Lecture 1. Introduction. Probability Theory

COMP90051 Statistical Machine Learning

Sem1 2020
Lecturer: Trevor Cohn

THE UNIVERSITY OF MELBOURNE

This lecture

- Machine learning: why and what?
- About COMP90051
- Review: ML basics, Probability theory

Why Learn Learning?

Motivation

- *"We are drowning in information,*
  but we are starved for knowledge"
  - John Naisbitt, *Megatrends*

- Data = raw information
- Knowledge = patterns or models behind the data

Solution: Machine learning

- Hypothesis: pre-existing data repositories contain a lot of potentially valuable knowledge
- Mission of learning: find it
- Definition of learning:

  (semi-)automatic extraction of **valid**, **novel**, **useful** and **comprehensible** knowledge – in the form of rules, regularities, patterns, constraints or models – from arbitrary sets of data

Applications of ML are deep and prevalent

- Online ad selection and placement
- Risk management in finance, insurance, security
- High-frequency trading
- Medical diagnosis
- Mining and natural resources
- Malware analysis
- Drug discovery
- Search engines

…

Draws on many disciplines

- Artificial Intelligence
- Statistics
- Continuous optimisation
- Databases
- Information Retrieval
- Communications/information theory
- Signal Processing
- Computer Science Theory
- Philosophy
- Psychology and neurobiology

…

Job$

Many companies across all industries hire ML experts:

Data Scientist
Analytics Expert
Business Analyst
Statistician
Software Engineer
Researcher
…

About this Subject

(refer also to LMS)

Vital statistics

Lecturer & Coordinator	Trevor Cohn (DMD3, trevor.cohn@unimelb.edu.au) Prof, Computing & Information Systems *Statistical Machine Learning, Natural Language Processing*
Co-lecturer:	Parvin Eskikand (DMD3, pzarei@unimelb.edu.au) Cognitive Computing for Medical Technologies
Tutors:	Justin Tan (Head Tutor; justan@student.unimelb.edu.au) Kazi Abir Adnan, Xudong Han, Jun Wang *Contact info: LMS → Modules → Welcome*
Contact:	*Weekly you should attend: 2x Lectures & 1x Workshop*
Office Hours	*TBD; will run on demand*

First port of call: LMS Discussion Board
Our aim half business day latency!

About me (Trevor)

- PhD 2006 – Melbourne
- Several years in research
  - UK: Edinburgh U, Sheffield U.
  - Australia: Melbourne U.
- Interests: Structured prediction; graphical models; probabilistic modelling (Bayesian); deep learning; transfer learning
- Applications to language: e.g., structure parsing / induction, translation, sequential tagging

Subject content

- The subject will cover topics from

  Foundations of statistical learning, linear models, non-linear bases, kernel approaches, neural networks, Bayesian learning, probabilistic graphical models (Bayes Nets, Markov Random Fields), cluster analysis, dimensionality reduction, regularisation and model selection

- Theory in lectures; hands-on experience with range of toolkits in workshop pracs and projects

- Vs COMP90049: much depth, much rigor, so wow

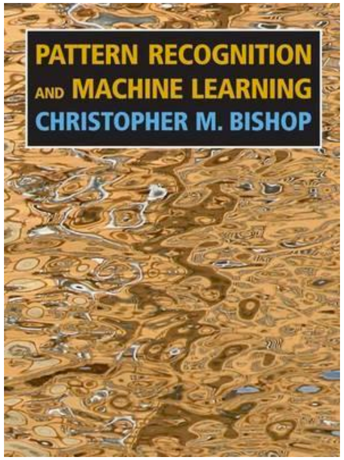
Advanced ML: Expected Background

- Why a challenge: Diverse math methods + CS + coding
- ML: COMP90049; refresher deck on LMS → *Modules* → *Resources*
- Alg & complexity: big-oh, termination; basic data structures & algorithms; solid coding ideally experience in Python
- Maths: Refreshers but really need **solid** understanding in advance
  "Matrix **A** *is symmetric & positive definite, hence its eigenvalues…"*
- Probability theory: probability calculus; discrete/continuous distributions; multivariate; exponential families; Bayes rule
- Linear algebra: vector inner products & norms; orthonormal bases; matrix operations, inverses, eigenvectors/values
- Calculus & optimisation: partial derivatives; gradient descent; convexity; Lagrange multipliers

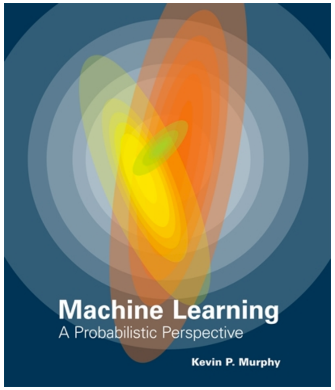
Subject objectives

- Develop an appreciation for the role of statistical machine learning, both in terms of foundations and applications
- Gain an understanding of a representative selection of ML techniques
- Be able to design, implement and evaluate ML systems
- Become a discerning ML consumer

