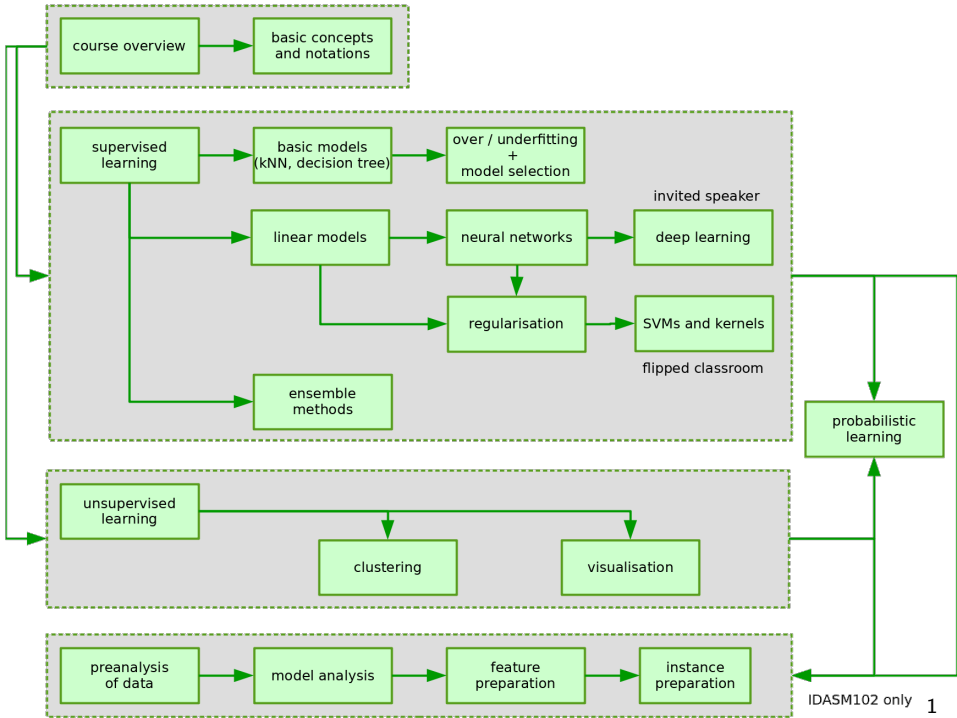


Machine Learning: Lesson 3

Introduction to Supervised Learning

Benoît Frénay - Faculty of Computer Science



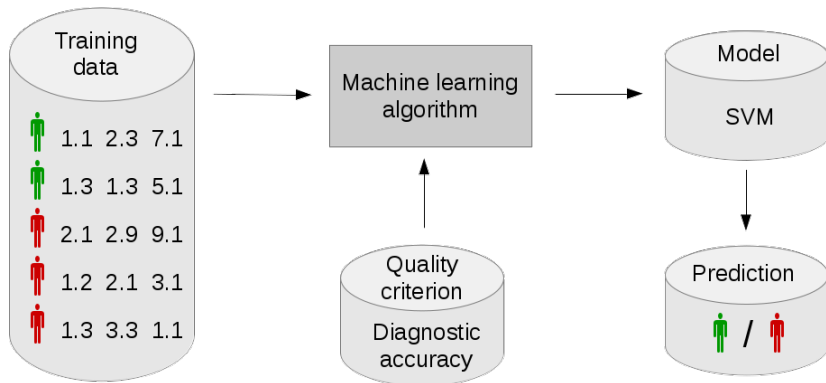


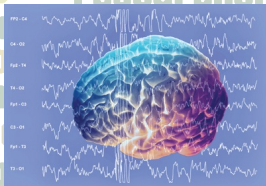
Outline of this Lesson

- what is supervised learning
- what is generalisation

What is Supervised Learning?

Supervised Learning in a Nutshell





Definition of Supervised Learning

Artificial Intelligence: a Modern Approach by Russel and Norvig

The type of feedback available for learning is usually the most important factor in determining the nature of the learning problem that the agent faces. The problem of supervised learning involves learning a function from examples of its inputs and outputs.

