Introduction to Machine Learning and Data Mining

Last revision: 28 Feb 201

Acknowledgements

Material derived from slides for the book
"Elements of Statistical Learning (2nd Ed.)" by T. Hastie,
R. Tibshirani & J. Friedman. Springer (2009)
http://statweb.stanford.edu/~tibs/ElemStatLearn

Material derived from slides for the book
"Machine Learning: A Probabilistic Perspective" by P. Murphy
MIT Press (2012)
http://www.c

Material derived f
"Machine Learn
Cambridge Univ
http://cs.br

Material derived from slides for the book
"Bayesian Reasoning and Machine Learning" by D. Barber
Cambridge University Press (2012)
http://www.cs.ucl.ac.uk/staff/d.barber/brml

Material derived from slides for the book
"Machine Learning" by T. Mitchell
McGraw-Hill (1997)
http://www-2.cs.cmu.edu/~tom/mlbook.html

Material derived from slides for the course
"Machine Learning" by A. Srinivasan
BITS Pilani, Goa, India (2016)

Aims

This lecture will introduce you to machine learning, giving an overview of some of the key ideas and approaches we will cover in the course. Following it you should be able to describe some of the main concepts and outline so

- cat
- wid
- batch vs. online settings
- parametric vs. non-parametric approaches
- generalisation in machine learning
- training, validation and testing phases in applications
- limits on learning

What we will cover

- cor
- fou
- rele
- practical applications

What we will NOT cover

- lots o
- lots o
- "bi
- commercial and business aspects of "analytics"
- ethical aspects of AI and ML

although all of these are interesting and important topi

Some history

One can imagine that after the machine had been in operation for some time, the instructions would have been altered out of recognition, but nevertheless still be such that one would have to ad

calc

desi

efficient manner. In such a case one would have to ad

the progress of the machine had not been foresee

original instructions were put in. It would be like a p

learnt much from his master, but had added much own work.

From A. M. Turing's lecture to the London Mathematical Society. (1947)

Some definitions

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve from experience.

"Machine L

Machine learning, then, is about making computers modify or adapt their actions (whether these actions are m predictions, or controlling a robot) so that these a more accurate, where accuracy is measured by h chosen actions reflect the correct ones.

"Machine Learning". S. Marsland (2015)

Some definitions

Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience.

Machine Le

The t

meaningful patterns in data.

"Understanding Machine Learning". S. Shalev-Shwartz an

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data.

"Data Mining". I. Witten *et al.* (2016)

Machine Learning is ...

Trying to g

R. Kohn (2015)

How is Machine Learning different from ...

Machine learning comes originally from Artificial Intelligence (AI), where the motivation is to build intelligent agents, capable of acting autonomously. Learning is a characteristic of intelligence, so to be successf *derstand* and *com*

These are not requirements in:

- statistics — the results are typically mathemat
- data mining — the results are typically models of humans

These criteria are often also necessary, but not always sufficient, for machine learning.

Machine Learning for Human-level Artifical Intelligence

ML for human-level AI, right ?

ML for human-level AI, right ? Not so fast . . .

Supervised and unsupervised learning

The most widely used categories of machine learning algorithms are:

- *Sup*
- *Uns*

There are a
strategies to acquire data, such as reinforcement lear
learning.

Note: output class can be real-valued or discrete, scalar
structure . . .

Supervised and unsupervised learning

Supervised learning tends to dominate in applications.

Why ?

Generally, because it is much easier to define the problem and develop an error measure (loss function) to evaluate different alg settings, data transformations etc. for supervised le unsupervised learning.

Supervised and unsupervised learning

Unfortunately . . .

In the real w
sufficien

So in such cases unsupervised learning is really what you

but currently, finding good unsupervised learning al
machine learning tasks remains a research challenge

Machine learning models

Machine learning models can be distinguished according to their main intuition

- *Geo*
(hy
- *Probabilistic* models view learning as a process of reducing uncertainty, modelled by means of probabilit
- *Logical* models are defined in terms of easily inte
expressions.

Machine learning models

Alternatively, can be characterised by *algorithmic properties*:

- *Reg*
- *Cla*
- *Neu*
- *Local models* predict in the local region of a que
- *Tree-based models* partition the data to ma
- *Ensembles* learn multiple models and combi

Linear regression

Given 2 real-valued variables X_1, X_2, labelled with a real-valued variable Y, find "line of best fit" that captures the dependency of Y on X_1, X_2.

Learning here is by minimizing MSE, i.e., the average of the squared vertical distances of values of Y from the learned function $\hat{Y} = \hat{f}(\mathbf{X})$.

Linear regression

A question: is it possible to do better than the line of best fit?

Maybe. Linear regression assume that the $(\mathbf{x}_i, y_i)$ examples in the data are "generat

So any trai $) = f(\mathbf{x})$

But what if f is non-linear ?

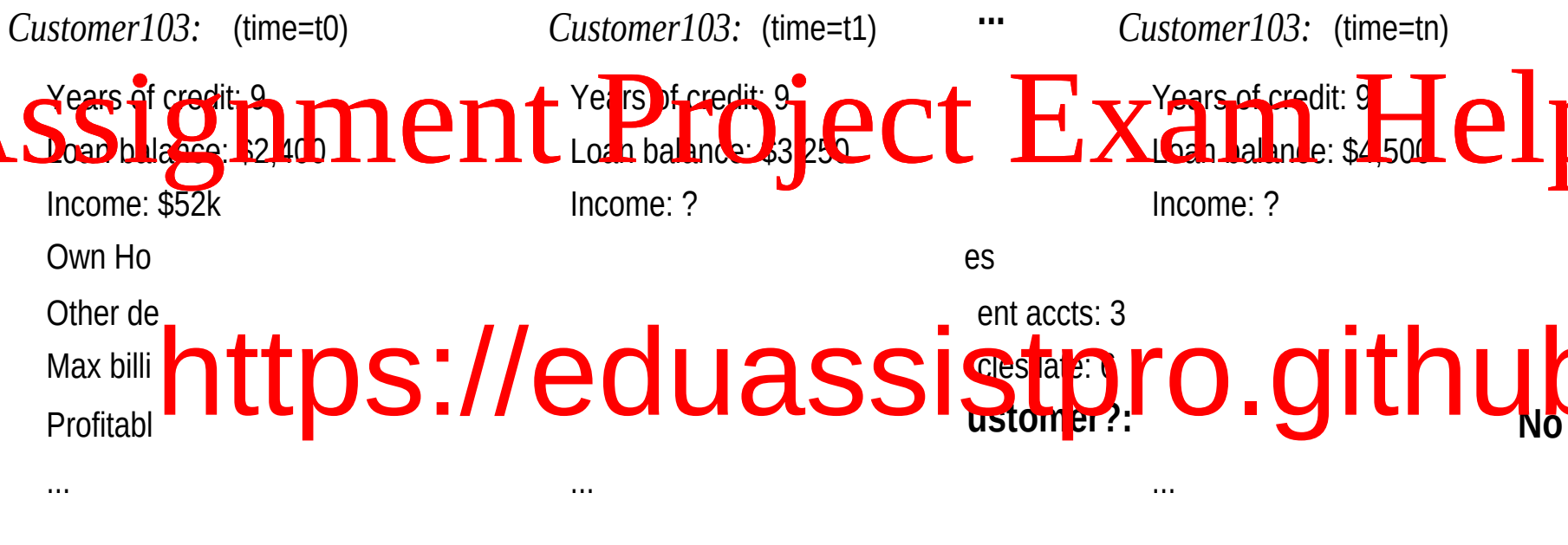
We may be able to reduce the mean squared error (MSE) v $\sum_i (y_i - \hat{y})^2$ by trying a different function.

