02450: Introduction to Machine Learning and Data Mining

Decision trees and linear regression

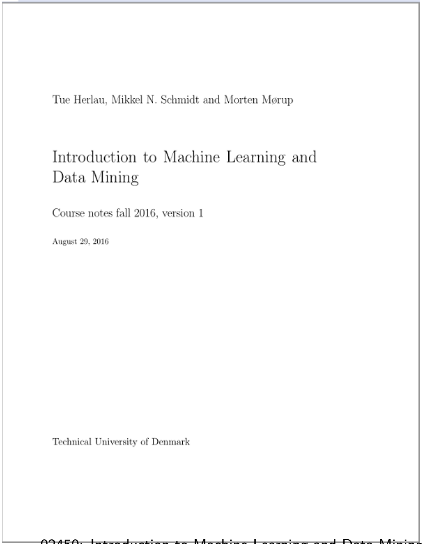
DTU Compute
Department of Applied Mathematics and Computer Science

Reading Material

Reading material:
C7, C8

Feedback Groups of the day:

- Albert Juhl, Mathias Henriksen
- Marianne Helenius, Niklas Refsgaard, Martin Haubro
- Martin Petersson, Christoffer Jensen
- Mads Okholm Bjørn, Rasmus Liebst, Johan Bloch Madsen
- Line Maj Thomsen, Johan Lassen
- Ramiro Mata, Bianca Burger, Marsela Fallah
- Martin Simon, Péter Semság
- Thomas Masquart, Julien Hoareau


Tue Herlau, Mikkel N. Schmidt and Morten Mørup

Introduction to Machine Learning and Data Mining

Course notes fall 2016, version 1

August 29, 2016

Technical University of Denmark

Lecture Schedule

1. Introduction
   30 August: C1

Data: Feature extraction, and visualization

2. Data and feature extraction
   6 September: C2, C3
3. Measures of similarity and summary statistics
   13 September: C4
4. Data Visualization and probability
   20 September: C5, C6

Supervised learning: Classification and regression

5. **Decision trees and linear regression**
   27 September: C7, C8 **(Project 1 due before 13:00)**
6. Overfitting and performance evaluation
   4 October: C9
7. Nearest Neighbor, Bayes and Naive Bayes
   11 October: TBA
8. Artificial Neural Networks and Bias/Variance
   25 October: TBA
9. AUC, ensemble methods and multi-class classifiers
   1 November: TBA

Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    8 November: TBA **(Project 2 due before 13:00)**
11. Mixture models and association mining
    15 November: TBA
12. Density estimation and anomaly detection
    22 November: TBA

Recap

13. Recap and discussion of the exam
    29 November: TBA **(Project 3 due before 13:00)**

Probabilities (revisited from last week)

- Basic rules of probability
  - Sum rule

$$p(x) = \sum_y p(x, y)$$

  - Product rule

$$p(x, y) = p(x|y)p(y)$$

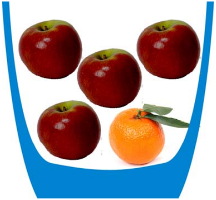
  - Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

	2	5	7
	4	1	5
	6	6	12



Apple taken from: https://upload-wikimedia.org/wikipedia/commons/3/32/Dark_apple.png
Orange (clementine) taken from: https://commons.wikimedia.org/wiki/File:Clementine_orange.jpg

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

$$= \frac{5/7 \cdot 7/12}{6/12} = 5/6$$

Apple taken from: https://upload-wikimedia.org/wikipedia/commons/3/32/Dark_apple.png
Orange (clementine) taken from: https://commons.wikimedia.org/wiki/File:Clementine_orange.jpg

The news paper “Economist”

News media agency Reuters Bureau sends stories to the Economist:

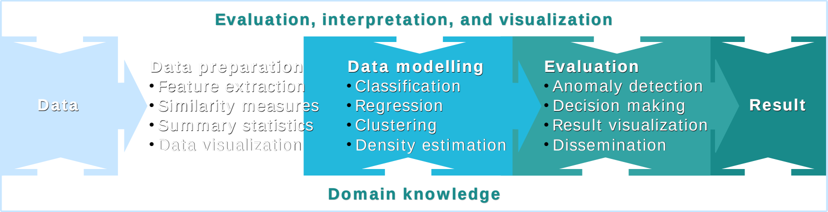
- 80% of the news stories from Reuters are positive and 20% of the news stories are negative.
- 90% of the negative news stories are published in the Economist while only 5% of the positive stories are published.

Consider a story from Reuters. What is the probability it is positive given it is published in the Economist?