The Elements of Statistical Learning by Hastie et al.

Supervised learning is called "supervised" because of the presence of the outcome variable to guide the learning process.

Definition of Supervised Learning

Artificial Intelligence: a Modern Approach by Russel and Norvig

The type of feedback available for learning is usually the most important factor in determining the nature of the learning problem that the agent faces. The problem of supervised learning involves learning a function from examples of its inputs and outputs.

The Elements of Statistical Learning by Hastie et al.

Supervised learning is called "supervised" because of the presence of the outcome variable to guide the learning process.

Definition of Supervised Learning

Pattern Recognition and Machine Learning by Bishop

Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems. Cases in which the aim is to assign each input vector to one of a finite number of discrete categories are classification problems. If the desired output consist of one or more continuous variables, then the task is regression.

Introduction to Machine Learning by Alpaydin

Both classification and regression are supervised learning algorithms where there is an input X and the output Y , and the task is to learn the mapping from the input to the output. We assume a model defined up to a set of parameters: $y = g(\mathbf{x}|\theta)$.

Definition of Supervised Learning

Pattern Recognition and Machine Learning by Bishop

Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems. Cases in which the aim is to assign each input vector to one of a finite number of discrete categories are classification problems. If the desired output consist of one or more continuous variables, then the task is regression.

Introduction to Machine Learning by Alpaydin

Both classification and regression are supervised learning algorithms where there is an input X and the output Y , and the task is to learn the mapping from the input to the output. We assume a model defined up to a set of parameters: $y = g(\mathbf{x}|\boldsymbol{\theta})$.

Examples of Supervised Learning Problems

Classification

- face recognition
- emotion recognition
- spam filtering
- automated diagnosis
- elderly fall detection

Regression

- artificial pancreas controller
- house price estimation
- survival analysis

Formalisation of Supervised Learning

Supervised learning in machine learning

goal = find the best parameters of the model $y = f(\mathbf{x}|\boldsymbol{\theta})$ for the assigned task by minimising an error criterion on the training data $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$

- $f(\mathbf{x}|\boldsymbol{\theta})$ = function that takes \mathbf{x} as input
- $\boldsymbol{\theta}$ = set of parameters of the model
- the value of the error depends on both \mathcal{D} and $\boldsymbol{\theta}$

Learning under supervision

the supervision is given by the targets **and** the error

- targets tell us what we should learn (e.g. find cats in pictures)
- error criteria tell us what errors we should care about (e.g. it is not so serious to not detect an hairless sphynx, but do not confuse with dogs)

Formalisation of Supervised Learning

Supervised learning in machine learning

goal = find the best parameters of the model $y = f(\mathbf{x}|\boldsymbol{\theta})$ for the assigned task by minimising an error criterion on the training data $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$

- $f(\mathbf{x}|\boldsymbol{\theta})$ = function that takes \mathbf{x} as input
- $\boldsymbol{\theta}$ = set of parameters of the model
- the value of the error depends on both \mathcal{D} and $\boldsymbol{\theta}$

Learning under supervision

the supervision is given by the targets and the error

- targets tell us what we should learn (e.g. find cats in pictures)
- error criteria tell us what errors we should care about (e.g. it is not so serious to not detect an hairless sphynx, but do not confuse with dogs)

Formalisation of Supervised Learning

Supervised learning in machine learning

goal = find the best parameters of the model $y = f(\mathbf{x}|\boldsymbol{\theta})$ for the assigned task by minimising an error criterion on the training data $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$

- $f(\mathbf{x}|\boldsymbol{\theta})$ = function that takes \mathbf{x} as input
- $\boldsymbol{\theta}$ = set of parameters of the model
- the value of the error depends on both \mathcal{D} and $\boldsymbol{\theta}$

Learning under supervision

the supervision is given by the targets and the error

- targets tell us what we should learn (e.g. find cats in pictures)
- error criteria tell us what errors we should care about (e.g. it is not so serious to not detect an hairless sphynx, but do not confuse with dogs)

Formalisation of Supervised Learning

Supervised learning in machine learning

goal = find the best parameters of the model $y = f(\mathbf{x}|\boldsymbol{\theta})$ for the assigned task by minimising an error criterion on the training data $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$