Can "decompose" MSE to aid in selecting a better function.

Nearest Neighbour

Nearest Neighbour is a regression or classification algorithm that predicts whatever is the output value of the nearest data point to some query.

Classification



If $OtherDelinquentAccounts > 2$, and
$NumberDelinquentBillingCycles > 1$
Then $ProfitableCustomer = No$

If $OtherDelinquentAccounts = 0$, and
$Income > 30k$ OR $YearsOfCredit > 3$
Then $ProfitableCustomer = Yes$

Assassinating spam e-mail

SpamAssassin is a widely used open-source spam filter. It calculates a score for an incoming e-mail, based on a number of built-in rules or 'tests' in SpamAssassin's terminology, and adds a 'junk' flag and a summary report to the e-mail's headers if the score is 5 or more.


```
-0.1 RCVD_IN_M

 0.6 HTML_IMA
 1.2 TVD_FW_G
 0.0 HTML_MESS                BODY: HTML in
 0.6 HTML_FONx_FACE_BAD       BODY: HTML font face is not a word
 1.4 SARE_GIF_ATTACH          FULL: Email has a inline gif
 0.1 BOUNCE_MESSAGE           MTA bounce message
 0.1 ANY_BOUNCE_MESSAGE       Message is some kind of bounce message
 1.4 AWL                      AWL: From: address is on the auto white-list
```


From left to right you see the score attached to a particular test, the test identifier, and a short description including a reference to the relevant part of the e-mail. As you see, scores for individual tests can be negative (indicating evidence suggesting the e-mail is ham rather than spam) as well as positive. The overall score of 5.3 suggests the e-mail might be spam.

Linear classification

Suppose we have only two tests and four training e-mails, one of which is spam. Both tests succeed for the spam e-mail; for one ham e-mail neither test succeeds, for another the first test succeeds and the second doesn't, and for the third ham e-mail the first test fails and the second succeeds.

It is easy to see that these four e

introduced above we could describe this classifier as $4x + 4x > 5$ or $(4, 4) \cdot (x_1, x_2) > 5$.

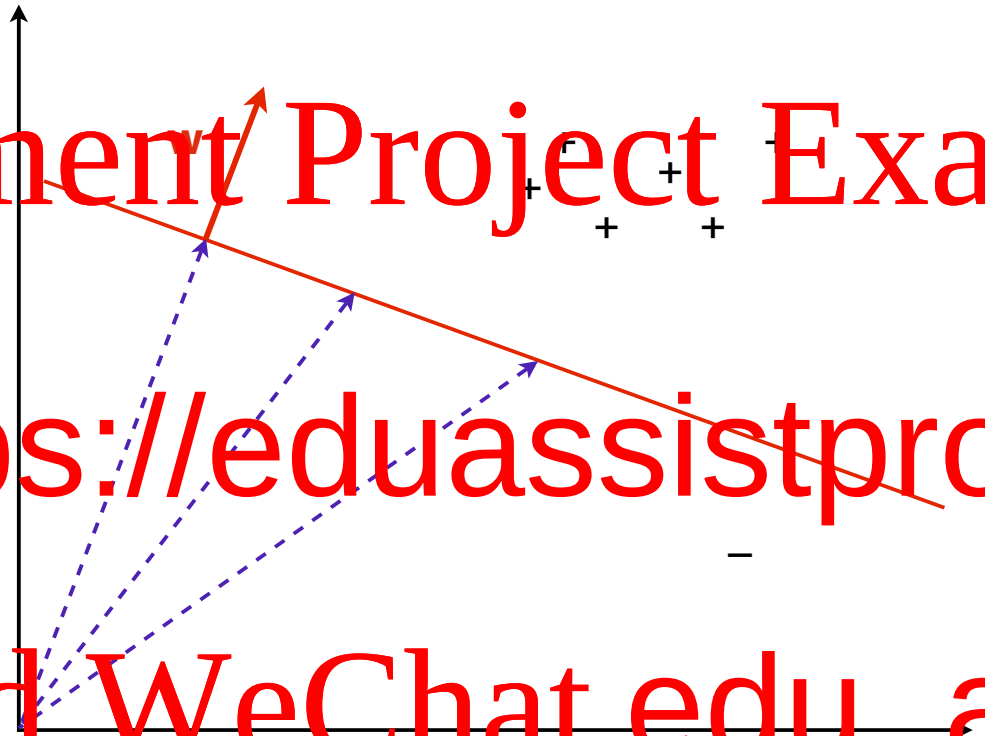
In fact, any weight between 2.5 and 5 will ensure that the th only exceeded when both tests succeed. We could even consider assigning different weights to the tests – as long as each weight is less than 5 and their sum exceeds 5 – although it is hard to see how this could be justified by the training data.

Spam filtering as a classification task

The columns marked x_1 and x_2 indicate the results of two tests on four different e-mails. The fourth column indicates which of the e-mails are spam. The right-most column demonstrates that by thresholding the function

E-mail	x_1	x_2	Spam?	$4x_1 + 4x_2$
1	1	1	1	8
2	0	0	0	0
3	1	0	0	4
4	0	1	0	4

Linear classification in two dimensions



- straight line separates positives from negatives
- defined by $\mathbf{w} \cdot \mathbf{x}_i = t$
- $\mathbf{w}$ is perpendicular to decision boundary
- $\mathbf{w}$ points in direction of positives
- t is the decision threshold

Linear classification in two dimensions

Note: $\mathbf{x}_i$ points in th $\mathbf{x}_0$ $\mathbf{w} \cdot \mathbf{x}_0 = ||\mathbf{w}||\ ||\mathbf{x}_0|| = t$ (where $||\mathbf{x}||$ denotes the length of the vector $\mathbf{x}$).

Homogeneous coordinates

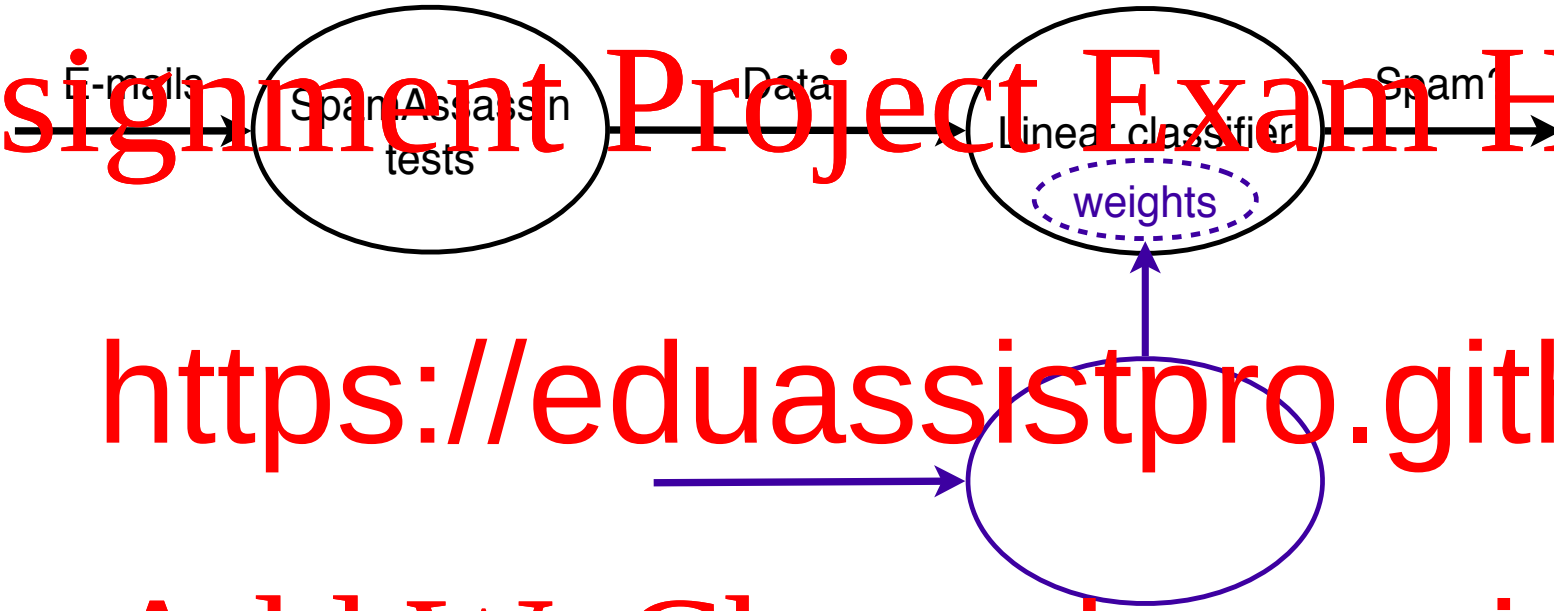
It is sometimes convenient to simplify notation further by introducing an extra constant 'variable' $x_0 = 1$, the weight of which is fixed to $w_0 = -t$.