Hints:

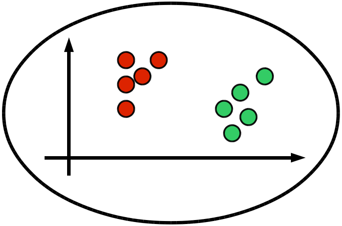
- *Sum Rule* $$p(x) = \sum_y p(x,y)$$
- *Product rule* $$p(x,y) = p(x|y)p(y)$$
- *Bayes theorem* $$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Data modeling framework



After today you should be able to:
Explain what supervised learning is
Explain the difference between classification and regression
Be able to evaluate classifiers in terms of the confusion matrix, error rate and accuracy
Understand the principles behind decision trees and Hunt's algorithm
Apply and interpret decision trees, linear regression and logistic regression

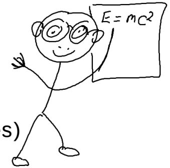
Supervised and Unsupervised learning

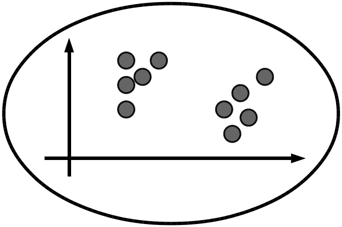


Supervised Learning

Input data $\mathbf{x}_n$ and output y_n

(Generalize from known examples)



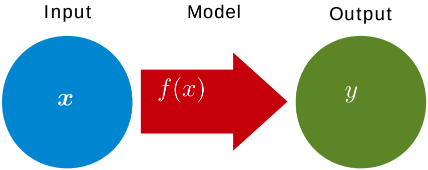


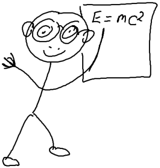
Unsupervised Learning

Input data $\mathbf{x}_n$ alone

(Exploratory analysis)

Supervised learning





- **Data**
  - Inputs and outputs *(this is what we are given)*

$$\{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$$

- **Model**
  - Function that maps inputs to outputs *(what we are trying to determine)*

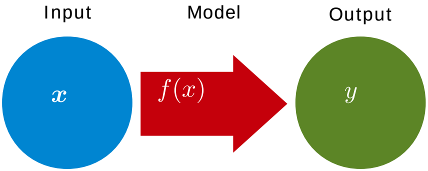
$$f(\boldsymbol{x})$$

- **Cost function**
  - Dissimilarity measure between observation and prediction *(how we tell if a model is good or bad)*

$$d\big(y, f(\boldsymbol{x})\big)$$

- **Types of supervised learning**
  - Regression: Continuous output **y**
  - Classification: Discrete output **y**

Give an example of a classification and a regression problem and explain what the model *f(x)* can be used for.



- **Mapping between domains**
  - Classification: Discrete (nominal) output
  - Regression: Continuous output

Classification

- **Definition:** Learning a function that maps a data object to a discrete class

- **Why classify?**
  - Descriptive modeling
    - Explain / understand the relation between attributes and class
  - Predictive modeling
    - Predict the class of a new data object

Confusion matrix

- Visualization of actual versus predicted class labels

- **Accuracy**

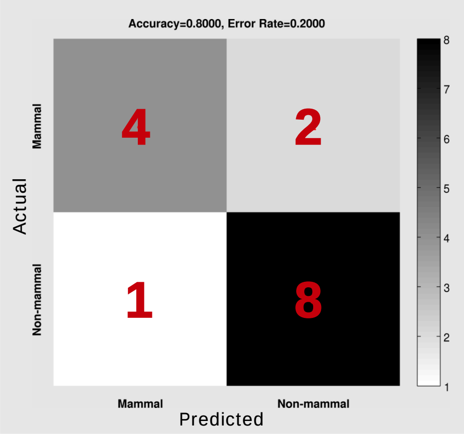
  (Number of correctly predicted observations divided by the total number of observations)

$$\frac{4+8}{4+2+1+8} = 80\%$$

- **Error rate**

  (Number of in-correctly predicted observations divided by the total number of observations)

$$\frac{2+1}{4+2+1+8} = 20\%$$



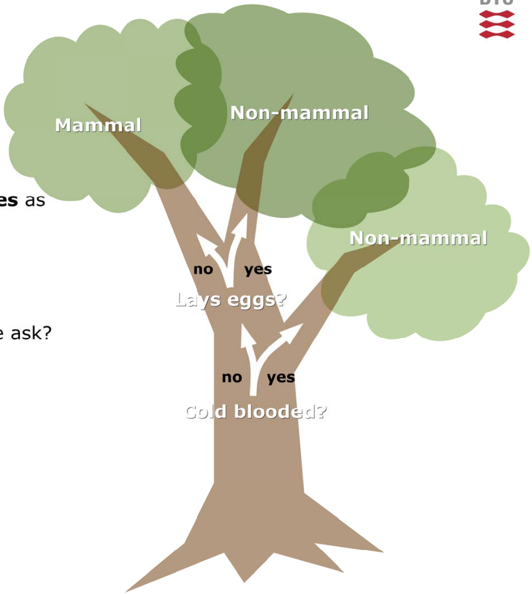
Decision trees

- Remember the game “20 questions to the professor”? (see also www.20q.new)

Q1. Is it an Animal? Yes.
Q2. Can you hold it? No.
Q3. Does it live in groups (gregarious)? Yes.
Q4. Are there many different sorts of it? No.
Q5. Can it jump? Yes.
Q6. Does it eat seeds? No.
Q7. Is it white? Sometimes.
Q8. Is it black and white? No.
Q9. Does it have paws? Yes.
Q10. Can you see it in a zoo? Yes.
Q11. Does it roar? Yes.
Q12. Is it worth a lot of money? Yes.
Q13. Does it have spots? Yes.
Q14. Is it multicoloured? Yes.
Q15. Can you make money by selling it? Yes.
Q16. Does it live in the jungle? Yes.
Q17. I guessed that it was a leopard? Wrong.
Q18. Does it like to play? Yes.
Q19. I guessed that it was a cheetah? Wrong.
Q20. I am guessing that it is a siberian tiger? Correct.