- $f(\mathbf{x}|\boldsymbol{\theta})$ = function that takes \mathbf{x} as input
- $\boldsymbol{\theta}$ = set of parameters of the model
- the value of the error depends on both \mathcal{D} and $\boldsymbol{\theta}$

Learning under supervision

the supervision is given by the targets and the error

- targets tell us what we should learn (e.g. find cats in pictures)
- error criteria tell us what errors we should care about (e.g. it is not so serious to not detect an hairless sphynx, but do not confuse with dogs)

Formalisation of Supervised Learning

Supervised learning in machine learning

goal = find the best parameters of the model $y = f(\mathbf{x}|\boldsymbol{\theta})$ for the assigned task by minimising an error criterion on the training data $\mathcal{D} = \{(\mathbf{x}_i, t_i)\}$

- $f(\mathbf{x}|\boldsymbol{\theta})$ = function that takes \mathbf{x} as input
- $\boldsymbol{\theta}$ = set of parameters of the model
- the value of the error depends on both \mathcal{D} and $\boldsymbol{\theta}$

Learning under supervision

the supervision is given by the targets **and** the error

- targets tell us what we should learn (e.g. find cats in pictures)
- error criteria tell us what errors we should care about (e.g. it is not so serious to not detect an hairless sphynx, but do not confuse with dogs)

Example of Supervised Learning Problem

Prediction of student grades with a linear model

- 1 you collect $\underbrace{x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}}}_{\text{features}}$ and $\underbrace{t_{\text{grade}}}_{\text{target}}$ for each student
- 2 you assume that $f(x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}} | \underbrace{w_{\text{lesson}}, w_{\text{study}}, w_{\text{Facebook}}}_{\text{parameters}}) =$
 $\underbrace{w_{\text{lesson}} x_{\text{lesson}} + w_{\text{study}} x_{\text{study}} + w_{\text{Facebook}} x_{\text{Facebook}}}_{\text{model prediction = weighted sum of feature values}}$ can model students

Possible models

- $f(x|w) = 0.5 x_{\text{Facebook}}$
- $f(x|w) = 0.5 x_{\text{lesson}}$
- $f(x|w) = 0.2 x_{\text{lesson}} + 0.3 x_{\text{study}} + 0.01 x_{\text{Facebook}}$

criterion to choose model parameters = mean error on predicted grades /
error rate to predict academic success / % of errors smaller than 2/20 / ...

Example of Supervised Learning Problem

Prediction of student grades with a linear model

- 1 you collect $x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}}$ and t_{grade} for each student

$\underbrace{x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}}}_{\text{features}}$

$\underbrace{t_{\text{grade}}}_{\text{target}}$
- 2 you assume that $f(x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}} | \underbrace{w_{\text{lesson}}, w_{\text{study}}, w_{\text{Facebook}}}_{\text{parameters}}) =$
 $\underbrace{w_{\text{lesson}} x_{\text{lesson}} + w_{\text{study}} x_{\text{study}} + w_{\text{Facebook}} x_{\text{Facebook}}}_{\text{model prediction = weighted sum of feature values}}$ can model students

Possible models

- $f(x|w) = 0.5 x_{\text{Facebook}}$
- $f(x|w) = 0.5 x_{\text{lesson}}$
- $f(x|w) = 0.2 x_{\text{lesson}} + 0.3 x_{\text{study}} + 0.01 x_{\text{Facebook}}$

criterion to choose model parameters = mean error on predicted grades /
error rate to predict academic success / % of errors smaller than 2/20 / ...