The exten ended weight ve $\mathbf{w}^\circ \cdot \mathbf{x}^\circ$

Thanks to these so-called *homogeneous coo* dary passes through the origin of the extended coordinate s expense of needing an additional dimension.

Note: this doesn't really affect the data, as all data points and the 'real' decision boundary live in the plane $x_0 = 1$.

Machine learning for spam filtering



At the top we see how SpamAssassin approaches the spam e-mail classification task: the text of each e-mail is converted into a data point by means of SpamAssassin's built-in tests, and a *linear classifier* is applied to obtain a 'spam or ham' decision. At the bottom (in blue) we see the bit that is done by machine learning.

A Bayesian classifier I

Bayesian spam filters maintain a *vocabulary* of words and phrases – potential spam or ham indicators – for which statistics are collected from a *training set*.

- For i
  e-m
  tha
  this e-mail is spam are 4:1, or the probability of it being spam is 0.80 and the probability of it being ham is

- The situation is slightly more subtle because we h
  account the prevalence of spam. Suppose that I re
  one spam e-mail for every six ham e-mails. This means that I would estimate the odds of an unseen e-mail being spam as 1:6, i.e., non-negligible but not very high either.

A Bayesian classifier II

- If I then learn that the e-mail contains the word 'Viagra', which occ
two
sim
pro

In this way you are combining two independent pieces o
concerning the prevalence of spam, and the other conc
occurrence of the word 'Viagra', pulling in opposite dir

A Bayesian classifier III

The nice thing about this 'Bayesian' classification scheme is that it can be repeated i

in favour of s

and suppo

combined odds are 4:1 times 3:1 is 12:1, which is ample to o

1:6 odds associated with the low prevalence of spam (to

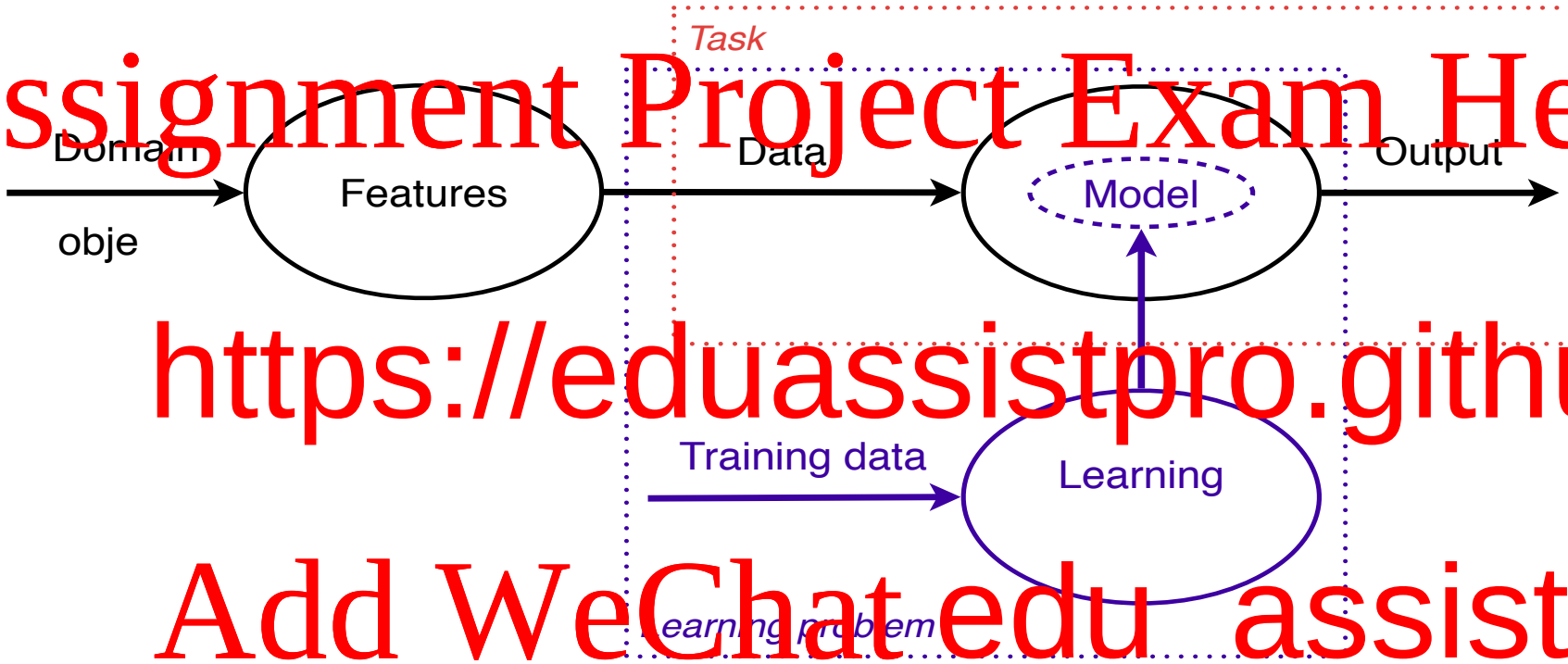
a spam probability of 0.67, up from 0.40

A rule-based classifier

- if the e-mail contains the word 'Viagra' then estimate the odds of spam as 4:1;
- oth
  of sp
- oth

The first rule covers all e-mails containing the word 'Via
whether they contain the phrase 'blue pill', so no overco
The second rule *only* covers e-mails containin
not the word 'Viagra', by virtue of the 'otherwise' clause
covers all remaining e-mails: those which neither contain neither 'Viagra' nor 'blue pill'.

How machine learning helps to solve a task



An overview of how machine learning is used to address a given task. A task (red box) requires an appropriate mapping – a model – from data described by features to outputs. Obtaining such a mapping from training data is what constitutes a learning problem (blue box).

Some terminology I

Tasks are a
learning al

Some terminology II

Machine l
right mod

Some terminology III

Models le
give it unity

Some terminology IV

Does the al
of learnin gorithm.

If however, it can continue to learn a new data arrives, it is an **online learning** algorithm.