Decision trees

- Ask a series of questions until a conclusion is reached
- **Example**: Classify **vertebrates** as
  - **Mammal** or
  - **Non-mammal**
- **Learning task**
  - Which questions should we ask?



Hunts algorithm

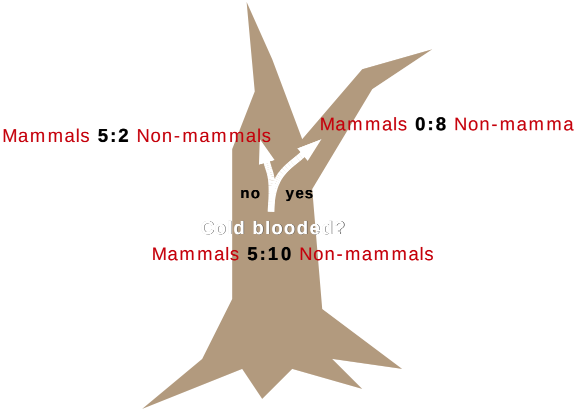
- Assign all data objects to the root

Hunts algorithm

- Select an attribute test condition
  - Find a good question to ask

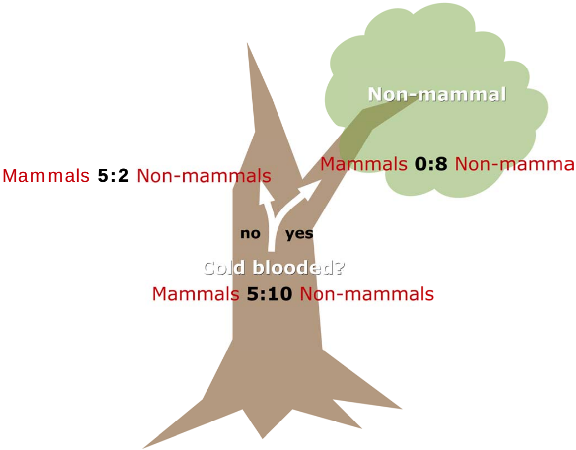
Hunt's Algorithm

- Partition the data objects into subsets according to the test condition



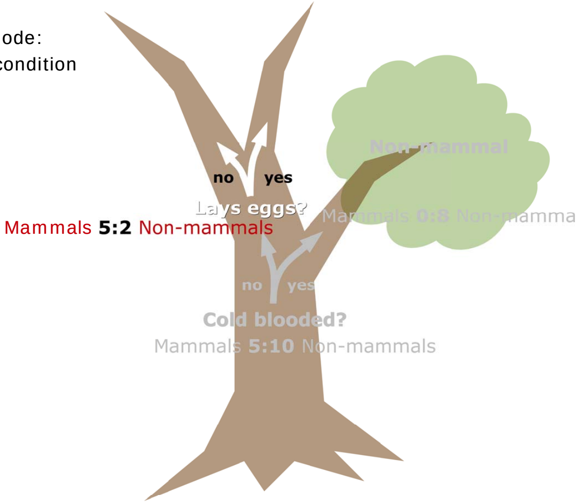
Hunts algorithm

- If all data objects belong to the same class
  - Create a leaf node



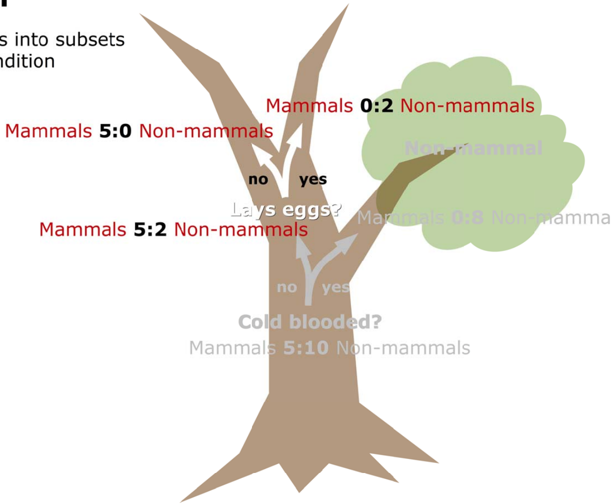
Hunts algorithm

- Repeat for each non-leave node:
  - Select an attribute test condition



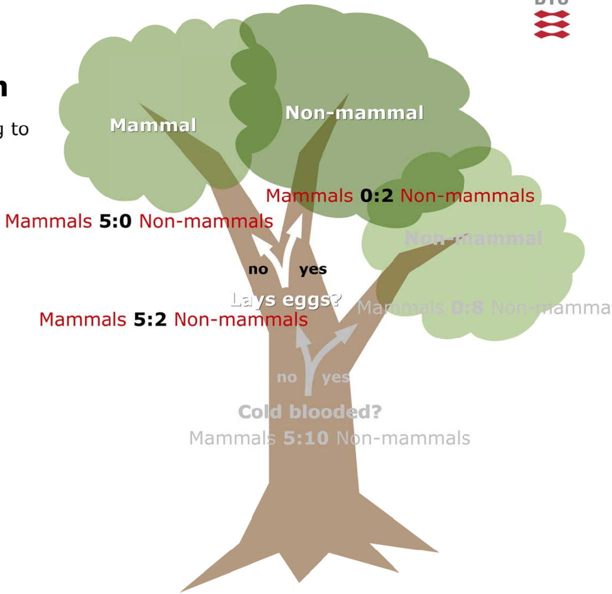
Hunts algorithm

- Partition the data objects into subsets according to the test condition



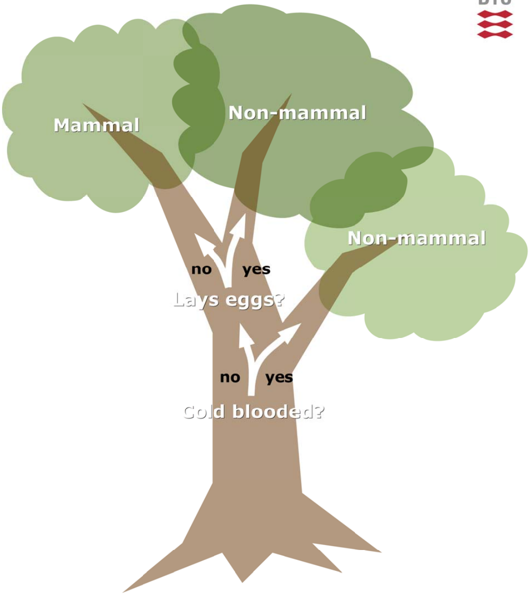
Hunts algorithm

- If all data objects belong to the same class
  - Create a leaf node



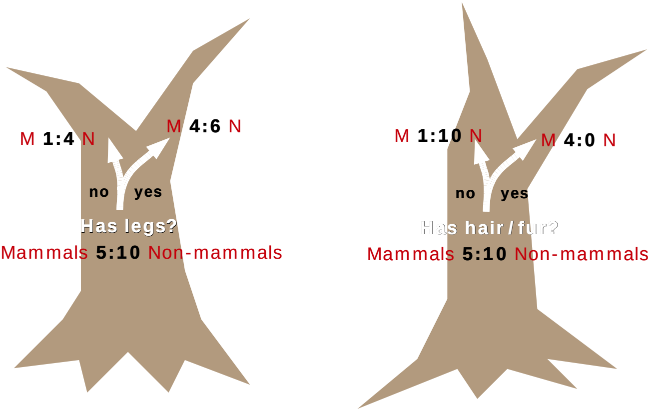
Hunts algorithm

- But how do we find the **best question** at each step?