Textbooks

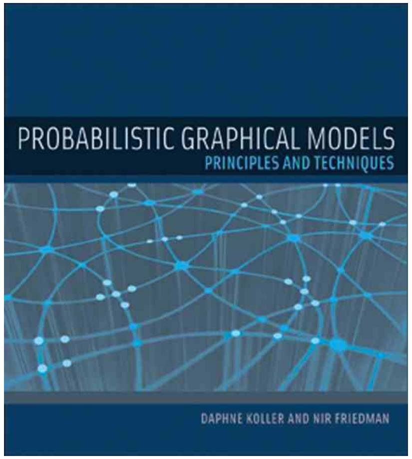


- Primarily references to
  - Bishop (2007) *Pattern Recognition and Machine Learning*
- Other good general references:
  - Murphy (2012) *Machine Learning: A Probabilistic Perspective* [read free ebook using 'ebrary' at http://bit.ly/29SHAQS]
  - Hastie, Tibshirani, Friedman (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction* [free at http://www-stat.stanford.edu/~tibs/ElemStatLearn]



Textbooks

- Also relevant for PGM component
  - Koller, Friedman (2009) *Probabilistic Graphical Models: Principles and Techniques*



Assessment

- Assessment components
  - Two projects – one released early (w4-7), one late (w8-11); will have ~3 weeks to complete
    - Each (25%)
    - Latter will be a group project
  - Final Exam (50%)