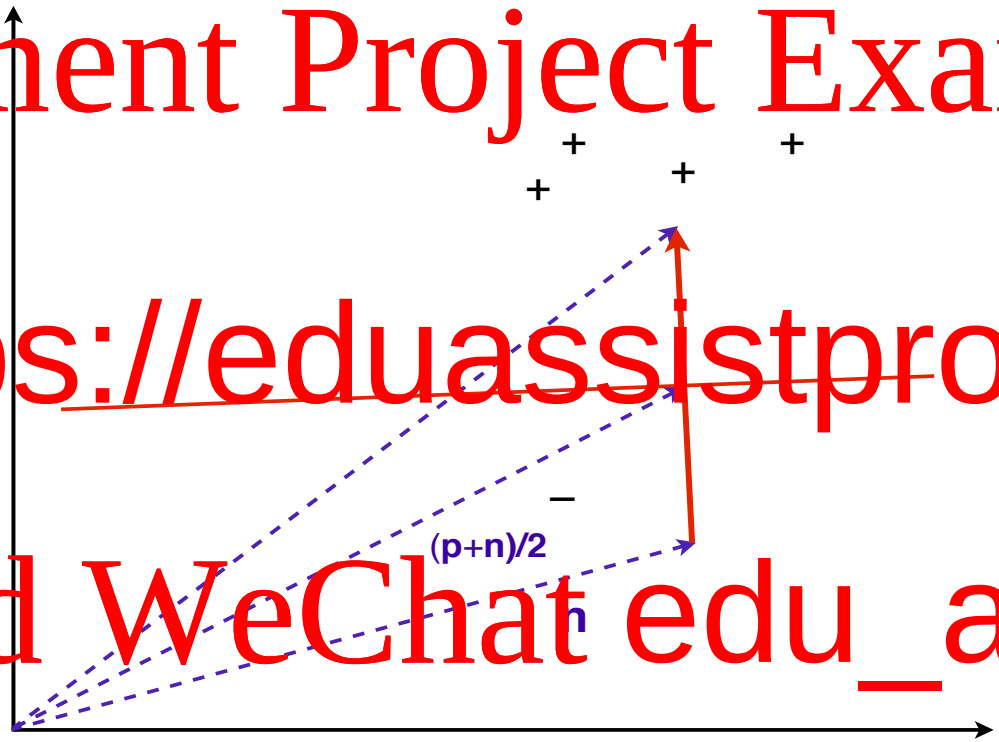
Some terminology V

If the mode

parame

Otherwise, if the number of parameters grows with the amount of training data it is categorised as **non-parametric**.

Basic linear classifier I



The basic linear classifier constructs a decision boundary by half-way intersecting the line between the positive and negative centres of mass.

Basic linear classifier II

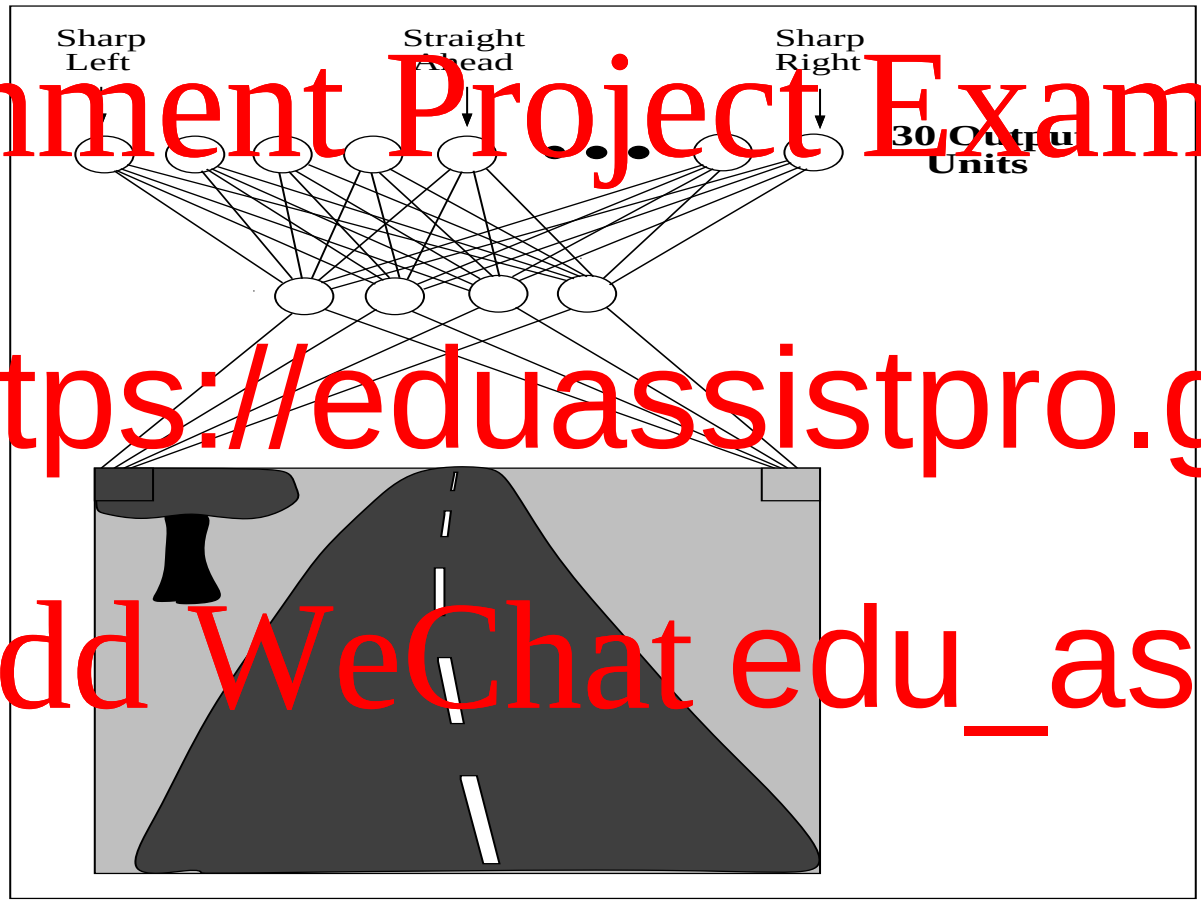
The basic li t, with
$\mathbf{w} = \mathbf{p}$ $(\mathbf{p} + \mathbf{n})/2$
is on the deci
$t = (\mathbf{p} - \mathbf{n}) \cdot (\mathbf{p} + \mathbf{n})/2 = (||\mathbf{p}|| - ||\mathbf{n}||)/2$, where $\mathbf{x}$ denotes the length of vector $\mathbf{x}$.

