Machine Learning 2007: Slides 1

Instructor: Tim van Erven (Tim.van.Erven@cwi.nl)
Website: www.cwi.nl/˜erven/teaching/0708/ml/

September 6, 2007

Overview

- **Course Organisation**
- Tentative Course Outline
- What is Machine Learning?
- This Lecture versus Mitchell
- Supervised versus Unsupervised Learning
- The Most Important Supervised Learning Problems
  - Prediction
  - Regression
  - Classification
- Hypotheses and Hypothesis Spaces
- Least Squares

People

Instructor: Tim van Erven

- E-mail: Tim.van.Erven@cwi.nl
- Bio:
  - Studied AI at the University of Amsterdam
  - Currently a PhD student at the Centrum voor Wiskunde en Informatica (CWI) in Amsterdam
  - Research focuses on the Minimum Description Length (MDL) principle for learning and prediction

Teaching Assistent: Rogier van het Schip

- E-mail: rsp400@few.vu.nl
- Bio:
  - 6th year AI student
  - Intends to start graduation work this year

Course Materials

Materials:

- “Machine Learning” by Tom M. Mitchell, McGraw-Hill, 1997
- Extra materials (on course website)
- Slides (on course website)

Course Website:

`www.cwi.nl/~erven/teaching/0708/ml/`

Important Note:

I will not always stick to the book. Don’t forget to study the slides and extra materials!

Grading

Part	Relative Weight
Homework assignments	40%
Intermediate exam	20%
Final exam ($\geq$ 5.5)	40%

- $5 \leq$ average grade $\leq 6 \Rightarrow$ round to whole point
- Else $\Rightarrow$ round to half point
- To pass: rounded average grade $\geq$ 6 **AND** final exam ≥ 5.5

Homework Assignments

- Should be submitted using Blackboard before the deadline (on the assignment)
- Late submissions:
  - Solutions discussed in class $\Rightarrow$ reject
  - Else $\Rightarrow$ minus half a point per day
- Exclude lowest grade
- Average assignment grades, no rounding
- Unsubmitted $\Rightarrow 1$

Homework Assignments

- Usually theoretical exercises (math or theory)
- One practical assignment using Weka
- One essay assignment near the end of the course

Overview

- Course Organisation
- **Tentative Course Outline**
- What is Machine Learning?
- This Lecture versus Mitchell
- Supervised versus Unsupervised Learning
- The Most Important Supervised Learning Problems
  - Prediction
  - Regression
  - Classification
- Hypotheses and Hypothesis Spaces
- Least Squares

Tentative Course Outline

Date	Topic
Sept. 6, 13	**Basic concepts**, list-then-eliminate algorithm, decision trees
Sept. 20	Neural networks
Sept. 27	Instance-based learning: k-nearest neighbour classifier
Oct. 4	Naive Bayes
Oct. 11	Bayesian learning
Oct. 18	Minimum description length (MDL) learning
?	**Intermediate Exam**
Oct. 31	Statistical estimation (don't read Mitchell sect. 5.5.1!)
Nov. 7	Support vector machines
Nov. 14	Computational learning theory: PAC learning, VC dimension
Nov. 21	Graphical models
Nov. 28	Unsupervised learning: clustering
Dec. 5	-
Dec. 12	The grounding problem, discussion, questions
?	**Final exam**

Overview

- Course Organisation
- Tentative Course Outline
- **What is Machine Learning?**
- This Lecture versus Mitchell
- Supervised versus Unsupervised Learning
- The Most Important Supervised Learning Problems
  - Prediction
  - Regression
  - Classification
- Hypotheses and Hypothesis Spaces
- Least Squares

Machine Learning

"Machine Learning is the study of computer algorithms that improve automatically through experience." – T. M. Mitchell

For example:

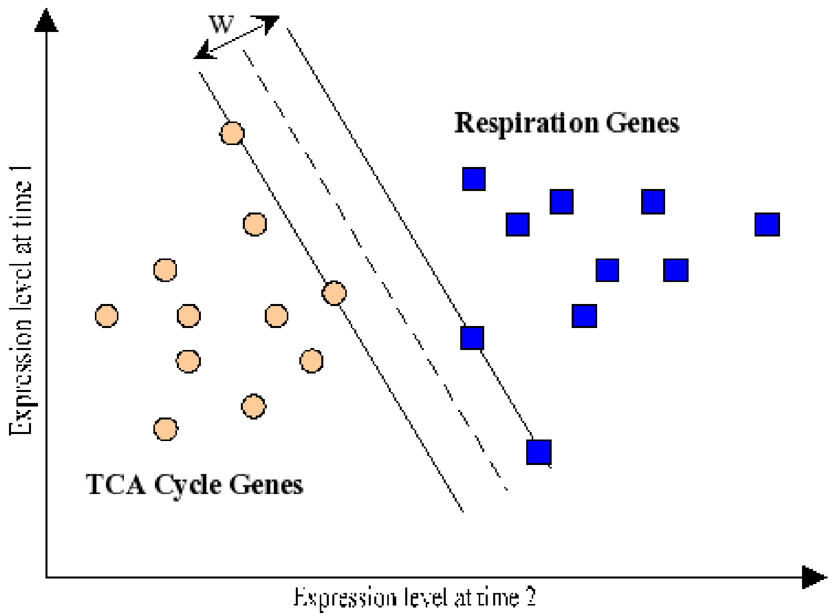
- Handwritten digit recognition: examples from MNIST database (figure taken from [LeCun et al., 1998])