Example of Supervised Learning Problem

Prediction of student grades with a linear model

- 1 you collect $x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}}$ and t_{grade} for each student
features target
- 2 you assume that $f(x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}} | \underbrace{w_{\text{lesson}}, w_{\text{study}}, w_{\text{Facebook}}}_{\text{parameters}}) =$
 $\underbrace{w_{\text{lesson}} x_{\text{lesson}} + w_{\text{study}} x_{\text{study}} + w_{\text{Facebook}} x_{\text{Facebook}}}_{\text{model prediction = weighted sum of feature values}}$ can model students

Possible models

- $f(x|w) = 0.5x_{\text{Facebook}}$

- $f(x|w) = 0.5x_{\text{lesson}}$

- $f(x|w) = 0.2x_{\text{lesson}} + 0.3x_{\text{study}} - 0.01x_{\text{Facebook}}$

criterion to choose model parameters = mean error on predicted grades /
error rate to predict academic success / % of errors smaller than 2/20 /

Example of Supervised Learning Problem

Prediction of student grades with a linear model

- 1 you collect $x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}}$ and t_{grade} for each student
features target
- 2 you assume that $f(x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}} | \underbrace{w_{\text{lesson}}, w_{\text{study}}, w_{\text{Facebook}}}_{\text{parameters}}) =$
 $\underbrace{w_{\text{lesson}} x_{\text{lesson}} + w_{\text{study}} x_{\text{study}} + w_{\text{Facebook}} x_{\text{Facebook}}}_{\text{model prediction = weighted sum of feature values}}$ can model students

Possible models

- $f(x|w) = 0.5 x_{\text{Facebook}}$
- $f(x|w) = 0.5 x_{\text{lesson}}$
- $f(x|w) = 0.2 x_{\text{lesson}} + 0.3 x_{\text{study}} - 0.01 x_{\text{Facebook}}$

criterion to choose model parameters = mean error on predicted grades /
error rate to predict academic success / % of errors smaller than 2/20 / ...

Example of Supervised Learning Problem

Prediction of student grades with a linear model

- 1 you collect $x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}}$ and t_{grade} for each student
features target
- 2 you assume that $f(x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}} | \underbrace{w_{\text{lesson}}, w_{\text{study}}, w_{\text{Facebook}}}_{\text{parameters}}) =$
 $\underbrace{w_{\text{lesson}} x_{\text{lesson}} + w_{\text{study}} x_{\text{study}} + w_{\text{Facebook}} x_{\text{Facebook}}}_{\text{model prediction = weighted sum of feature values}}$ can model students

Possible models

- $f(\mathbf{x}|\mathbf{w}) = 0.5 x_{\text{Facebook}}$
- $f(\mathbf{x}|\mathbf{w}) = 0.5 x_{\text{lesson}}$
- $f(\mathbf{x}|\mathbf{w}) = 0.2 x_{\text{lesson}} + 0.3 x_{\text{study}} - 0.01 x_{\text{Facebook}}$

criterion to choose model parameters = mean error on predicted grades /
error rate to predict academic success / % of errors smaller than 2/20 / ...

Example of Supervised Learning Problem

Prediction of student grades with a linear model

- 1 you collect $x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}}$ and t_{grade} for each student
features target
- 2 you assume that $f(x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}} | \underbrace{w_{\text{lesson}}, w_{\text{study}}, w_{\text{Facebook}}}_{\text{parameters}}) =$
 $\underbrace{w_{\text{lesson}} x_{\text{lesson}} + w_{\text{study}} x_{\text{study}} + w_{\text{Facebook}} x_{\text{Facebook}}}_{\text{model prediction = weighted sum of feature values}}$ can model students

Possible models

- $f(\mathbf{x}|\mathbf{w}) = 0.5 x_{\text{Facebook}}$
- $f(\mathbf{x}|\mathbf{w}) = 0.5 x_{\text{lesson}}$
- $f(\mathbf{x}|\mathbf{w}) = 0.2 x_{\text{lesson}} + 0.3 x_{\text{study}} - 0.01 x_{\text{Facebook}}$

criterion to choose model parameters = mean error on predicted grades /
error rate to predict academic success / % of errors smaller than 2/20 / ...