Neural Networks I

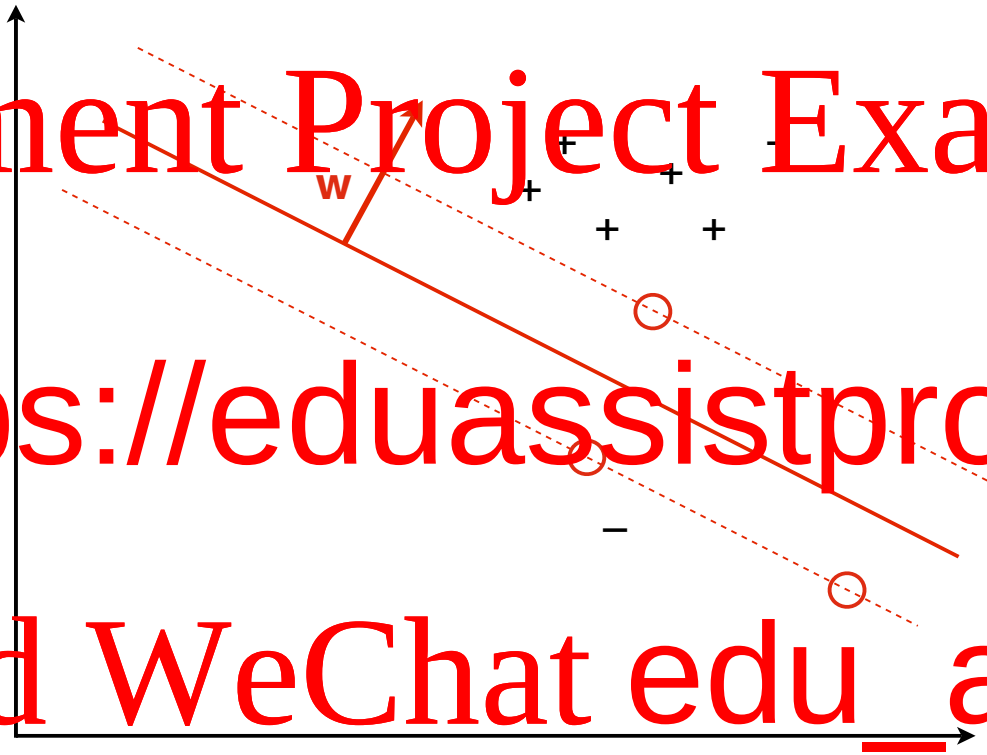
Neural Networks II

Neural Networks III

Neural Networks IV



Support vector machine



The decision boundary learned by a support vector machine from the linearly separable data from before. The decision boundary maximises the margin, which is indicated by the dotted lines. The circled data points are the support vectors.

A simple probabilistic model

'Viagra' and 'lottery' are two Boolean features; Y is the class variable, with value
indicate i

Viagra	lotter		ottery)
0	0	0.31	**0.69**
0	1	**0.65**	
1	0	**0.80**	
1	1	0.40	

Decision rule

Assuming that X and Y are the only variables we know and care about, the posterior distribution $P(Y\ X)$ helps us to answer many questions of interest.

- For
  'Via
  probability $P(Y = \text{spam}|\text{Viagra, lotte}$
  probability exceeds 0.5 and ham otherwi

- Such a recipe to predict a value of Y X
  and the posterior distribution $P(Y|X)$ is called a *decision rule*.

Missing values I

Suppose we skimmed an e-mail and noticed that it contains the word 'lottery' but we haven't looked closely enough to determine whether it uses the word 'Viagra'. This means that we don't know whether to use the second or t
would pre
and ham if it d

The solution is to average these two rows, using the prob
occurring in any e-mail (spam or not):

$$P(Y|\text{lottery}) = P(Y|\text{Viagra} = 0, \text{l}$$
$$+ P(Y|\text{Viagra} = 1, \text{lottery})P(\text{Viagra} = 1)$$

Missing values II

For instance, suppose for the sake of argument that one in ten e-mails contain th

$P(\text{Viagr}$

$P(Y = \text{s}$ nd

$P(Y = \text{ham}|\text{lottery} = 1) = 0.35 \cdot 0.90 + 0$ the

occurrence of 'Viagra' in any e-mail is relatively rare, the r distribution deviates only a little from the second row in t

Likelihood ratio

As a matter of fact, statisticians work very often with different conditional probabilities, given by the *likelihood function* $P(X|Y)$.

- I like to think of these as thought experiments: if somebody were to send me a spam e-mail, how likely would it be that it contains exactly the w
  ha
- Wh
  their ratio: how much more likely is it to observe this combination of words in a spam e-mail than it is in a non-spam e-mail

- For instance, suppose that for a particular e-mai X we have $P(X|Y = \text{spam}) = 3.5 \cdot 10^{-5}$ a $\cdot 10^{-6}$, then observing X in a spam e-mail is nearly five times more likely than it is in a ham e-mail.
- This suggests the following decision rule: predict spam if the likelihood ratio is larger than 1 and ham otherwise.

When to use likelihoods

Use likelih
uniform, a

Posterior odds

$$\frac{P(Y = \text{spam}|\text{Viagra} = 0, \text{lottery} = 0)}{P(Y = \text{ham}|\text{Viagra} = 0, \text{lottery} = 0)} = \frac{0.31}{0.69} = 0.45$$

$$\frac{P(Y = \text{spam}|\text{Viagra} = 1, \text{lottery}}{P(Y = \text{ham}|\text{Viagra} = 0, \text{lottery}}$$

$$\frac{P(Y = \text{spam}|\text{Viagra} = 1, \text{lotter}}{P(Y = \text{ham}|\text{Viagra} = 1, \text{lottery}}$$

Using a MAP decision rule we predict ham in the top two cases and spam in the bottom two. Given that the full posterior distribution is all there is to know about the domain in a statistical sense, these predictions are the best we can do: they are *Bayes-optimal*.

Example marginal likelihoods

Y	$P(\text{Viagra} = 1 \mid Y)$	$P(\text{Viagra} = 0 \mid Y)$
		0.6
		0.88

Y	$P(\text{lottery} = 1 \mid Y)$	
spam	0.21	0.79
ham	0.13	0.87

Using marginal likelihoods

Using the marginal likelihoods from before, we can approximate the likelihood ratios (the previously calculated odds from the full posterior distribution are shown in brackets):

$$\frac{P(\text{Viagra} = 0 \mid Y = \text{spam})}{P(\text{Viagr} \quad \frac{P(\text{lottery} = 0 \mid Y = \text{spam})}{} \quad \frac{0.60}{} \; \frac{0.79}{} \quad .62 \quad (0.45)$$

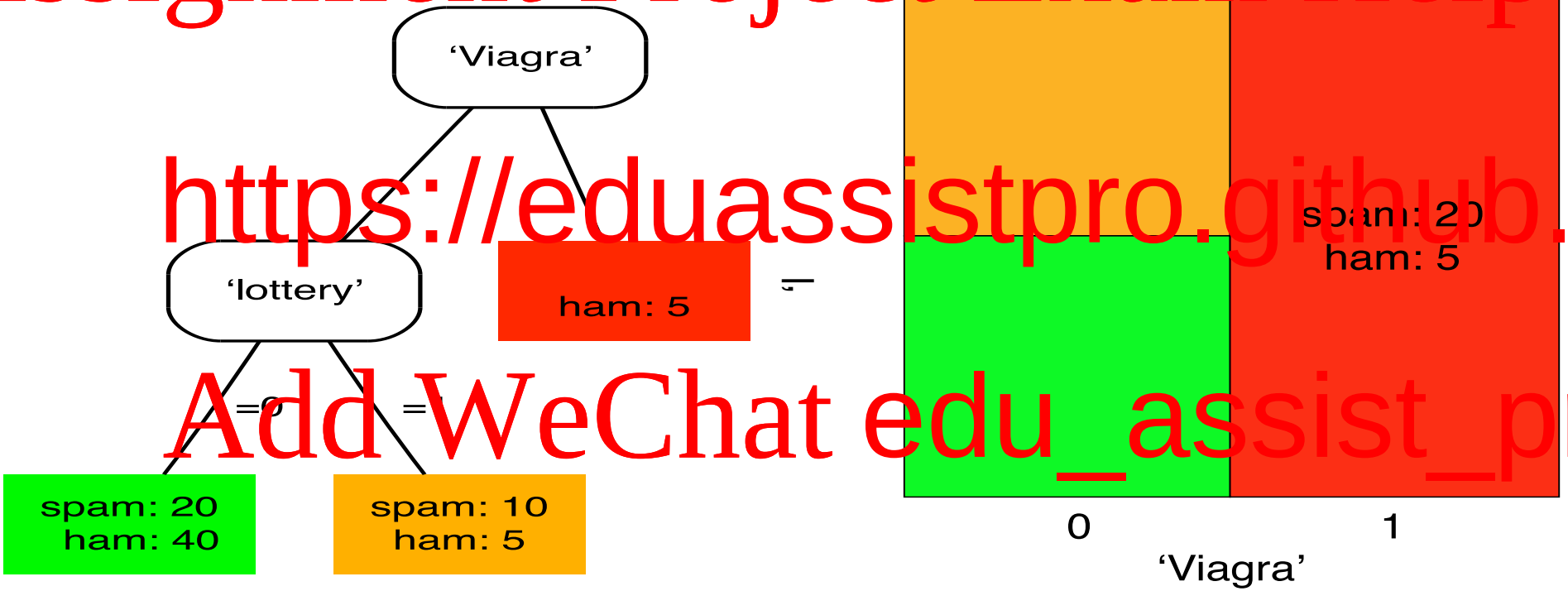
$$\frac{P(\text{Viagr}}{P(\text{Viagr} \quad \text{https://eduassistpro.github.i} \quad .1 \quad .9)$$