- 50% Hurdle applies to both **exam** and **ongoing assessment**

Machine Learning Basics

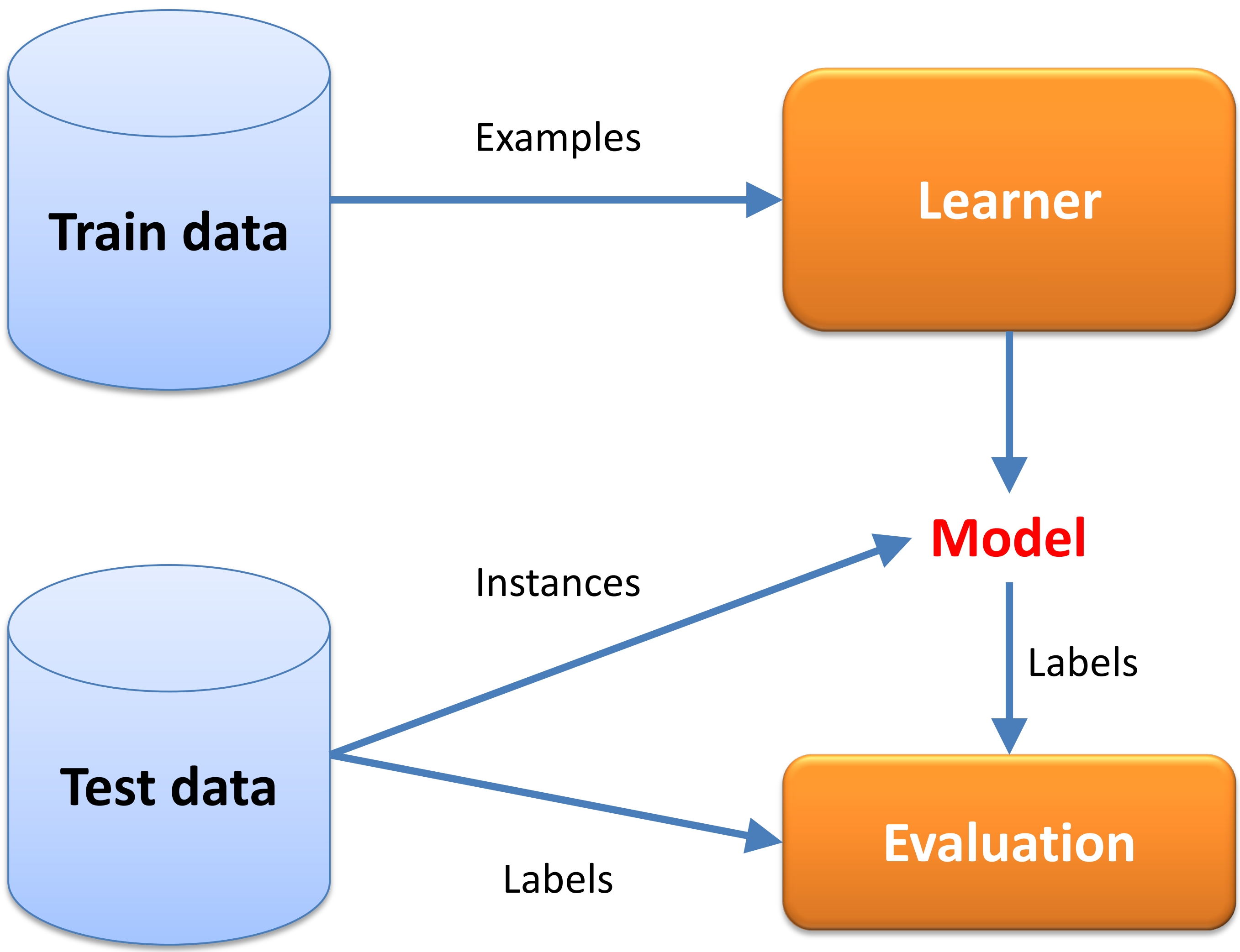
Terminology

- Input to a machine learning system can consist of
  - Instance: measurements about individual entities/objects
    a loan application
  - Attribute (aka Feature, explanatory var.): component of the instances
    the applicant’s salary, number of dependents, etc.
  - Label (aka Response, dependent var.): an outcome that is categorical, numeric, etc.
    forfeit vs. paid off
  - Examples: instance coupled with label
    <(100k, 3), “forfeit”>
  - Models: discovered relationship between attributes and/or label

Supervised vs unsupervised learning

	Data	Model used for
Supervised learning	Labelled	Predict labels on new instances
Unsupervised learning	Unlabelled	Cluster related instances; Project to fewer dimensions; Understand attribute relationships

Architecture of a supervised learner



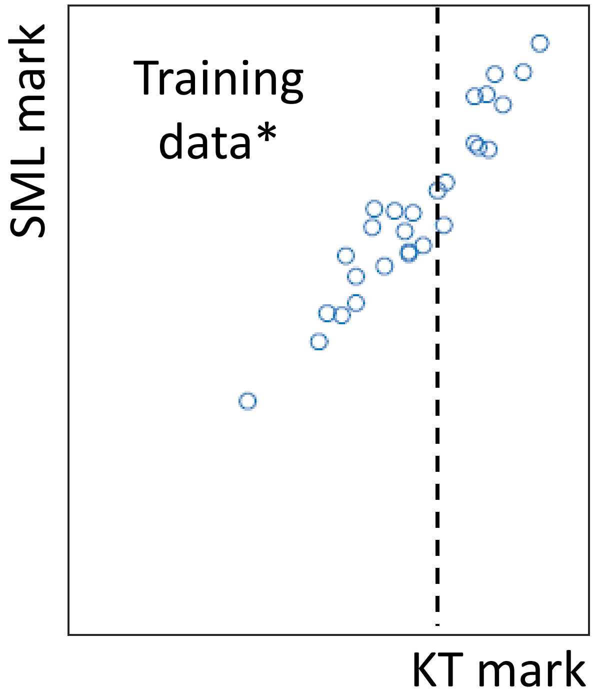
Evaluation (supervised learners)

- How you measure quality depends on your problem!
- Typical process
  - Pick an evaluation metric comparing label vs prediction
  - Procure an independent, labelled test set
  - "Average" the evaluation metric over the test set
- Example evaluation metrics
  - Accuracy, Contingency table, Precision-Recall, ROC curves
- When data poor, cross-validate

Probability Theory

(This should be a) brief refresher

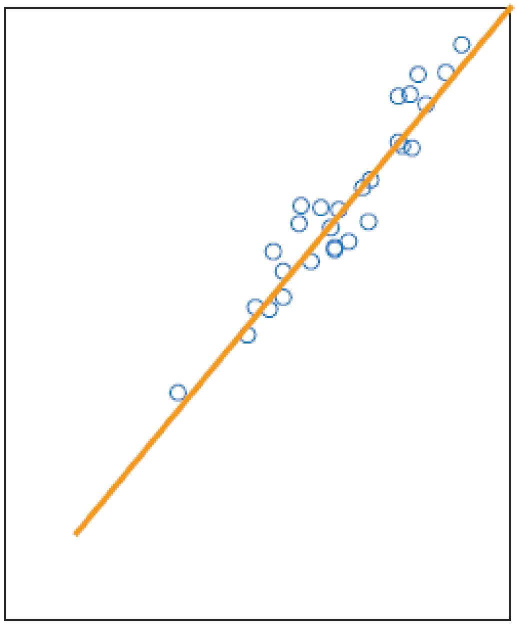
Data is noisy (almost always)



- Example:
  - given mark for Knowledge Technologies (KT)
  - predict mark for Stat Machine Learning (SML)