Machine Learning

"Machine Learning is the study of computer algorithms that improve automatically through experience." – T. M. Mitchell

For example:

- Handwritten digit recognition: examples from MNIST database (figure taken from [LeCun et al., 1998])
- Classifying genes by gene expression (figure taken from [Molla et al.])



Machine Learning

"Machine Learning is the study of computer algorithms that improve automatically through experience." – T. M. Mitchell

For example:

- Handwritten digit recognition: examples from MNIST database (figure taken from [LeCun et al., 1998])
- Classifying genes by gene expression (figure taken from [Molla et al.])
- Evaluating a board state in checkers based on a set of board features. E.g. the number of black pieces on the board. (c.f. Mitchell)

Deduction versus Induction

We will (mostly) consider induction rather than deduction.

Deduction: a particular case from general principles

1. You need at least a 6 to pass this course. ($A \rightarrow B$)
2. You have achieved at least a 6. (A)
3. Hence, you pass this course. (Therefore B)

Induction: general laws from particular facts

Name	Average Grade	Pass?
Sanne	7.5	Yes
Sem	6	Yes
Lotte	5	No
Ruben	9	Yes
Sophie	7	Yes
Daan	4	No
Lieke	6	Yes
Me	8	?

Why Machine Learning?

- Too much data to analyse by humans (e.g. ranking websites, spam filtering, classifying genes by gene expression)
- Too difficult data representations (e.g. 3D brain scans, angle measurements on joints of an industrial robot)
- Algorithms for machine learning keep improving
- Computation is cheap; humans are expensive
- Some jobs are too boring for humans (e.g. spam filtering)
- . . .

Overview

- Course Organisation
- Tentative Course Outline
- What is Machine Learning?
- **This Lecture versus Mitchell**
- Supervised versus Unsupervised Learning
- The Most Important Supervised Learning Problems
  - Prediction
  - Regression
  - Classification
- Hypotheses and Hypothesis Spaces
- Least Squares

This Lecture versus Mitchell

Mitchell, Chapter 1 and Chapter 2 up to section 2.2

- Very abstract and general, but non-standard framework for machine learning programs (Figures 1.1 and 1.2)
- Hard to see similarities between different machine learning algorithms in this framework

This Lecture

- Important in science: Separate the problem from its solution
- Standard categories of machine learning problems
- Less general than Mitchell, but provides more solid ground (I hope you will see what I mean by that)

What should you study? Both.

Overview

- Course Organisation
- Tentative Course Outline
- What is Machine Learning?
- This Lecture versus Mitchell
- **Supervised versus Unsupervised Learning**
- The Most Important Supervised Learning Problems
  - Prediction
  - Regression
  - Classification
- Hypotheses and Hypothesis Spaces
- Least Squares

Supervised versus Unsupervised Learning

- Unsupervised learning: only unlabeled training examples
  - We have data $D = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$
  - Find interesting patterns
  - E.g. group data into clusters
- Supervised learning: labeled training examples
  - We have data $D = \begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ \mathbf{x}_n \end{pmatrix}$
  - Learn to predict a label y for any unseen case $\mathbf{x}$
- Semi-supervised learning: some of the training examples have been labeled

Overview

- Course Organisation
- Tentative Course Outline
- What is Machine Learning?
- This Lecture versus Mitchell
- Supervised versus Unsupervised Learning
- The Most Important Supervised Learning Problems
    - **Prediction**
    - Regression
    - Classification
- Hypotheses and Hypothesis Spaces
- Least Squares

Prediction

Definition:

Given data

$$D = y_1, \ldots, y_n,$$

predict how the sequence continues with

$$y_{n+1}$$

- Prediction is supervised learning: we only get the labels. There are no feature vectors $\mathbf{x}$.