Selecting the best split

- Which of these two questions is best and why?



Selecting the best split

- Consider a large number of possible splits
- Compute a measure of impurity after the proposed split
  - For each new branch of the tree
  - Compute weighted average impurity
- Choose split that reduces impurity most

Selecting the best split: Impurity measures

- Compute the purity gain, Δ

$\mathrm{Entropy}(t) = -\sum_{i=0}^{c-1} p(i\|t) log_2 p(i\|t)$	$\Delta = ?$	$\Delta = ?$
$\mathrm{Gini}(t) = 1 - \sum_{i=0}^{c-1} \left(p(i\|t)\right)^2$	$\Delta = ?$	$\Delta = ?$
$\mathrm{Class.\ error}(t) = 1 - \max_i p(i\|t)$	$\Delta = ?$	$\Delta = ?$

$p(i|t)$ Fraction of objects that belong to class i

$N(v_j)/N$ Fraction of animals in branch v_j

$$\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

Selecting the best split: Impurity measures

- Compute the purity gain, Δ

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i|t) log_2 p(i|t)$$

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} \left(p(i|t)\right)^2$$

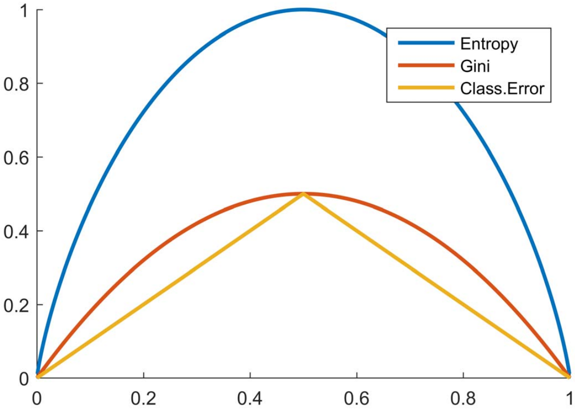
$$\text{Class. error}(t) = 1 - \max_i p(i|t)$$