$$\frac{P(\text{Viagra} = 1 \mid Y = \text{spam})}{P(\text{Viagra} = 1 \mid Y = \text{ham})} \quad \frac{P(\text{lottery} = 0 \mid Y =}{} \quad \frac{0.40}{} \; \frac{0.79}{} \quad .0)$$

$$\frac{P(\text{Viagra} = 1 \mid Y = \text{spam})}{P(\text{Viagra} = 1 \mid Y = \text{ham})} \quad \frac{P(\text{lottery} = 1 \mid Y =}{} \quad \frac{}{} \; \frac{}{} \quad .67)$$

We see that, using a maximum likelihood decision rule, our very simple model arrives at the *Bayes-optimal* prediction in the first three cases, but not in the fourth ('Viagra' and 'lottery' both present), where the marginal likelihoods are actually very misleading.

A classification tree I



A classification tree combining two Boolean features.

A classification tree II

Each internal node or split is labelled with a feature, and each edge emanating from a split is labelled with a feature value.

Each leaf th

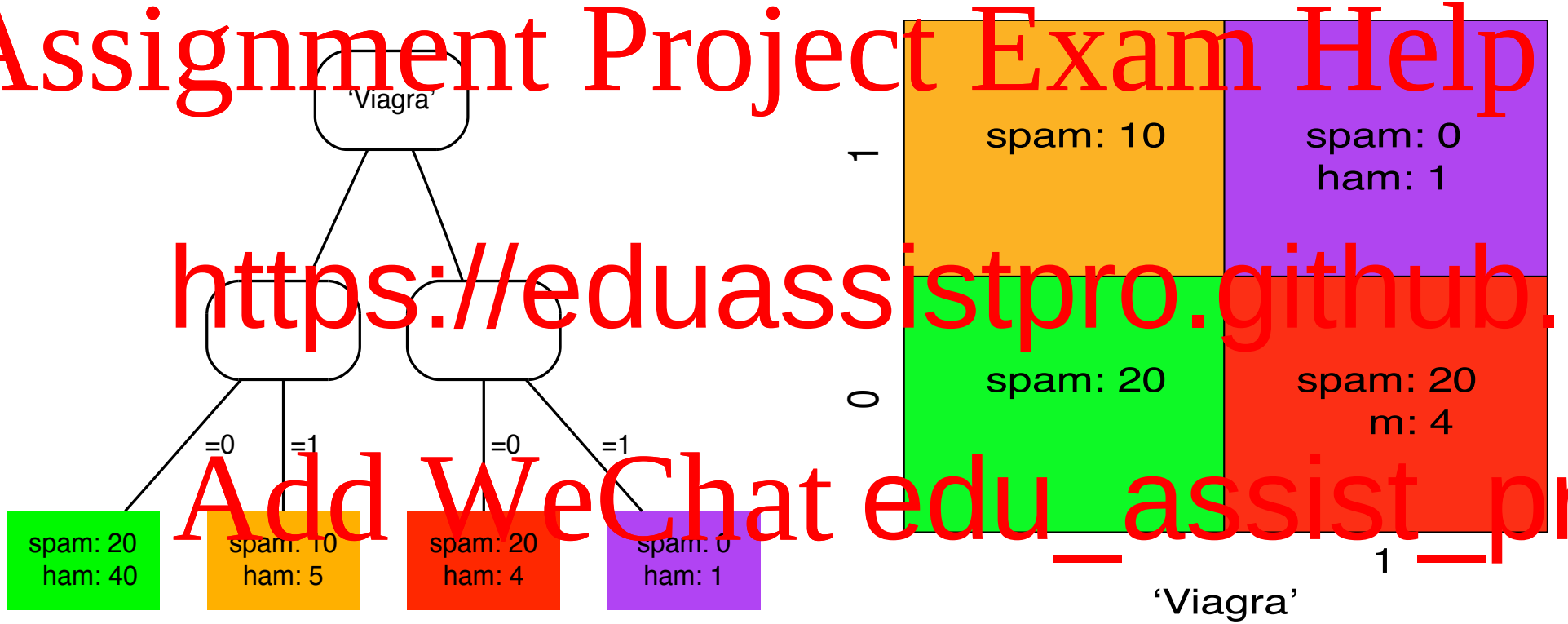
Also indic
training set.

A classification tree partitions the instance space into r
one for each leaf. We can clearly see that the majority of ham lives in the lower left-hand corner.

Labelling a classification tree

- The leaves of the classification tree could be labelled, from left to righ
  maj
- Alte
  occurring in each leaf: from left to right, 1/3, 2/3, and 4/5.
- Or, if our task was a regression task, we could label th
  predicted real values or even linear functions of so
  real-valued features.

A complete classification tree



A complete classification tree built from two Boolean features. The corresponding instance space partition is the finest partition that can be achieved with those two features.

Two uses of features

Suppose $1 \leq x \leq 1$.