Prediction Examples (deterministic)

A simple sequence:

- $D = 2, 4, 6, \ldots$

Prediction Examples (deterministic)

A simple sequence:

- $D = 2, 4, 6, \ldots$

But wait, suppose I tell you a few more numbers:

- $D = 2, 4, 6, 10, 16, \ldots$

Prediction Examples (deterministic)

A simple sequence:

- $D = 2, 4, 6, \ldots$

But wait, suppose I tell you a few more numbers:

- $D = 2, 4, 6, 10, 16, \ldots$

Another easy one:

- $D = 1, 4, 9, 16, 25, \ldots$

Prediction Examples (deterministic)

A simple sequence:

- $D = 2, 4, 6, \ldots$

But wait, suppose I tell you a few more numbers:

- $D = 2, 4, 6, 10, 16, \ldots$

Another easy one:

- $D = 1, 4, 9, 16, 25, \ldots$

I doubt whether you will get this one:

- $D = 1, 4, 2, 2, 4, 1, 0, 1, 4, 2, \ldots$

Prediction Examples (deterministic)

A simple sequence:

- $D = 2, 4, 6, \dots$

But wait, suppose I tell you a few more numbers:

- $D = 2, 4, 6, 10, 16, \dots$

Another easy one:

- $D = 1, 4, 9, 16, 25, \dots$

I doubt whether you will get this one:

- $D = 1, 4, 2, 2, 4, 1, 0, 1, 4, 2, \dots$ (squares modulo 7)

Doesn't have to be numbers:

- $D = a, b, b, a, a, a, b, b, b, b, a, a, \dots$

The Necessity of Bias

We have seen that $D = 2, 4, 6, \ldots$ can continue as

$$D = 2, 4, 6, \ldots \begin{cases} \ldots, 8, 10, 12, 14, \ldots \\ \ldots, 10, 16, 26, 42, \ldots \end{cases}$$

- Why did you prefer the first continuation when you clearly also accepted the second one?
- What about $\ldots, 2, 4, 6, 2, 4, 6, 2, 4, \ldots$?
- Why not $\ldots, 7, 1, 9, 3, 3, 3, 3, 3, \ldots$?

The Necessity of Bias

We have seen that $D = 2, 4, 6, \ldots$ can continue as

$$D = 2, 4, 6, \ldots \begin{cases} \ldots, 8, 10, 12, 14, \ldots \\ \ldots, 10, 16, 26, 42, \ldots \end{cases}$$

- Why did you prefer the first continuation when you clearly also accepted the second one?
- What about $\ldots, 2, 4, 6, 2, 4, 6, 2, 4, \ldots$?
- Why not $\ldots, 7, 1, 9, 3, 3, 3, 3, 3, \ldots$?

Bias is unavoidable!

Prediction Examples (statistical)

Independent and identically distributed (i.i.d.)

$$P(y_1) = P(y_2) = P(y_3) = \dots$$

- $D = 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, \dots$

Prediction Examples (statistical)

Independent and identically distributed (i.i.d.)

$P(y_1) = P(y_2) = P(y_3) = \ldots$

- $D = 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, \ldots$ ($P(y = 1) = 1/6$)
- $D = 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, \ldots$

Prediction Examples (statistical)

Independent and identically distributed (i.i.d.)

$P(y_1) = P(y_2) = P(y_3) = \dots$

- $D = 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, \dots$ $(P(y = 1) = 1/6)$
- $D = 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, \dots$ $(P(y = 1) = 1/2)$

Dependent on the previous outcome (Markov Chain)

$P(y_{i+1}|y_1, \dots, y_i) = P(y_{i+1}|y_i)$

- $D = 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, \dots$

Prediction Examples (statistical)

Independent and identically distributed (i.i.d.)

$P(y_1) = P(y_2) = P(y_3) = \ldots$

- $D = 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, \ldots$
  $(P(y = 1) = 1/6)$
- $D = 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, \ldots$
  $(P(y = 1) = 1/2)$

Dependent on the previous outcome (Markov Chain)

$P(y_{i+1}|y_1, \ldots, y_i) = P(y_{i+1}|y_i)$

- $D = 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, \ldots$
  $P(y_{i+1} = y_i|y_i) = 5/6$

Prediction Examples (real world 1)

What will be the outcome of the next horse race?

$D =$

		Race				
Horse	Owner	1	2	3	4	5
Jolly Jumper	Lucky Luke	4th	1st	4th	4th	4th
Lightning	Old Shatterhand	2nd	2nd	3rd	2nd	2nd
Sleipnir	Wodan	1st	4th	1st	1st	1st
Bucephalus	Alex. the Great	3rd	3rd	2nd	3rd	3rd

Prediction Examples (real world 1)

What will be the outcome of the next horse race?

$D =$

Horse	Owner	Race 1	Race 2	Race 3	Race 4	Race 5
Jolly Jumper	Lucky Luke	4th	1st	4th	4th	4th
Lightning	Old Shatterhand	2nd	2nd	3rd	2nd	2nd
Sleipnir	Wodan	1st	4th	1st	1st	1st
Bucephalus	Alex. the Great	3rd	3rd	2nd	3rd	3rd

- Is there any deterministic or statistical regularity?
- Can we say that there is a true distribution that determines these outcomes?

(Okay, I made up this example, but this way is more fun than taking the results from a real race.)

