Lecture 14: Classification

Sarah Erfani and Karin Verspoor, CIS

Semest

THE UNIVERSITY OF MELBOURNE

What is Classification?

Classification involves predicting a discrete class or classes. Those classes are defined in advance.

Binary (yes/no)

- Deciding whether a lone application is risky or not
-
-
- Will a student skip class on Friday?

Multi-class

- Categorise a document into newspa
  entertainment, health)
- Recognise images of digits (0-9)
- Discriminating between different species of e.g. a kind of plant or an insect.
- Predicting type of cancer from gene expression data.

What are (Supervised) Classifiers?

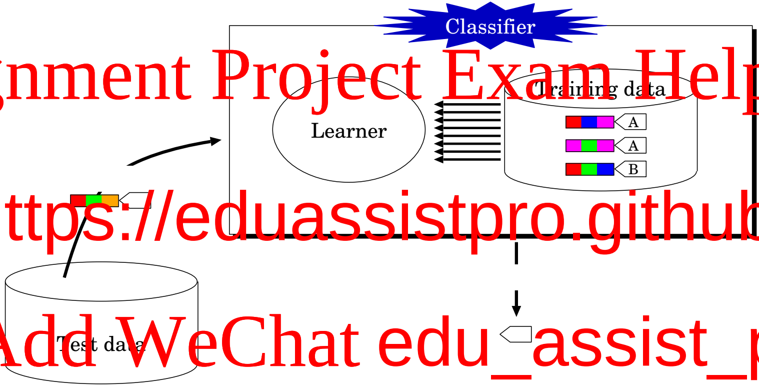
- Given:
  1. a fixed representation language of *attributes*

the category of a novel inpu

- Model:

  discover the function that p ven a previously unseen x

Supervised classification paradigm



The goal of learning from examples is not to **memorise** but rather to **generalise**, e.g., predict.

Example: Supervised Learning (Regression)

Can we predict housing prices?

A friend has a house which is 750 square feet – h
expect to get?

(draw a straight line vs. fit a curve)

Linear regression, mathematically

Linear regression captures a relationship between two variables or attributes.

It makes the assumption that there is a *linear* relationship between the two variables.

- le, or
-

At its most basic, the relationship can be expr deterministic function:

$$y = f(x)$$
$$y = \beta_0 + \beta_1 * x$$
$$y = \beta \cdot x \text{ (given } x_0 = 1\text{)}$$

A simple assumption!

Linear functions are more basic than non-linear functions (mathematically).

They capture that changes in one variable correlate linearly with

[The more umbrellas you sell, the more mon
money you make is directly proportional to h
sell.]

Applicability: Regression can be applied when all variables/attributes are real numbers.

Explore the relationship

From Schutt & O'Neil, *Doing Data Science*

Explore the relationship

From Schutt & O'Neil, *Doing Data Science*

Fitting the model

Want to choose the *best* line.

Operationally, the line that minimises the *distance* between all points and the line.

$\hat{y}_i$s

and observed y_is. Put another way, we w ... duces $\hat{y}_i$ for each x_i that is closest to the known

Minimise the Residual Sum of Squares (RS

(aka Sum of Squares Due to Error (SSE)):

$$RSS(\beta) = \sum_i (y_i - \beta x_i)^2$$

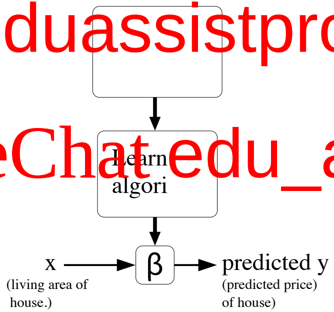
THE UNIVERSITY OF MELBOURNE

Prediction

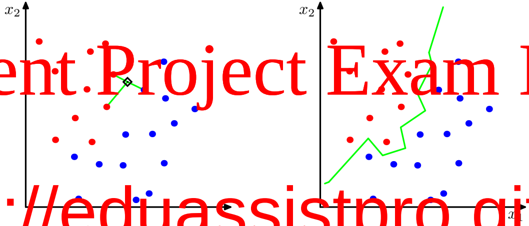
Armed with a linear model $y = \beta_0 + \beta_1 * x$, we can straightforwardly predict a continuous valued output for y given a value of x.