A linear ap

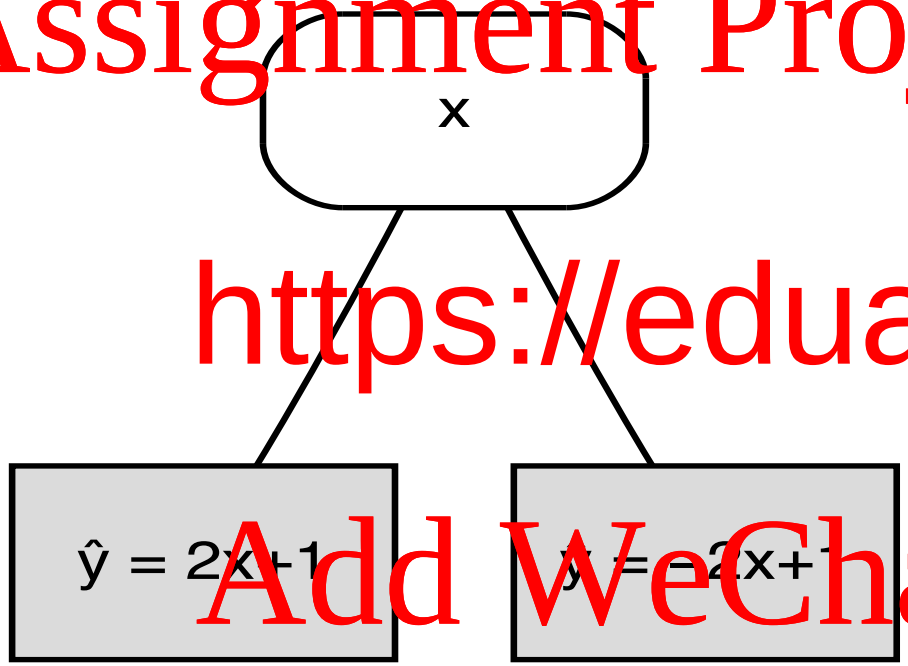
$y = 0$. Ho and

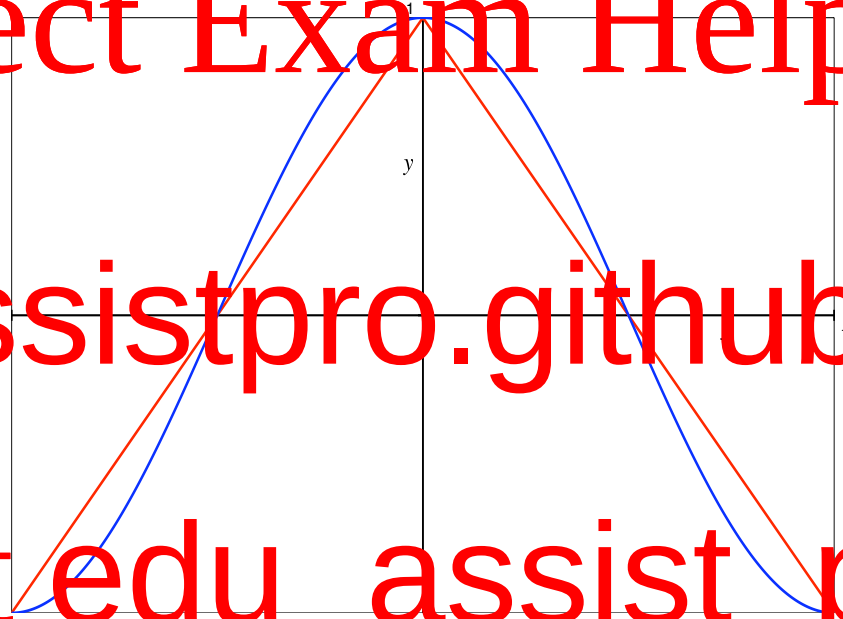
$0 \leq x \leq$

interval. We can achieve this by using x b

a regression variable.

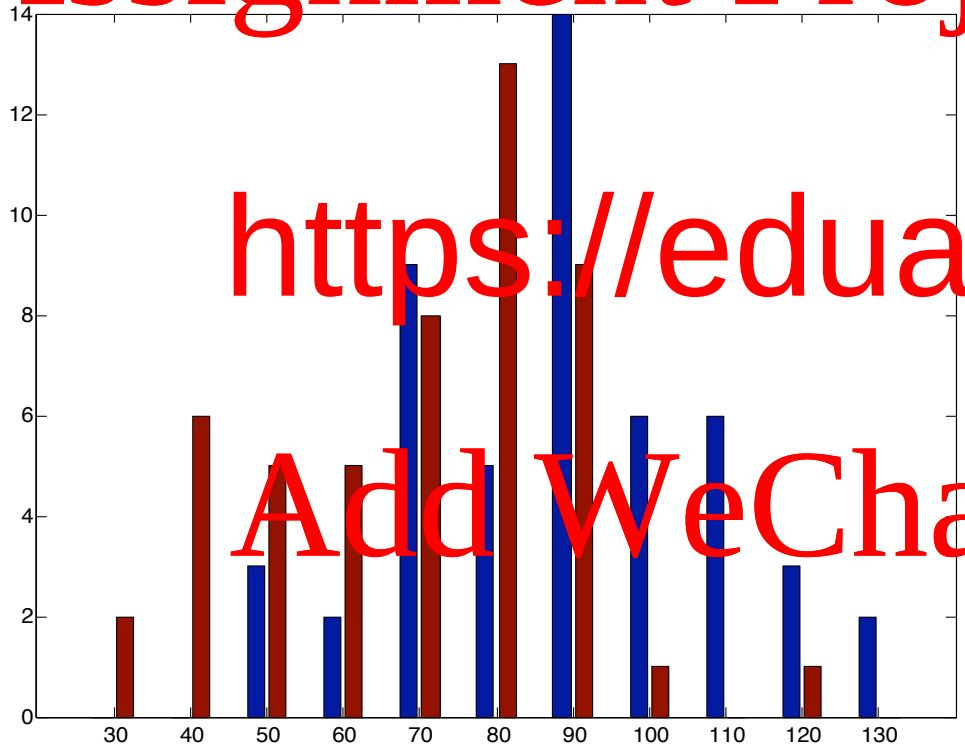
A small regression tree





A regression tree combining a one-split feature tree with linear regression models in the leaves. Note: x used as both a splitting feature and regression variable. At right, function $y = \cos \pi x$ on the interval $-1 \leq x \leq 1$, and piecewise linear approximation by regression tree.

Class-sensitive discretisation I



Class-sensitive discretisation II

Artificial data depicting a histogram of body weight measurements of people wit
of 10 kilogr

By joining t
eighth, ninth and tenth intervals, we obtain a discretis
proportion of diabetes cases increases from left to right
discretisation makes the feature more useful in predic

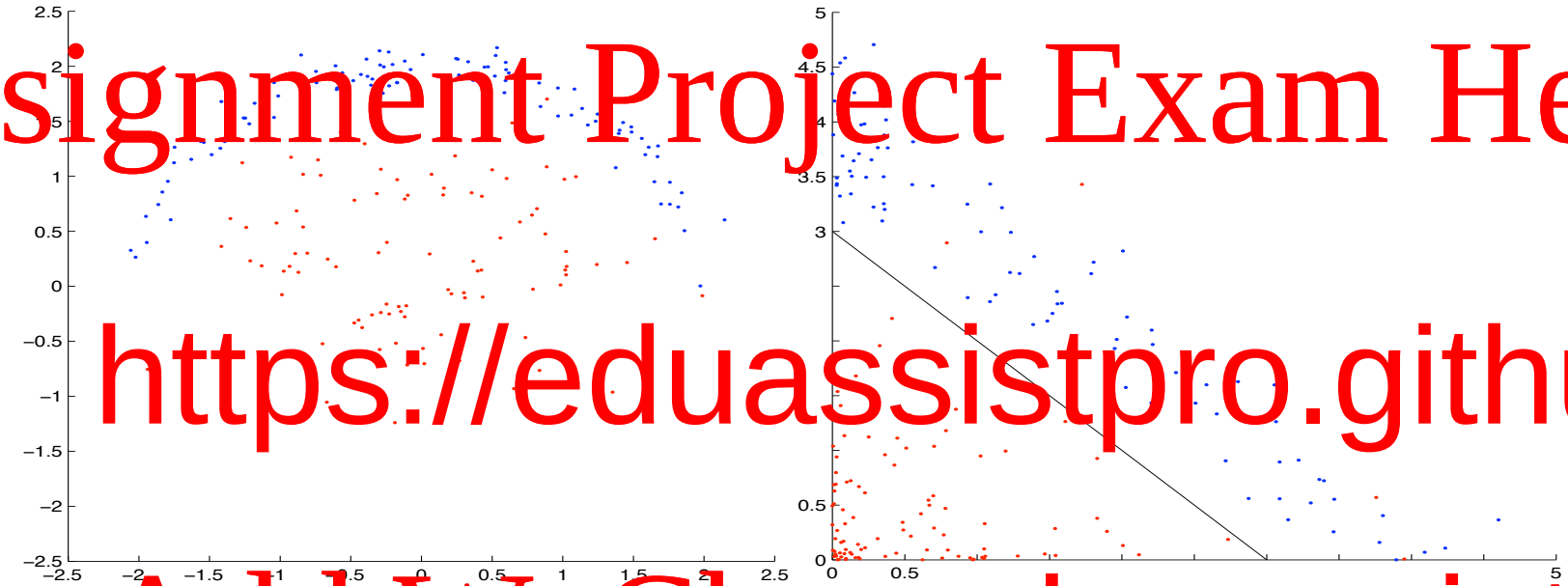
The kernel trick

Let $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x}_2 = (x_2, y_2)$ be two data points, and consider the mapping $(x, y) \mapsto (x^2, y^2, \sqrt{2}xy)$ to a three-dimensional feature space. The points in feature space corresponding to $\mathbf{x}_1$ and $\mathbf{x}_2$ are $\mathbf{x}'_1 = (x$ f these two f