Example of Supervised Learning Problem

Prediction of student grades with a linear model

- 1 you collect $x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}}$ and t_{grade} for each student
features target
- 2 you assume that $f(x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}} | \underbrace{w_{\text{lesson}}, w_{\text{study}}, w_{\text{Facebook}}}_{\text{parameters}}) =$
 $\underbrace{w_{\text{lesson}} x_{\text{lesson}} + w_{\text{study}} x_{\text{study}} + w_{\text{Facebook}} x_{\text{Facebook}}}_{\text{model prediction = weighted sum of feature values}}$ can model students

Possible models

- $f(\mathbf{x}|\mathbf{w}) = 0.5 x_{\text{Facebook}}$
- $f(\mathbf{x}|\mathbf{w}) = 0.5 x_{\text{lesson}}$
- $f(\mathbf{x}|\mathbf{w}) = 0.2 x_{\text{lesson}} + 0.3 x_{\text{study}} - 0.01 x_{\text{Facebook}}$

criterion to choose model parameters = mean error on predicted grades /
error rate to predict academic success / % of errors smaller than 2/20 / ...

Example of Supervised Learning Problem

Prediction of student grades with a linear model

- 1 you collect $\underbrace{x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}}}_{\text{features}}$ and $\underbrace{t_{\text{grade}}}_{\text{target}}$ for each student
- 2 you assume that $f(x_{\text{lesson}}, x_{\text{study}}, x_{\text{Facebook}} | \underbrace{w_{\text{lesson}}, w_{\text{study}}, w_{\text{Facebook}}}_{\text{parameters}}) =$
 $\underbrace{w_{\text{lesson}} x_{\text{lesson}} + w_{\text{study}} x_{\text{study}} + w_{\text{Facebook}} x_{\text{Facebook}}}_{\text{model prediction = weighted sum of feature values}}$ can model students

Possible models

- $f(\mathbf{x}|\mathbf{w}) = 0.5 x_{\text{Facebook}}$
- $f(\mathbf{x}|\mathbf{w}) = 0.5 x_{\text{lesson}}$
- $f(\mathbf{x}|\mathbf{w}) = 0.2 x_{\text{lesson}} + 0.3 x_{\text{study}} - 0.01 x_{\text{Facebook}}$

criterion to choose model parameters = mean error on predicted grades /
error rate to predict academic success / % of errors smaller than 2/20 / ...

Common error criteria

in regression, mean square error for model f =

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - t_i)^2$$



in classification, error/misclassification rate for model f =

$$\text{error rate} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[f(\mathbf{x}_i) \neq t_i]$$



Outline of the Next Lessons

Basic algorithms

- k -nearest neighbours for classification and regression
- decision trees (ID3 algorithm) for classification



Linear models and friends

- linear models for classification and regression
- single-layer and multi-layer artificial neural networks
- kernel machines (kernelised ridge regression and SVMs)

Other models

- ensemble methods (random forests, AdaBoost)
- advanced probabilistic models (HMMs, GPs, etc.)

What is generalisation?

Predictive Models and Generalisation

Naive algorithm for machine learning

- learning: store target t_i for training instance x_i (e.g. in a huge table)
- prediction: retrieve target t for new instance x (e.g. from the table)
- what if x has not been previously seen (likely if many features)?
- what if we are facing exabytes of data (e.g. Google, Facebook, etc.)?

Definition of generalisation

ability of a model to make prediction for new, unseen objects

Inductive learning hypothesis (we need good and useful models)

if a model approximates the target process well over a large set of training instances, it will also approximate well over new, unobserved instances

Predictive Models and Generalisation

Naive algorithm for machine learning

- learning: store target t_i for training instance x_i (e.g. in a huge table)
- prediction: retrieve target t for new instance x (e.g. from the table)
- what if x has not been previously seen (likely if many features)?
- what if we are facing exabytes of data (e.g. Google, Facebook, etc.)?

Definition of generalisation

ability of a model to make prediction for new, unseen objects

Inductive learning hypothesis (we need good and useful models)

if a model approximates the target process well over a large set of training instances, it will also approximate well over new, unobserved instances

Predictive Models and Generalisation

Naive algorithm for machine learning

- learning: store target t_i for training instance x_i (e.g. in a huge table)
- prediction: retrieve target t for new instance x (e.g. from the table)
- what if x has not been previously seen (likely if many features)?
- what if we are facing exabytes of data (e.g. Google, Facebook, etc.)?