$p(i\|t)$	Fraction of objects that belong to class i
$N(v_j)/N$	Fraction of animals in branch v_j

$$\Delta = I(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

	Has legs?	Has hair/fur?
Entropy	I(Parent)=-5/15·log(5/15)-10/15·log(10/15) = 0.9183 I(left) =-1/5·log(1/5)-4/5·log(4/5) = 0.7219 I(right) =-4/10·log(4/10)-6/10·log(6/10) = 0.9710 Δ=0.9183-5/15·0.7219-10/15·0.9710 = 0.0303	I(Parent)=-5/15·log(5/15)-10/15·log(10/15) = 0.9183 I(left) =-1/11·log(1/11)-10/11·log(10/11) = 0.4395 I(right) =-4/4·log(4/4)-0/4·log(0/4) = 0 Δ=0.9183-11/15·0.4395-4/15·0 = 0.5960
Gini	I(Parent)=1-(5/15)²-(10/15)² =0.4444 I(left) =1-(1/5)²-(4/5)² =0.3200 I(right) =1-(4/10)²-(6/10)² =0.4800 Δ=0.4444-5/15·0.3200-10/15·0.4800 =0.0177	I(Parent)=1-(5/15)²-(10/15)² =0.4444 I(left) =1-(1/11)²-(10/11)² =0.1653 I(right) =1-(4/4)²-(0/4)² =0 Δ=0.4444-11/15·0.1653-4/15·0 =0.3232
Class. error	I(Parent)=1-10/15 =5/15 I(left) =1-4/5 =1/5 I(right) =1-6/10 =4/10 Δ =5/15-5/15·1/5-10/15·4/10 =0	I(Parent)=1-10/15 = 5/15 I(left) =1-10/11 = 1/11 I(right) =1-4/4 = 0 Δ = 5/15-11/15·1/11-4/15·0 = 0.2667