We derive that linear model by estimating it from training examples.

Given examples $(x_0, y_0), (x_1, y_1), ...(x_n, y_n)$, we determine β through



$k-$Nearest Neighbour methods in Classification



Given class assignments for existing data (black).

(a) According to the class membership of the K closest data points.
(b) For $k = 1$, the induced decision boundary.

See: Charles Elkan, UCSD, 2011 lecture notes (posted on LMS)

$k-$Nearest Neighbour classification strategies

[1-NN]: Classify the test input according to the class of the closest training instance.

k

ights are based on similarity of the input to each of th

[offset-weighted k-NN]: Classify the weighted accumulative class of the re weights are based on similarity of the input to e hbours, factoring in an offset to indicate the prior expectation of a test input being classified as being a member of that class.

$k-$Nearest Neighbour classification implementation

The most naive neighbour search implementation involves the brute-force computation of distances between all pairs of points in the dataset.

For N samples in D dimensions, this approach scales as $O(DN^2)$.

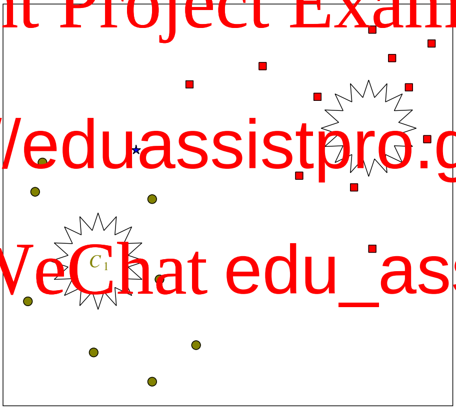
- ive

Alternative: tree-based data structures

- These structures attempt to reduce th calculations by efficiently encoding for the sample.
- The basic idea is that if point A is very distant from point B, and point B is very close to point C, then we know that points A and C are very distant, without having to explicitly calculate their distance.
- In this way, the computational cost of a nearest neighbours search can be reduced to $O(DN \log(N))$ or better.

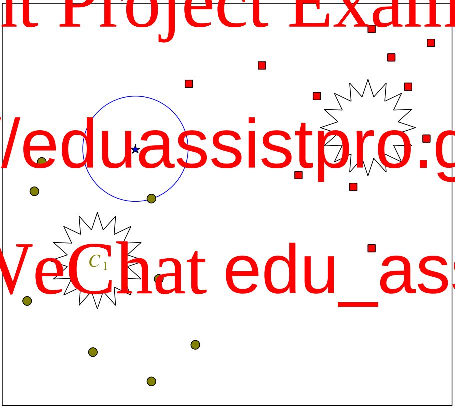
Visualisation of $k-$Nearest Neighbour classification

The nearest neighbour approach corresponds to classification by "hyper-spheres" (or "hyper-ellipsoids")



Visualisation of $k-$Nearest Neighbour classification

The nearest neighbour approach corresponds to classification by "hyper-spheres" (or "hyper-ellipsoids")



Visualisation of $k-$Nearest Neighbour classification

The nearest neighbour approach corresponds to classification by "hyper-spheres" (or "hyper-ellipsoids")

Strengths and Weaknesses of Nearest Neighbour methods

Strengths

- Simple
- Can handle arbitrarily many classes (multi-class and multi-label)

- We need some sort of averaging or voti
the labels of multiple training exampl
obvious to design.
- Expensive (in terms of index accesse
- Everything is done at run time (lazy learner)
- Prone to bias
- Arbitrary k value

Bayesian Methods

- uced
- er the

possible categories given a descripti

$$P(C|X) = \frac{P(}{}$$

Naive Bayes (NB) Classifiers

- Task: classify an instance $X = \langle x_1, x_2, ..., x_n \rangle$ according to one of the classes $c_j \in C$

$$c = \mathrm{argmax}_{c_j \in C} P(c_j | x_1, x_2, ..., x_n)$$

posterior $P(c_j | x_1, x_2, ..., x_n) =$ li

- Predicts X belongs to c_i iff the pr i t among all the $P(c_k|X)$ for all the K classes
- Since $P(x_1, x_2, ..., x_n)$ is constant for all classes, only $P(x_1, x_2, ..., x_n | c_j) P(c_j)$ needs to be maximised.