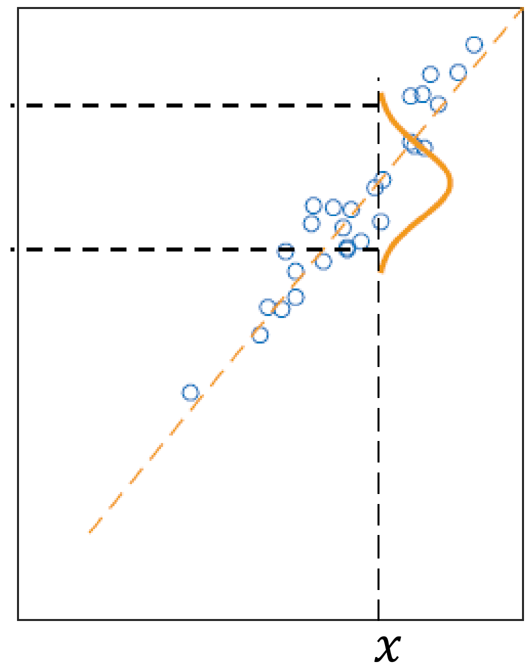
* synthetic data :)

Types of models



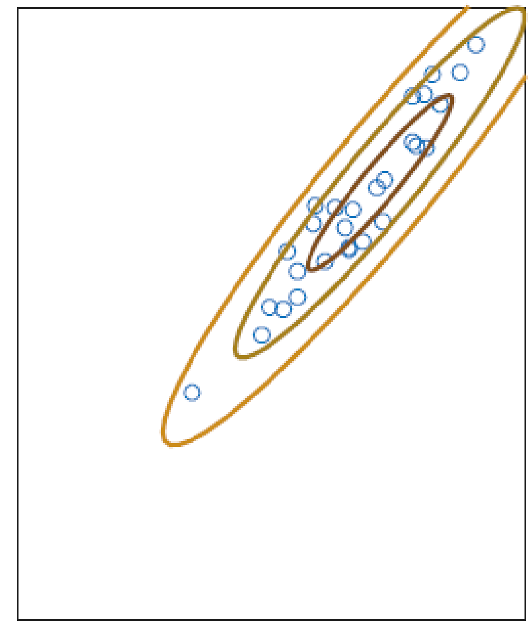
$\hat{y} = f(x)$

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



$P(y|x)$

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



$P(x, y)$

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

Basics of probability theory

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

Axioms of probability

1. $P(f) \geq 0$ for every event f in F

2. $P\left(\bigcup_f f\right) = \sum_f P(f)$ for all collections* of pairwise disjoint events

3. $P(\Omega) = 1$

* We won't delve further into advanced probability theory, which starts with measure theory. But to be precise, additivity is over collections of countably-many events.

Random variables (**r.v.'s**)

- A random variable X is a numeric function of outcome $X(\omega) \in \boldsymbol{R}$
- $P(X \in A)$ 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) = ½

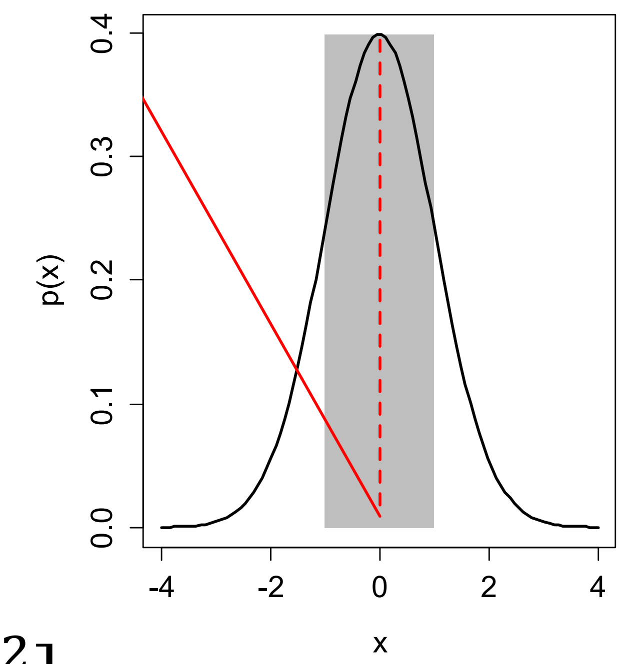
Discrete vs. continuous distributions

- 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: $Var(X) = E[(X - E[X])^2]$

Read more at https://en.wikipedia.org/wiki/Expected_value

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
  - X,Y independent equiv to $P(Y = y|X = x) = P(Y = y)$

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
- Bayesian statistical inference makes heavy use
  - Marginals: probabilities of individual variables
  - Marginalisation: summing away all but r.v.'s of interest
    $P(A) = \sum_b P(A, B = b)$

Summary

- Why study machine learning?
- COMP90051
- Machine learning basics
- Review of probability theory

Homework week #1: COMP90049 & linear algebra decks
Jupyter notebooks setup and launch (at home or in labs)

Next time: Statistical schools of thought - how many ML algorithms come to be