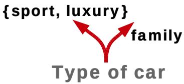
For a two class problem



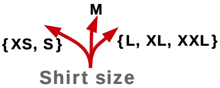
Which splits to consider

- Binary

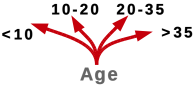
- Nominal

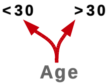


- Ordinal

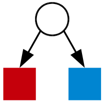


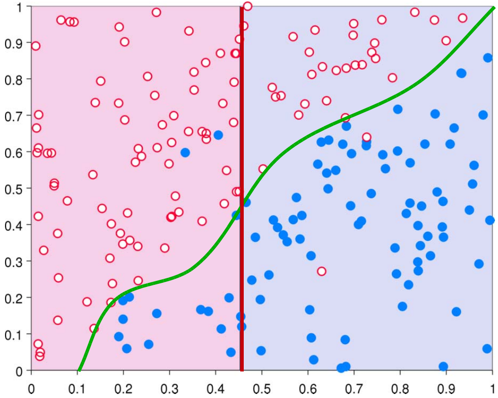
- Continuous



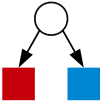


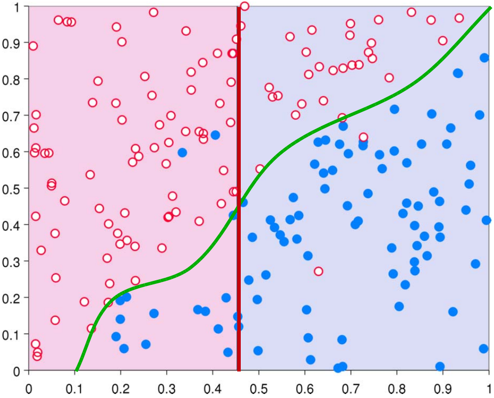
Classification Trees



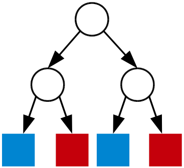


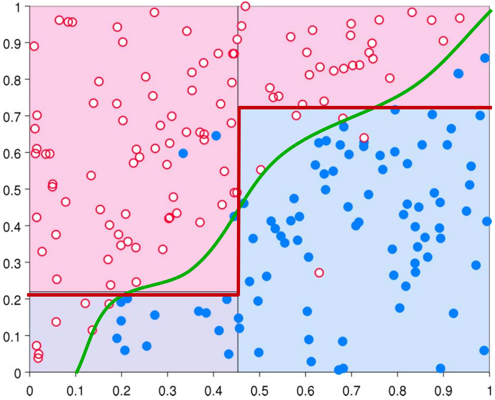
Classification Trees



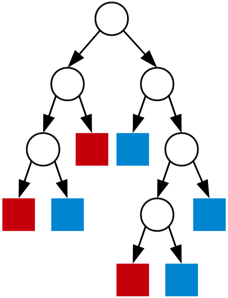


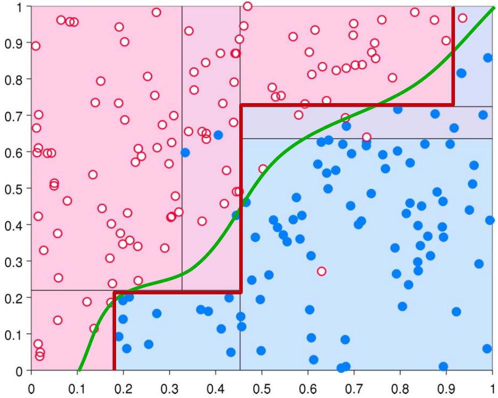
Classification trees





Classification trees



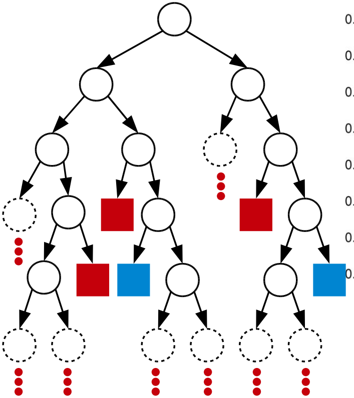


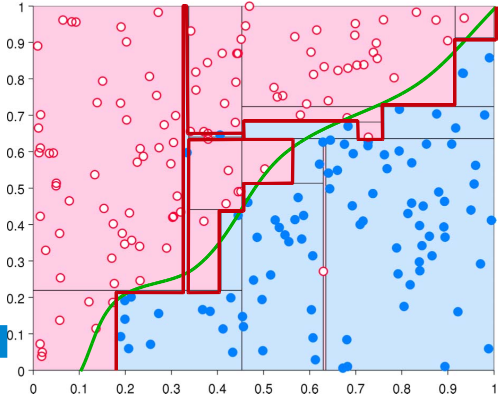
Classification trees

<u>Common stopping criteria:</u>

All records have the same class label

The number of observations have fallen below some minimum treshold





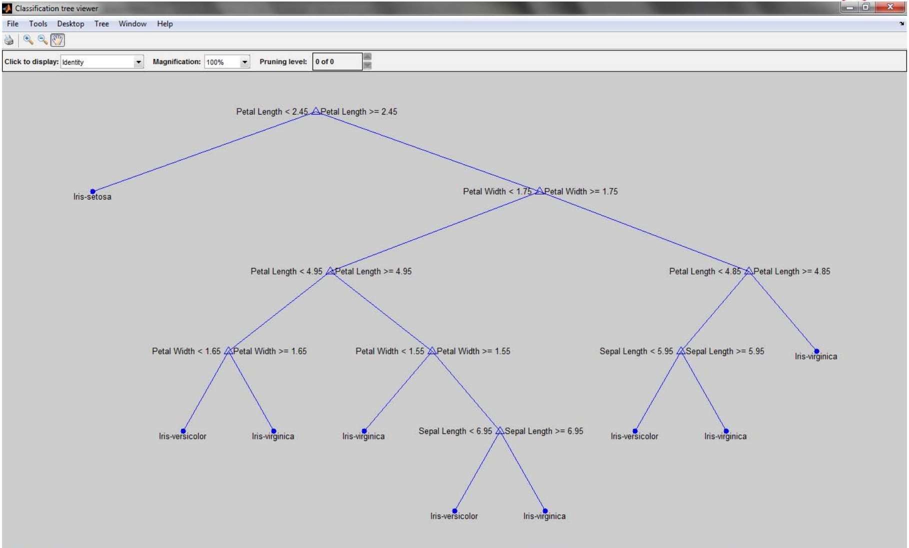
The iris data set

- **Three flowers**
  - 50 instances of each class, 150 in total

- **Attributes**
  - Sepal (outermost leaves)
    - length in cm
    - width in cm
  - Petal (innermost leaves)
    - length in cm
    - width in cm
  - Class of flower
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

Fower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	5.9	3.0	5.1	1.8

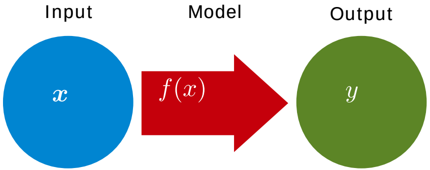
$X^{\text{Observation x Attribute}}$



What would the following iris flower be classsified as?

Sepal Length	Sepal Width	Petal Length	Petal Width
4.0	3.5	3.0	2.0

Supervised learning



- **Mapping between domains**
  - Classification: Discrete (nominal) output
  - Regression: Continuous output

Supervised learning

- **Data**
  - Inputs and outputs $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$

- **Model**
  - Function that maps inputs to outputs

$$f(\boldsymbol{x})$$

- **Cost function**
  - Dissimilarity measure between data and model

$$d\big(y, f(\boldsymbol{x})\big)$$

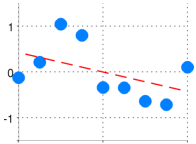
Regression

- **Definition:** Learning a function that maps a data object to a continuous-valued output

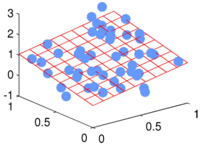
- **Why Regression?**
  - Descriptive modeling
    - Explain / understand the relation between attributes and continous-valued output
  - Predictive modeling
    - Predict the output value of a new data object

Linear regression

- 1-dimensional inputs
  $f(x) = w_0 + w_1 x$



- 2-dimensional inputs
  $f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2$



- K-dimensional inputs
  $f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$

Linear regression

- K-dimensional inputs

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

- Non-linearly transformed inputs

$$f(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_K x^K$$

$$f(x) = w_0 + w_1 sin(x) + w_2 cos(x)$$

Linear regression

- K-dimensional inputs

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

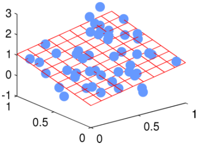
- Non-linearly transformed inputs

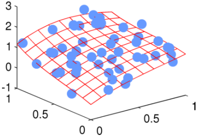
$$f(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_K x^K$$

$$f(x) = w_0 + w_1 sin(x) + w_2 cos(x)$$

- **Example**

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2$$





$$\begin{aligned} f(\boldsymbol{x}) = w_0 &+ w_1 x_1 + w_2 x_2 \\ &+ w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 \\ &+ w_6 x_1^3 + w_7 x_1^2 x_2 + w_8 x_1 x_2^2 + w_9 x_2^3 \end{aligned}$$