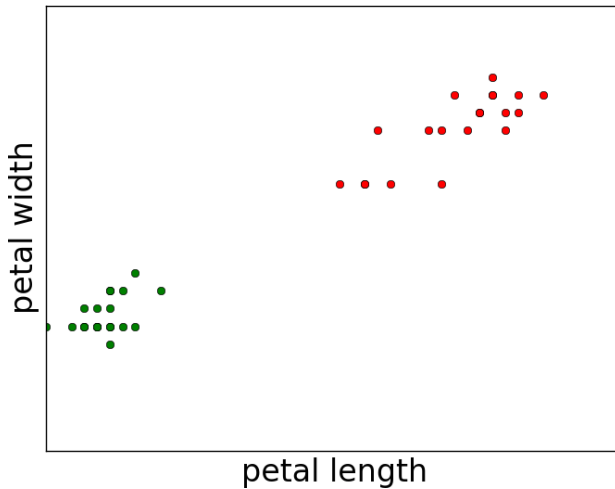
Definition of generalisation

ability of a model to make prediction for new, unseen objects

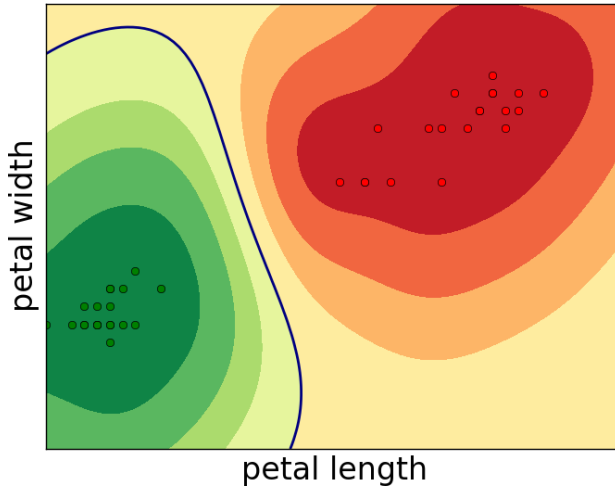
Inductive learning hypothesis (we need good and useful models)

if a model approximates the target process well over a large set of training instances, it will also approximate well over new, unobserved instances