Prediction Examples (real world 2)

$D =$ “The problem of inducing general functions from specific training ex. . . ”

Prediction Examples (real world 2)

$D =$ "The problem of inducing general functions from specific training ex. . . " (Mitchell, Ch.2)

- Is there any deterministic or statistical regularity?
- Can we say that there is one true distribution that determines the next outcome?
- Should we consider this sentence an instance of
  - the population of sentences in Mitchell's book,
  - the population of sentences written by Mitchell,
  - the population of books about Machine Learning,
  - the population of English sentences?

Prediction Examples (real world 2)

$D =$ "The problem of inducing general functions from specific training ex. . . " (Mitchell, Ch.2)

- Is there any deterministic or statistical regularity?
- Can we say that there is one true distribution that determines the next outcome?
- Should we consider this sentence an instance of
  - the population of sentences in Mitchell's book,
  - the population of sentences written by Mitchell,
  - the population of books about Machine Learning,
  - the population of English sentences?

All are possible and all have different statistical regularities. . .

Prediction Again (to help you remember)

Definition:

Given data

$$D = y_1, \ldots, y_n,$$

predict how the sequence continues with

$$y_{n+1}$$

- Simple example: $D = 1, 1, 2, 3, 5, 8, \ldots$ (Fibonacci sequence)

Overview

- Course Organisation
- Tentative Course Outline
- What is Machine Learning?
- This Lecture versus Mitchell
- Supervised versus Unsupervised Learning
- The Most Important Supervised Learning Problems
  - Prediction
  - **Regression**
  - Classification
- Hypotheses and Hypothesis Spaces
- Least Squares

Regression

Definition:

Given data

$$D = \begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ \mathbf{x}_n \end{pmatrix},$$

learn to predict the value of the label y for any new feature vector $\mathbf{x}$.

- Typically y can take infinitely many values (e.g. $y \in \mathbb{R}$).
- This may be viewed as prediction of y with extra side-information $\mathbf{x}$.
- Sometimes y is called the **regression** variable and $\mathbf{x}$ the **regressor** variable.
- Sometimes y is called the **dependent** variable and $\mathbf{x}$ the **independent** variable.

Regression Example

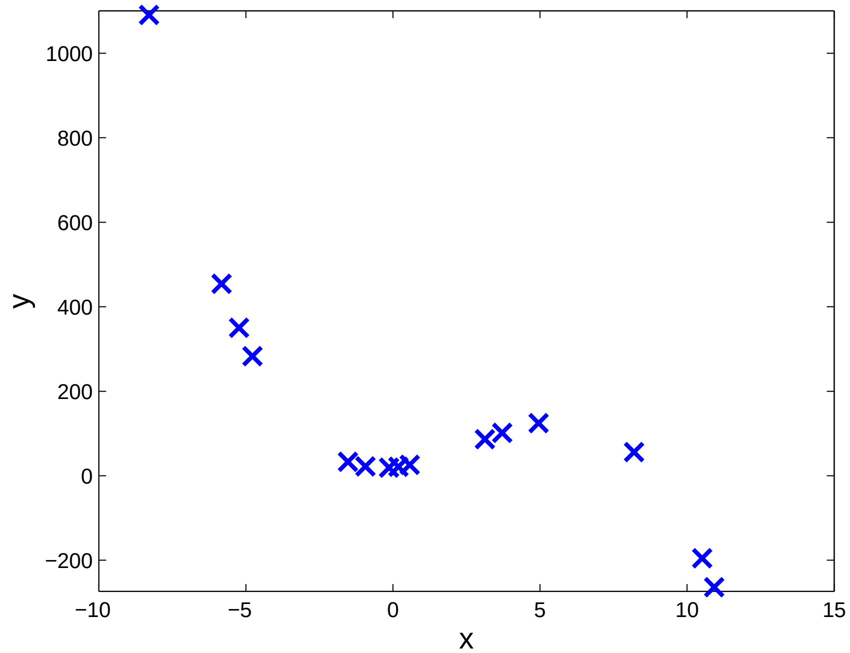
y	1090.5	350.4	283.1	454.5	19.3	33.2	25.9	22.2
$\mathbf{x}$	-8.3	-5.2	-4.8	-5.8	-0.1	-1.5	0.6	-0.9

y	21.4	86.5	101.4	56.0	124.4	-263.6	-195.3
$\mathbf{x}$	0.2	3.1	3.7	8.2	4.9	10.9	10.5