Vector notation

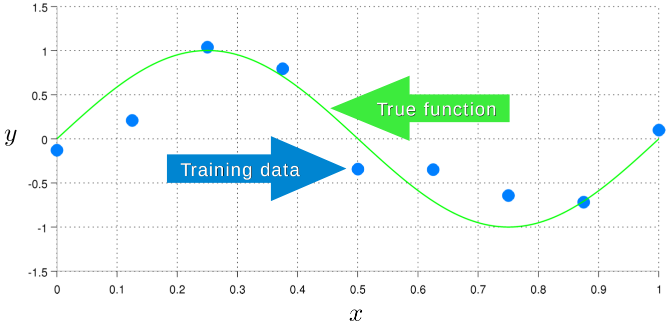
- The linear model can be written compactly using vector notation

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

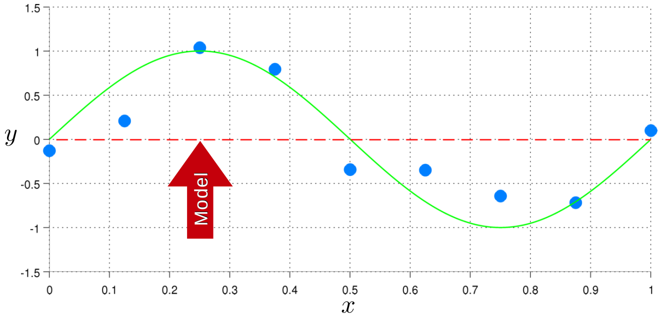
$$= \sum_{k=0}^{K} w_k x_k = \boxed{\boldsymbol{x}^\top \boldsymbol{w}}$$