Calculating the likelihood

(given a class).

For large n,

1. Typically not enough data to estimate this accurately.
2. Common to encounter the situation where there are no training examples for a particular combination.
3. This would likely lead to over-fitting (biased to combinations for which there are examples).

Simplifying Assumptions

- $P(c_j)$
  - can be estimated from the frequency of classes in the training

- Naive Bayes Conditional Independend
  - assume that the probability of obser equal to the product of the individual p *[hence "naive"]*

The Final NB Formulation

$$= \quad \text{argmax}_{c_j \in C} P(c_j) \quad P(x_i \; c_j)$$

Naive Bayes Example

Given a training data set, what are the probabilities we need to estimate?

Headache	Sore	Temperature	Cough	Diagnosis
severe	mild	high	yes	Flu
no	severe	normal	yes	Cold
				Flu
		n	no	Cold

Ann comes to the clinic with severe headache, no soreness, normal temperature and with cough. What does sh with highest probability.

$P(Flu|Headache = severe, Sore = no,$ $yes)$
$\sim\ P(Flu) * P(Headache = severe|Flu) * P(Sore = no|Flu) * P(Temperature = normal|Flu) * P(Cough = yes|Flu)$

$P(Cold|Headache = severe, Sore = no, Temperature = normal, Cough = yes)$
$\sim\ P(Cold) * P(Headache = severe|Cold) * P(Sore = no|Cold) * P(Temperature = normal|Cold) * P(Cough = yes|Cold)$

Estimating probabilities

$$P(Flu) = 3/5$$
$$P(Headache = severe|Flu) = 2/3$$
$$P(Headache = mild|Flu) = 1/3$$
$$P(Headache = no|Flu) = 0/3\ (= e)$$
$$P(Sore = severe|Flu) = 1/3$$
$$P(Sore = mild|Flu) = 2/3$$
$$P(Sore = no|Flu) = 0/3\ (= e)$$

$$P(Cold) = 2/5$$
$$P(Headache = severe|Cold) = 0/2\ (= e)$$
$$P(Headache = mild|Cold) = 1/2$$
$$P(Headache = no|Cold) = 1/2$$
$$P(Sore = severe|Cold) = 1/2$$
$$P(Sore = mild|Cold) = 0/2\ (= e)$$
$$P(Sore = no|Cold) = 1/2$$
$$) = 0/2\ (= e)$$
$$|Cold) = 2/2$$
$$old) = 1/2$$
$$old) = 1/2$$

Set $0/y$ to e, a small value like 10^{-7} (or l – of training instances)

$P(Flu|Headache = severe, Sore = no,$ $yes)$
$\sim\ P(Flu) * P(Headache = severe|Flu)$ $rature =$
$normal|Flu) * P(Cough = yes|Flu) = 3/5 * 2/3 * e * 2/3 * 3/3 = 0.26e$

$P(Cold|Headache = severe, Sore = no, Temperature = normal, Cough = yes)$
$\sim\ P(Cold) * P(Headache = severe|Cold) * P(Sore =$
$no|Cold) * P(Temperature = normal|Cold) * P(Cough = yes|Cold)$
$= 2/5 * e * 1/2 * 1 * 1/2 = 0.1e$ *Diagnosis is Flu*

Naive Bayes, analysis

Naive Bayes (NB) Classifier is very simple to build, extremely fast to make decisions, and easy to change the probabilities when the new

- Scales easily for large number of dimensions (100s) and data sizes.
- Easy to explain the reason for the decis
- One should apply NB first before launc classification techniques.

Summary

e

required to make the computation tra

Resources

stn.pdf

Witten, Frank, Hall (2011) Data Mining. Chapter 4. (*kD tree*, ball tree)