Regression Example

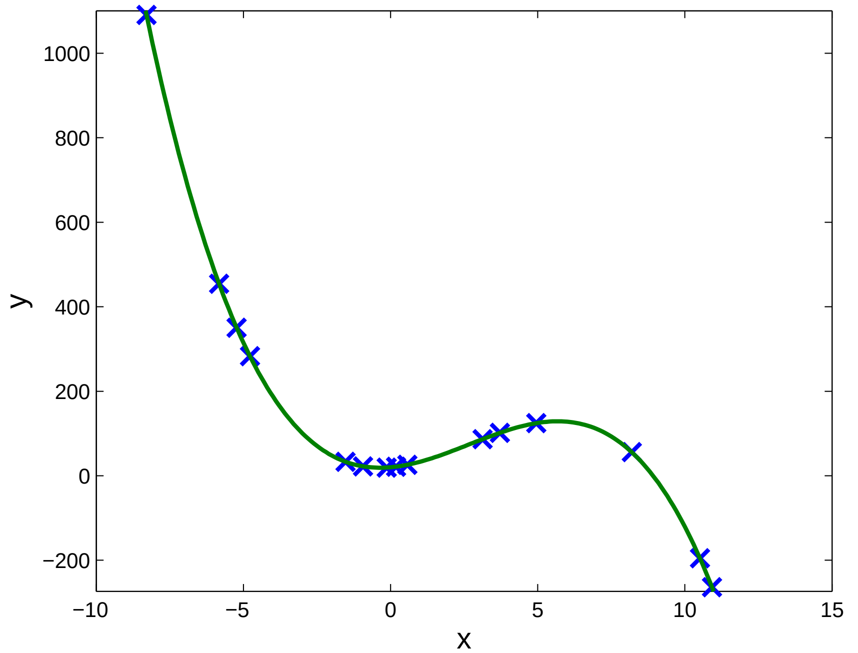
y	1090.5	350.4	283.1	454.5	19.3	33.2	25.9	22.2
$\mathbf{x}$	-8.3	-5.2	-4.8	-5.8	-0.1	-1.5	0.6	-0.9

y	21.4	86.5	101.4	56.0	124.4	-263.6	-195.3
$\mathbf{x}$	0.2	3.1	3.7	8.2	4.9	10.9	10.5

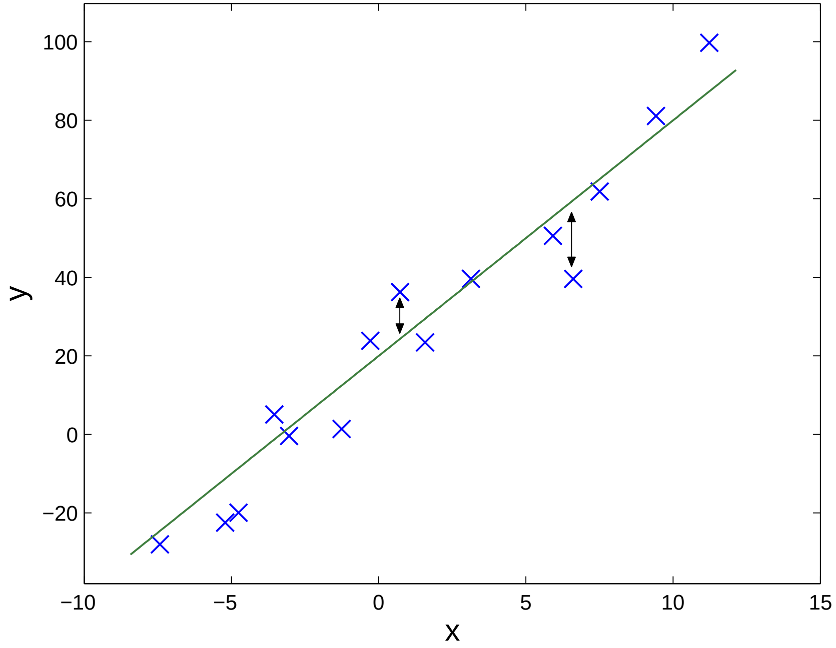


Regression Example

y	1090.5	350.4	283.1	454.5	19.3	33.2	25.9	22.2
$\mathbf{x}$	-8.3	-5.2	-4.8	-5.8	-0.1	-1.5	0.6	-0.9
y	21.4	86.5	101.4	56.0	124.4	-263.6	-195.3	
$\mathbf{x}$	0.2	3.1	3.7	8.2	4.9	10.9	10.5	



Example: A Linear Function with Noise



Data generated by a linear function plus Gaussian noise in y:

$$y = 6x + 20 + \mathcal{N}(0, 10)$$

Regression: Can we recover this function from the data alone?

Regression Repeated

Definition:

Given data

$$D = \begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \dots, \begin{pmatrix} y_n \\ \mathbf{x}_n \end{pmatrix},$$

learn to predict the value of the label y for any new feature vector $\mathbf{x}$.

- Typically y can take infinitely many values (e.g. $y \in \mathbb{R}$).

