likelihood $P(X|\theta)$
- * Observe data $X = x$
- * Update prior to posterior $P(\theta|X = x)$



Bayes

- Primary tools to obtain the posterior

- * **Bayes Rule**: reverses order of conditioning

$$P(\theta|X = x) = \frac{P(X = x|\theta)P(\theta)}{P(X = x)}$$

- * **Marginalisation**: eliminates unwanted variables

$$P(X = x) = \sum_t P(X = x, \theta = t)$$

This quantity
is called the
evidence

These are
general tools of
probability and
not specific to
Bayesian
stats/ML

Example

- Given:** The data comes from a **Normal distribution with variance 1** but **unknown mean θ** .
- Goal:** Find the **posterior over the mean** after seeing one data point where **$X=1$** . Assume a **Normal prior over θ with mean 0 and variance 1**.

$$P(\theta|X=1) = \frac{P(X=1|\theta)P(\theta)}{P(X=1)}$$

$$\propto P(X=1|\theta)P(\theta)$$

$$= \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1-\theta)^2}{2}\right) \right] \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right) \right]$$

$$\propto \exp\left(-\frac{(1-\theta)^2 + \theta^2}{2}\right) = \exp\left(-\frac{2\theta^2 - 2\theta + 1}{2}\right)$$

$$= \exp\left(-\frac{\theta^2 - \theta + \frac{1}{2}}{2 \times \frac{1}{2}}\right)$$

The Normal distribution:

$$\mathcal{N}(x; \mu, \sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Discard constants
wrt θ

Make leading numerator zero:
 $\times \frac{1}{2}$ on top and bottom

Example

$$P(\theta|X=1) = \exp\left(-\frac{\theta^2 - \theta + \frac{1}{2}}{2 \times \frac{1}{2}}\right)$$

$$= \exp\left(-\frac{(\theta - \frac{1}{2})^2 + \frac{1}{4}}{2 \times \frac{1}{2}}\right)$$

$$= \exp\left(-\frac{(\theta - \frac{1}{2})^2}{2 \times \frac{1}{2}}\right) \cdot \exp\left(-\frac{\frac{1}{4}}{2 \times \frac{1}{2}}\right)$$

$$\propto \exp\left(-\frac{(\theta - \frac{1}{2})^2}{2 \times \frac{1}{2}}\right)$$

$$\propto \mathcal{N}(0.5, 0.5)$$

The square completion trick!

https://en.wikipedia.org/wiki/Completing_the_square

$$\begin{aligned}\theta^2 - \theta + \frac{1}{2} &\rightarrow (\theta - ?)^2 \\ \theta^2 - \theta + \frac{1}{2} &= (\theta - \frac{1}{2})^2 + \frac{1}{4}\end{aligned}$$

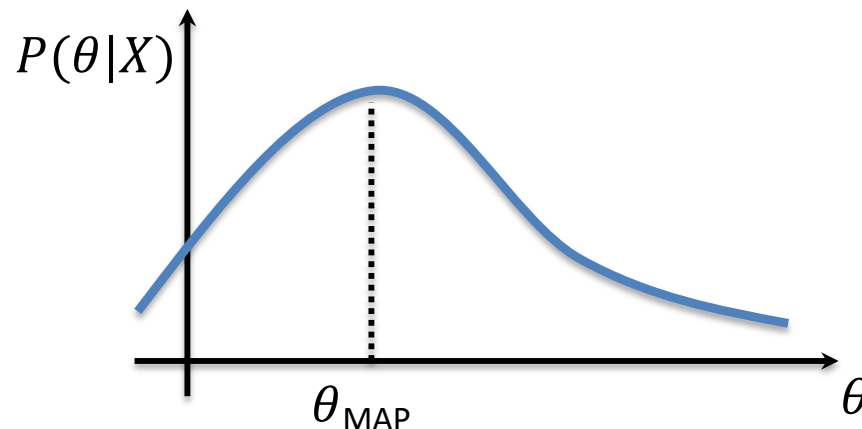
The Normal distribution:

$$\mathcal{N}(x; \mu, \sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

更新后验的过程 从 (0,1) \rightarrow 观察到一点数据 (0.5, 0.5)

How Bayesians make point estimates

- They don't, unless forced at gunpoint!
 - * The posterior carries full information, why discard it?
- But, there are common approaches
 - * Posterior mean $E_{\theta|X}[\theta] = \int \theta P(\theta|X) d\theta$
 - * Posterior mode $\underset{\theta}{\operatorname{argmax}} P(\theta|X)$ (^{MAP}max a posteriori or MAP)
 - * There're Bayesian decision-theoretic interpretations of these



MLE in Bayesian context

- MLE formulation: find parameters that best fit data

$$\hat{\theta} \in \operatorname{argmax}_{\theta} P(X = x | \theta)$$

- Consider the **MAP** under a Bayesian formulation

$$\hat{\theta} \in \operatorname{argmax}_{\theta} P(\theta | X = x)$$

$$= \operatorname{argmax}_{\theta} \frac{P(X = x | \theta) P(\theta)}{P(X = x)}$$

$$= \operatorname{argmax}_{\theta} P(X = x | \theta) P(\theta)$$

与MLE相比
先验权重
MAP 多了 $P(\theta)$

- Prior** $P(\theta)$ weights; **MLE like uniform** $P(\theta) \propto 1$

Frequentists vs Bayesians – Oh My!