$$\mathbf{x}'_1 \cdot \mathbf{x}_2 = x_1 x_2 + y_1 y_2 + 2x_1 y_1 x_2 y_2 = (x_1 x_2 + y_1 y_2) \quad = (\mathbf{x}_1 \quad \mathbf{x}_2)^2$$

That is, by squaring the dot product in the original space w dot product in the new space *without actua vectors*! A function that calculates the dot product in feature space directly from the vectors in the original space is called a *kernel* – here the kernel is $\boldsymbol{\kappa}(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 \cdot \mathbf{x}_2)^2$.

Non-linearly separable data



A *linear classifier* would perform poorly on this da
original (x, y) data into $(x', y') = (x^2, y^2)$, the data becomes more 'linear', and a linear decision boundary $x' + y' = 3$ separates the data fairly well. In the original space this corresponds to a circle with radius $\sqrt{3}$ around the origin.

Where do features come from ?

Feature engineering

Can features be learned ?

Yes, to some extent. For example, in the intermediate layers of a convolutional neural network[1].

[1] Image downloaded 28/2/18 from `http://devblogs.nvidia.com/deep-learning-nutshell-core-concepts/`

Where does data come from ?

AlphaZero played around 44 million chess games against an expert (itself).

Tasks for machine learning

The most common machine learning tasks are *predictive*, in the sense that they conc

- Bin
- Reg
- Clustering: hidden target
- Dimensionality reduction: intrinsic structur

Exploratory or *descriptive* tasks are concerne structure in the data.

Measuring similarity I

If our e-mails are described by word-occurrence features as in the text classification example, the similarity of e-mails would be measured in terms of the word

number of c

words occ *card coefficient*).

Suppose that one e-mail contains 42 (different) word contains 112 words, and the two e-mails have 23 words i their similarity would be $\frac{23}{42+112-23} = \frac{23}{130} = 0.18$.

Measuring similarity II

We can then cluster our e-mails into groups, such that the average similarity of an e-mail to the other e-mails in its group is much larger than the averag

While it wo
separated clusters corresponding to spam and ham – t
– the clusters may reveal some interesting and useful str
It may be possible to identify a particular kind of spam in th
subgroup uses a vocabulary, or language, not found in o

Clustering

We want to cluster similarly expressed genes in cancer samples.

How many clusters ? I

How many clusters ? II

Looking for structure I

Consider the following matrix:

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 2 & 2 \end{pmatrix}$$

Imagine these represent ratings by six different peopl
of 0 to 3, of four different films – say *The S* *The*
Usual Suspects, *The Godfather*, and *The Bi* m
left to right). *The Godfather* seems to be the most popular of the four with an average rating of 1.5, and *The Shawshank Redemption* is the least appreciated with an average rating of 0.5.
Can you see any structure in this matrix?

Looking for structure II

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & 0 & 1 \\ 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 1 \\ 0 & 2 & 2 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- The right-most matrix associates films (in colu rows). *The Shawshank Redemption* a elong to two different genres, say drama and crime, belongs to both, and *The Big Lebowski* is a crim new genre (say comedy).
- The tall, 6-by-3 matrix then expresses people's preferences in terms of genres.

Looking for structure III

- Fin
imp
pref

The philosophical problem

Deduction: derive specific consequences from general theories

Inductio

Deduction is well-founded (mathematical logic).

Induction is (philosophically) problematic – induct

often seems to work – an inductive argument !

Generalisation - the key objective of machine learning

What we are really interested in is *generalising* from the sample of data in our training set. This can be stated as:

The indu

> *Any well approximate the target function well over othe examples.*

A corollary of this is that it is necessary to make some assumptions about the type of target function in a task for an algorithm to go beyond the data, i.e., generalise or learn.

Cross-validation I

There are c
We use the d
use a separate *validation* or *development* set.

Cross-validation II

Cross-validation III

Cross-validation IV

Contingency table I

Two-class prediction case:

Yes	True Positive (TP)	
No	False Positive (FP)	

Contingency table II

Classification Accuracy on a sample of labelled pairs $(x, c\,x)$ given a learned classification model that predicts, for each instance x, a class value $\hat{c}(x)$:

$$\frac{\sum_{x \in \text{Test}} I[\hat{c}(x) = c\,x]}{}$$

where Test is a test set and $I[]$ is the indicat argument evaluates to true, and 0 otherwise.

Classification Error is 1-acc.

Overfitting

Imagine you are preparing for your *Machine Learning 101* exam. Helpfully, your professor has made previous exam papers and their worked answers available online. You begin by trying to answer the questions from previous papers an

Unfortu

the model a
completely consists of past questions, you are certain t
if the new exam asks different questions about the same
would be ill-prepared and get a much lower mark than wi
traditional preparation.

In this case, one could say that you were *overfitting* the past exam papers and that the knowledge gained didn't *generalise* to future exam questions.

The Bias-Variance Decomposition

Error (particularly MSE) can be shown to have two components:

Bias: e
learnabl

Variance: error due to variation between trainin
distribution on data generated by the target function

Inductive Bias

All m

Box & Drape

Inductive Bias

Confusingly, "inductive bias" is *NOT* the same "bias" as in the "bias-variance" decomposition.

"Inducti
on the mod

Essentially it means that the algorithm and model com
using to solve the learning problem is appropriate for th

Success in machine learning requires understandin
algorithms and models, and choosing them appropriately for the task[2].

[2] Even true for "deep learning", but watch Andrew Ng's talk on this at `http://www.youtube.com/watch?v=F1ka6a13S9I&t=48s`.

No Free Lunch

Wolpert (1996)

No Free Lunch

Owing to the "No Free Lunch" Theorem, there is no universally best machine learning algorithm.

Some lear
their assu
for the lear

On some other tasks, those assumptions, and hence th
perform much worse than others.

These assumptions form the inductive bias of the learning algorithm.

Ethics of machine learning

Machine learning algorithms are now widely available and increasingly used on "big data" in commercial, medical, legal and scientific applications.

These applications bring many benefits, but as they become more widespre

In some of th
machine learning must be considered

For example, can the use of machine learning be biased o
the training data does not adequately represent all of th
the model may be applied ? Here is a recent paper discussing some of these issues and giving some recommendations[3].

[3] Zook et al. (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. http://doi.org/10.1371/journal.pcbi.1005399.

Summary

The purpose of this introductory lecture was, from a high-level, to survey the landscape that we will explore in this course

We have motivated the subject matter, outlined what parts of machine learning w
and termi

Through
programming languages, machine learning tools a
These change very rapidly, although the core technic
challenges remain.

Following this lecture you should have a clear idea of the course scope. In the remaining lectures we will expand on the topics we have just introduced and go into more detail. You will also get to work on practical applications of what we have covered.