Overview

- Course Organisation
- Tentative Course Outline
- What is Machine Learning?
- This Lecture versus Mitchell
- Supervised versus Unsupervised Learning
- The Most Important Supervised Learning Problems
  - Prediction
  - Regression
  - **Classification**
- Hypotheses and Hypothesis Spaces
- Least Squares

Classification

Definition:

Given data

$$D = \begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \dots, \begin{pmatrix} y_n \\ \mathbf{x}_n \end{pmatrix},$$

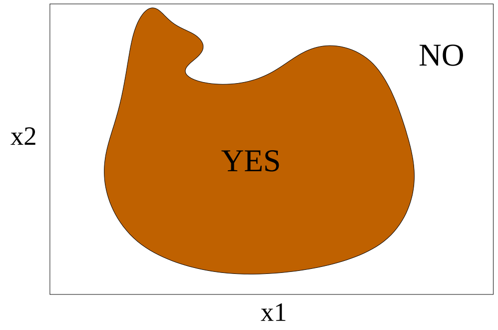
learn to predict the class label y for any new feature vector $\mathbf{x}$.

- The class label y only has a finite number of possible values, often only two (e.g. $y \in \{-1, 1\}$).
- Seems a special case of regression, but there is a difference:
- There is no notion of distance between class labels: Either the label is correct or it is wrong. You cannot be almost right.

Concept Learning

Definition:

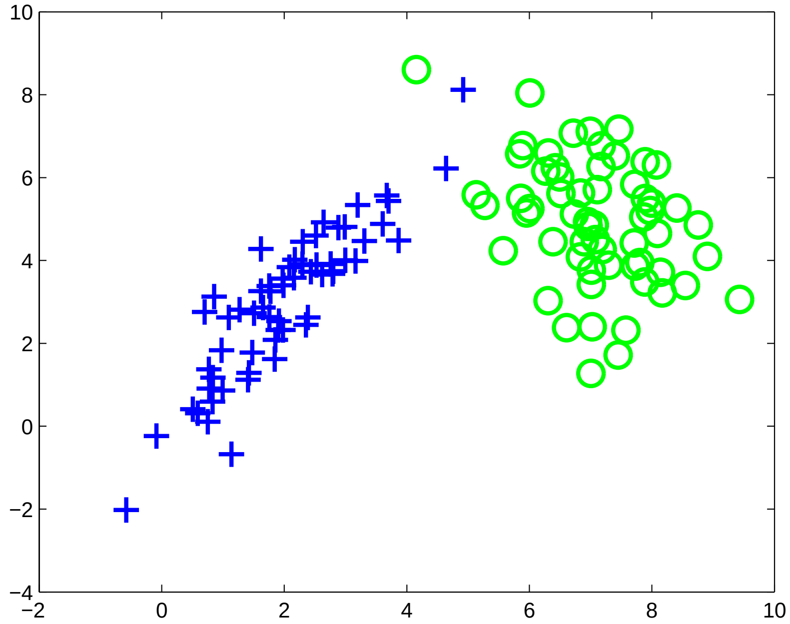
Concept learning is the specific case of classification where the label y can only take on two possible values: $\mathbf{x}$ is part of the concept or not.



Enjoysport Example

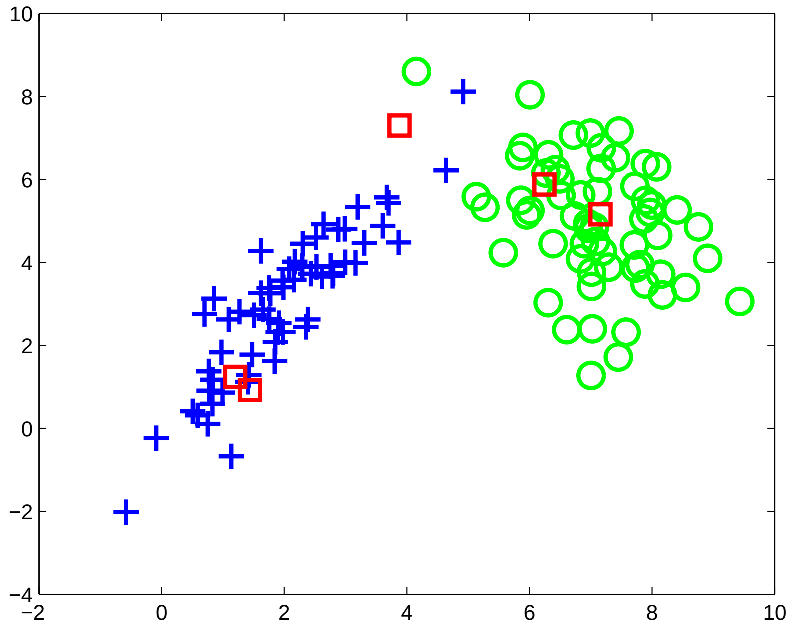
$\mathbf{x}$					y
Sky	**AirTemp**	**Humidity**	**Water**	**Forecast**	**EnjoySport**
Sunny	Warm	Normal	Warm	Same	Yes
Sunny	Warm	High	Warm	Same	Yes
Rainy	Cold	High	Warm	Change	No
Sunny	Warm	High	Cool	Change	?