- Two key schools of statistical thinking
 - * Decision theory complements both
- Past: controversy; animosity; almost a 'religious' choice
- Nowadays: deeply connected

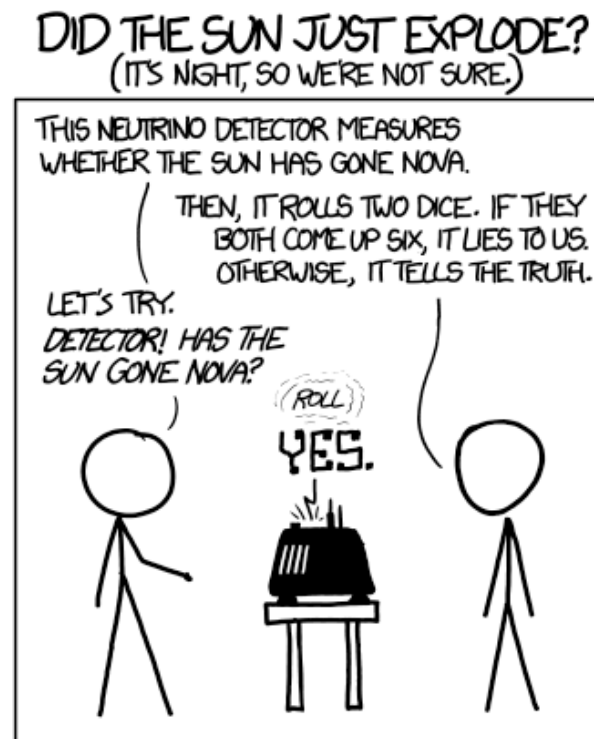
I declare the Bayesian vs. Frequentist debate over for data scientists

👤 Rafael Irizarry 📅 2014/10/13

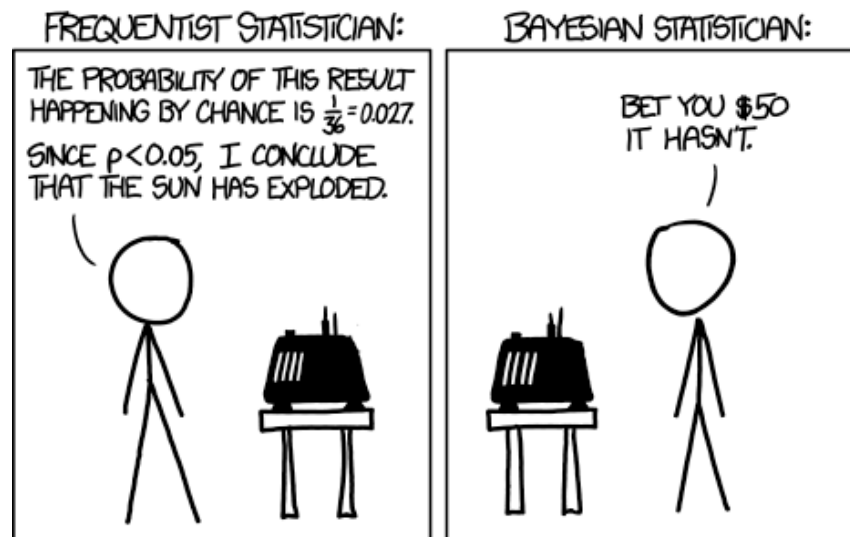
Are You a Bayesian or a Frequentist?

Michael I. Jordan
 Department of EECS
 Department of Statistics
 University of California, Berkeley

<http://www.cs.berkeley.edu/~jordan>



<https://xkcd.com/1132/> CC-NC2.5



(Some) Categories of Probabilistic Models

Parametric vs non-parametric models

参数模型: 随数据量增加而增加¹⁷

Parametric	Non-Parametric
Determined by <u>fixed, finite</u> number of parameters ^{linear regression}	^{tree KNN SVM} Number of parameters grows with data, potentially infinite
Limited flexibility	More flexible
Efficient statistically and computationally	Less efficient

Examples to come! There are non/parametric models in both the frequentist and Bayesian schools.

Generative vs. discriminative models

- X 's are instances, Y 's are labels (supervised setting!)
 - * Given: i.i.d. data $(X_1, Y_1), \dots, (X_n, Y_n)$
 - * Find model that can predict Y of new X

- Generative approach ^{Bayes, Markov, Gaussian}
 - * Model full joint $P(X, Y)$ ^{找到 X 与 Y 的联合分布} ^{理解完整的数据生成过程}
- Discriminative approach ^{logistic, linear}
 - * Model conditional $P(Y|X)$ only ^{只想告诉你我的模型对于 X 最大 predict Y from X}
- Both have pro's and con's ^{判别方法简单, 计算效率高 (但不提供完整数据生成模型) 无法理解数据生成过程}

Examples to come! There are generative/discriminative models in both the frequentist and Bayesian schools.

→ 生成模型: 能够提供新样本, 对底层数据分布提供更深理解 (更复杂, 计算量大) 判别模型

Mini Summary

- Bayesian paradigm: It's all in the prior!
- Bayesian point estimate: MAP
- Parametric vs Non-parametric models
- Discriminative vs. Generative models

Next: Linear regression (unlike you've ever seen before)

Submit Homework 0 (no credit)