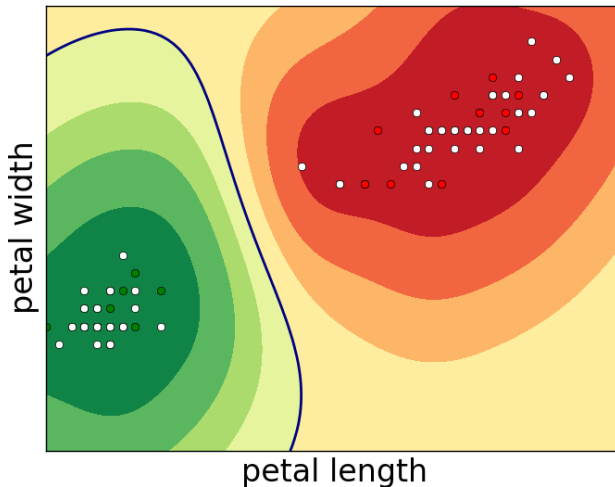
Examples of Classification Model



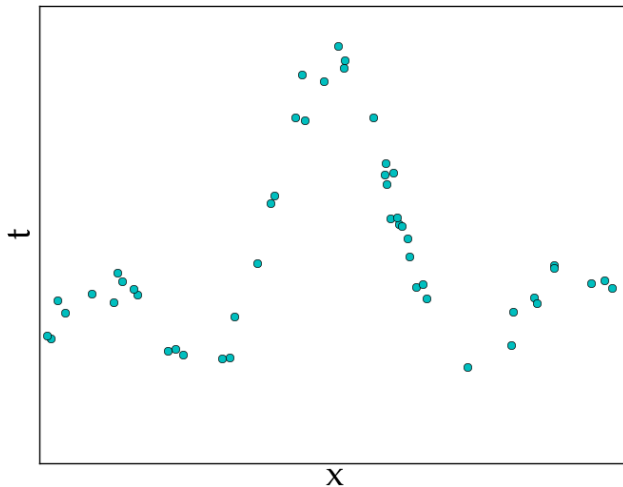
Examples of Classification Model



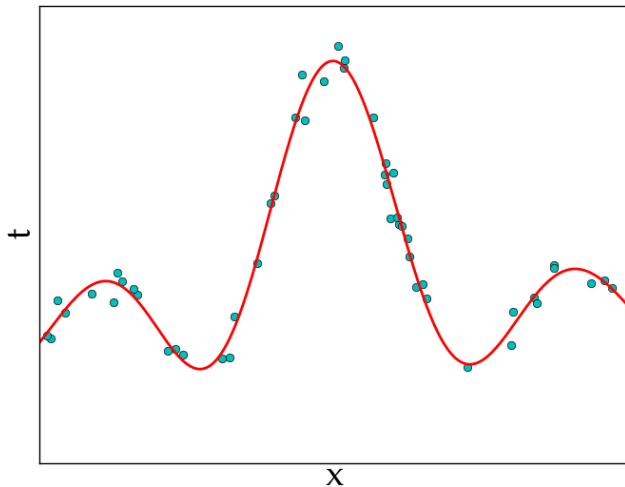
Examples of Classification Model



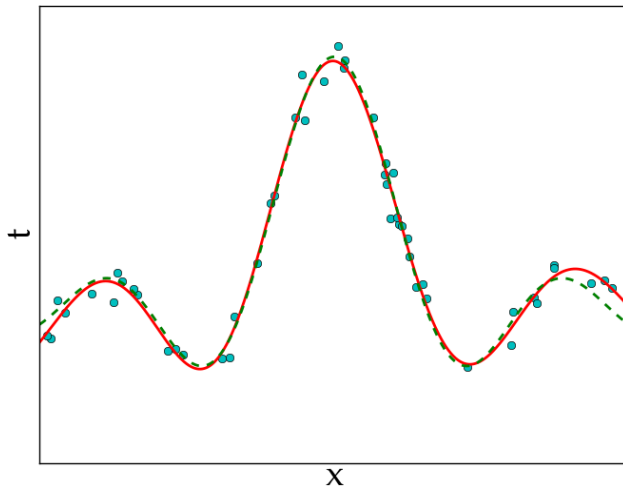
Examples of Regression Model



Examples of Regression Model



Examples of Regression Model



How do we Assess Generalisation?

Training vs. testing performances

models should always be assessed on new, unseen instances

- training instances have already been "seen" during learning → cannot be used to assess generalisation (memorised / learnt by heart)
- test instances = independent instances never "seen" during learning

assessment in ML is probabilistic (no specification / tests available)

Common criteria in supervised learning

in regression, the average square error for model f is used

$$\frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - t_i)^2$$

in classification, the misclassification rate is used (% correct predictions)

Implications of the Inductive Learning Hypothesis

Definition of the learning bias

in order to generalise, we need hypotheses about the generating process

- the learning bias is the set of hypotheses which allow generalisation
- it is impossible to generalise without any learning bias
- the learning bias is often implicit (and hard to notice)


Examples of learning bias

- training pairs (\mathbf{x}_i, t_i) are i.i.d. (independent and identically distributed)
- the process generating data is stable over time
- the function to be predicted is continuous
- the relationship between \mathbf{x} and t is linear
- the noise over the target value t is Gaussian

Importance of the Prior Knowledge

Definition of prior knowledge

knowledge about the process which can be obtained from experts

- allows choosing a set of suitable learning biases
- examples: type of noise in data (Gaussian vs. Poisson), proportionality assumptions (linear vs. logarithmic), known useful/useless features, prevalence of classes, ageing of sensors, unreliability of labels (crowd) 

Beware: choose the right bias

failure to use the right bias for your data can have disastrous consequences

- http://archive.wired.com/techbiz/it/magazine/17-03/wp_quant

Outline of this Lesson

- what is supervised learning
- what is generalisation

References