Classification Example



- NB Visualisation is different from regression example: the value of y is shown using colour, not as an axis. The feature vectors $\mathbf{x} \in \mathbb{R}^2$ are 2-dimensional.
- To which class do you think the red squares belong?

Classification Example



- NB Visualisation is different from regression example: the value of y is shown using colour, not as an axis. The feature vectors $\mathbf{x} \in \mathbb{R}^2$ are 2-dimensional.
- To which class do you think the red squares belong?

Summary of Machine Learning Categories

Prediction: Given data $D = y_1, \ldots, y_n$, predict how the sequence continues with y_{n+1}

Regression: Given data $D = \begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ \mathbf{x}_n \end{pmatrix}$, learn to predict the value of the label y for any new feature vector $\mathbf{x}$. Typically y can take infinitely many values. Acceptable if your prediction is close to the correct y.

Classification: Given data $D = \begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ \mathbf{x}_n \end{pmatrix}$, learn to predict the class label y for any new feature vector $\mathbf{x}$. Only finitely many categories. Your prediction is either correct or wrong.

- Not all machine learning problems fit into these categories.
- We will see a few more categories during the course .

Categorizing Machine Learning Problems

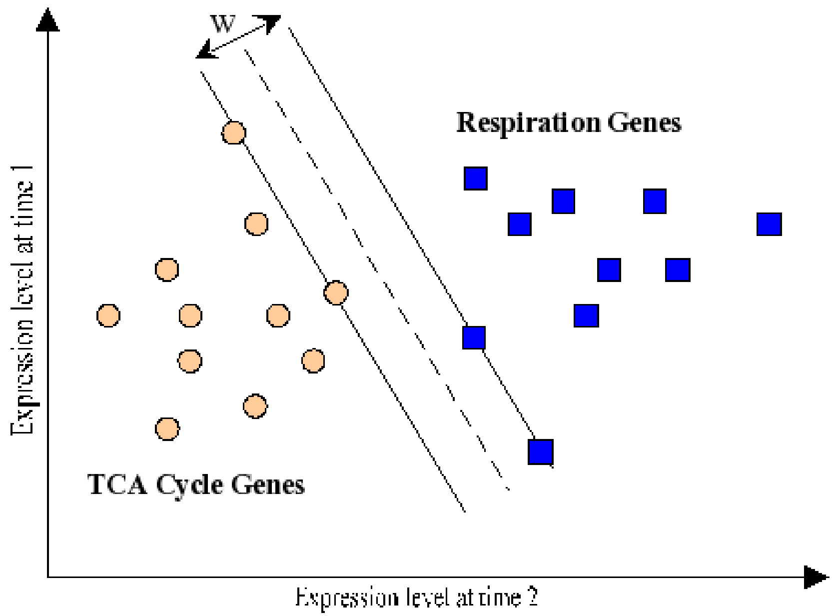
- Handwritten digit recognition
- Classifying genes by gene expression
- Evaluating a board state in checkers based on a set of board features

Categorizing Machine Learning Problems

- Handwritten digit recognition: classification
- Classifying genes by gene expression
- Evaluating a board state in checkers based on a set of board features

Categorizing Machine Learning Problems

- Handwritten digit recognition: classification
- Classifying genes by gene expression
- Evaluating a board state in checkers based on a set of board features



Categorizing Machine Learning Problems

- Handwritten digit recognition: classification
- Classifying genes by gene expression: classification
- Evaluating a board state in checkers based on a set of board features

Categorizing Machine Learning Problems

- Handwritten digit recognition: classification
- Classifying genes by gene expression: classification
- Evaluating a board state in checkers based on a set of board features: regression

Overview

- Course Organisation
- Tentative Course Outline
- What is Machine Learning?
- This Lecture versus Mitchell
- Supervised versus Unsupervised Learning
- The Most Important Supervised Learning Problems
  - Prediction
  - Regression
  - Classification
- **Hypotheses and Hypothesis Spaces**
- Least Squares

Hypotheses and Hypothesis Spaces

Definition of a Hypothesis:

A hypothesis h is a candidate description of the regularity or pattern in your data.

- Prediction example: $y_{n+1} = y_{n-1} + y_n$
- Regression example: $y = 5x$
- Classification example: $y = \begin{cases} +1 & \text{if } 3x + 20 > 0; \\ -1 & \text{else.} \end{cases}$

Definition of a Hypothesis Space:

A hypothesis space $\mathcal{H}$ is a set $\{h\}$ of hypotheses.

Example Hypothesis Space: Linear Regression

Linear Regression

In linear regression the goal is to select a linear hypothesis that best captures the regularity in the data.

Hypotheses

A linear hypothesis $h_{\mathbf{w},b} : \mathbb{R}^d \to \mathbb{R}$ is of the form:

$$y = h_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b,$$

where d denotes the number of features in/dimensionality of $\mathbf{x}$, and $\mathbf{w}$ and b are called **weights**.

Hypothesis Space

The corresponding hypothesis space $\mathcal{H}$ is

$$\mathcal{H} = \{h_{\mathbf{w},b} : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

Overview

- Course Organisation
- Tentative Course Outline
- What is Machine Learning?
- This Lecture versus Mitchell
- Supervised versus Unsupervised Learning
- The Most Important Supervised Learning Problems
  - Prediction
  - Regression
  - Classification
- Hypotheses and Hypothesis Spaces
- **Least Squares**

Least Squares for Linear Regression

Squared Error

For given $\mathbf{w}$ and b, we may evaluate the squared error of $h_{\mathbf{w},b}$ on a single data-item $(y_i, \mathbf{x}_i)^\top$:

$$(y_i - h_{\mathbf{w},b}(\mathbf{x}_i))^2 = (y_i - \mathbf{w}^\top \mathbf{x}_i - b)^2.$$

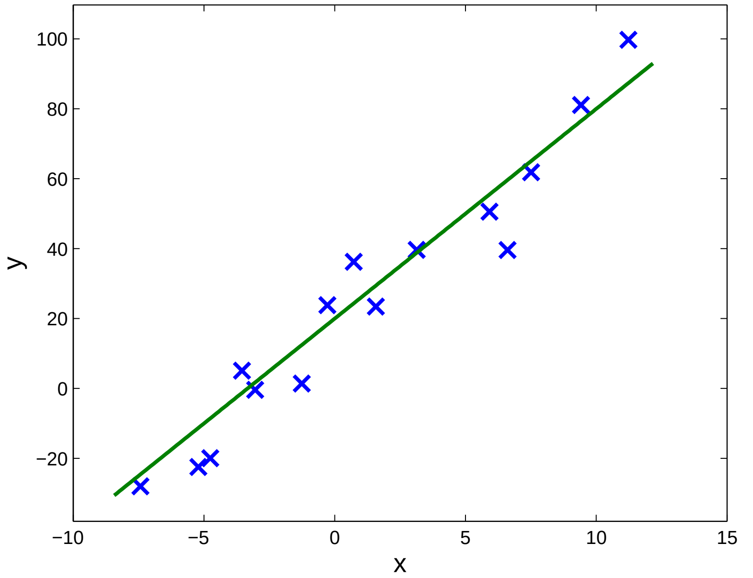
Least Squares Linear Regression

Select $\mathbf{w}$ and tb such that they minimize the sum of squared errors (SSE) on all data:

$$\min_{\mathbf{w},b} \text{SSE} = \min_{\mathbf{w},b} \sum_{i=1}^{n} (y_i - h_{\mathbf{w},b}(\mathbf{x}_i))^2.$$

Linear Regression Example

The previous example again:

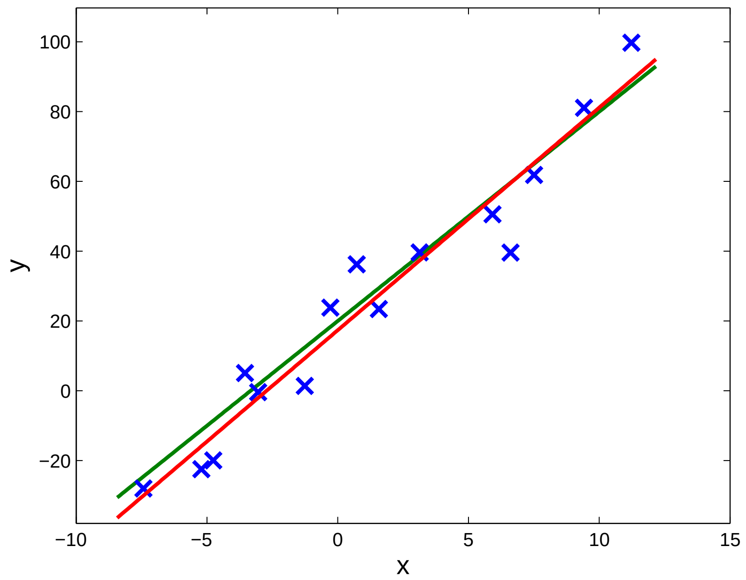


Original Function

$$y = 6x + 20$$

Linear Regression Example

The previous example again:



Original Function	Least Squares
$y = 6x + 20$	$y = 6.38x + 17.37$

Bibliography

- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ”Gradient-Based Learning Applied to Document Recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
- M. Molla, M. Waddell, D. Page & J. Shavlik (2004). Using Machine Learning to Design and Interpret Gene-Expression Microarrays. AI Magazine, 25, pp. 23-44. (To Appear in the Special Issue on Bioinformatics)
- N. Cristianini and J. Shawe-Taylor, “Support Vector Machines and other kernel-based learning methods,” Cambridge University Press, 2000