  – where $x_0 = 1$

Linear regresssion



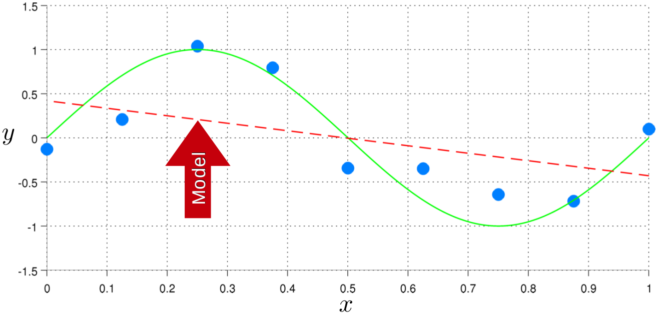
Linear regresssion



Model

$$f(x) = w_0$$

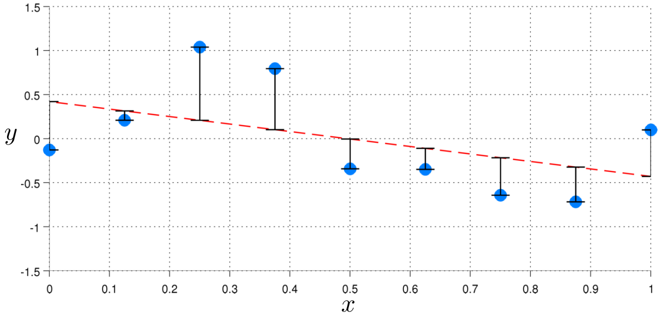
Linear regresssion



Model

$$f(x) = w_0 + w_1 x$$

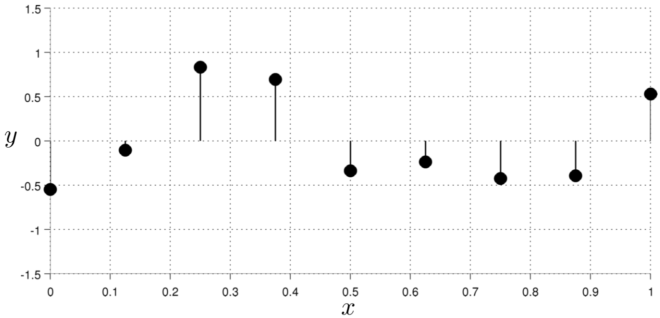
Residual error



Model

$$f(x) = w_0 + w_1 x$$

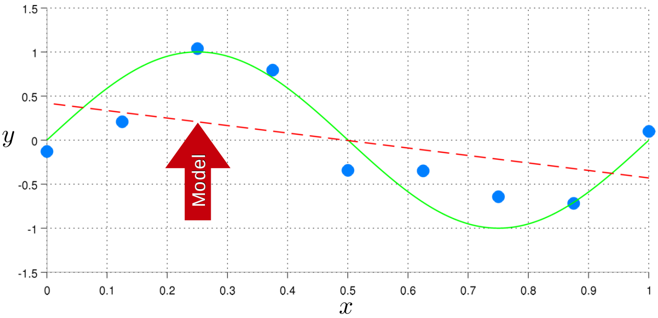
Residual error



Model

$$f(x) = w_0 + w_1 x$$

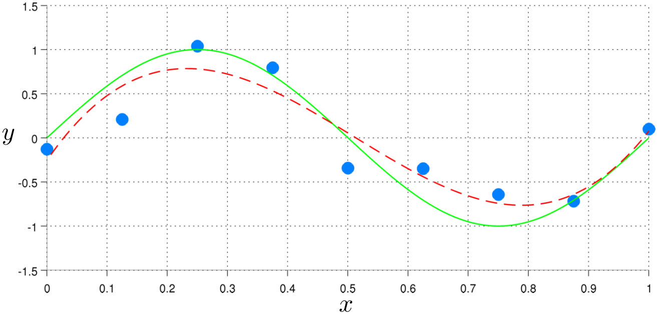
Linear regression



Model

$$f(x) = w_0 + w_1 x$$

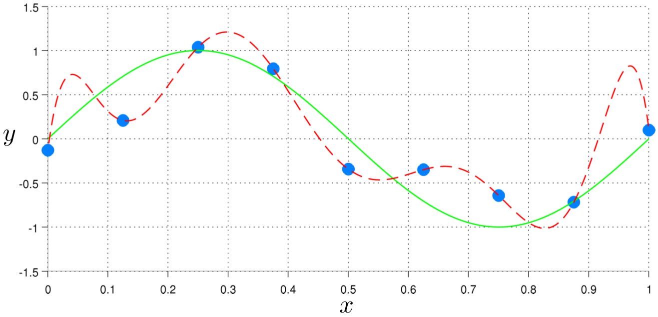
Linear regression



Model

$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

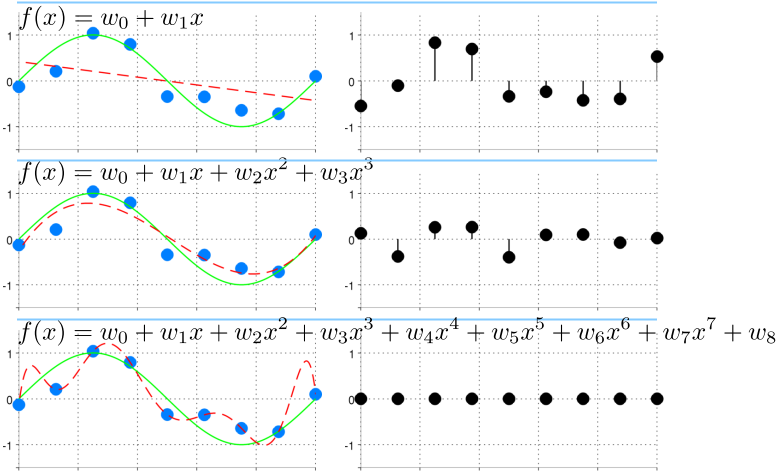
Linear regression



Model

$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + w_6 x^6 + w_7 x^7 + w_8 x^8$$

Model order



- Which model order
  - Gives the best fit?
  - Do you think is most “correct”?

Estimating parameters

- **How do we compute the parameters?**
  - Most simple approach: **Minimize cost function over data set**
  - **Data** $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$
  - **Model** $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{w}$
  - **Cost function** $d\big(y, f(\boldsymbol{x})\big)$
  - **Parameters** $\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^N d\big(y_n, f(\boldsymbol{x}_n)\big)$

Least Squares Regression

- **Cost function:** Squared error

$$d\big(y, f(\boldsymbol{x})\big) = \big(y - f(\boldsymbol{x})\big)^2$$

Model: Linear regression

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{w}$$

- **Parameters**

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} d\big(y_n, f(\boldsymbol{x}_n)\big) = \arg\min_{\boldsymbol{w}} \underbrace{\sum_{n=1}^{N} (y_n - \boldsymbol{x}_n^\top \boldsymbol{w})^2}_{E = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2}$$

$$\frac{\partial E}{\partial \boldsymbol{w}} = 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top \boldsymbol{X} = 0$$

$$\Rightarrow 2\boldsymbol{y}^\top \boldsymbol{X} = 2\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}$$

$$\Rightarrow \boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Logistic Regression
(for binary classification, y$\in\{0,1\}$)

- **Cost function:**
  negative log of the Bernoulli distribution

$$d\big(y, f(\boldsymbol{x})\big) = -y \log f(\boldsymbol{x}) - (1-y)\log(1 - f(\boldsymbol{x}))$$

Model: Logistic link function

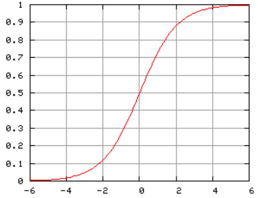
$$f(\boldsymbol{x}) = g_{\text{logistic}}(\boldsymbol{x}^\top \boldsymbol{w})$$

- **Parameters**

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} d\big(y_n, f(\boldsymbol{x}_n)\big)$$

$$g_{\text{logistic}}(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

$g_{\text{logistic}}(z)$



z

Interpretation of f(x): The probability that the observation belongs to class 1

Generalized linear model

- **Cost function:** Choose one

$$d\big(y, f(\boldsymbol{x})\big)$$

Model: Linear + non-linear link

$$f(\boldsymbol{x}) = g_{\text{link}}(\boldsymbol{x}^\top \boldsymbol{w})$$

- **Parameters:** Optimize using numerical optimization methods

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} d\big(y_n, f(\boldsymbol{x}_n)\big)$$

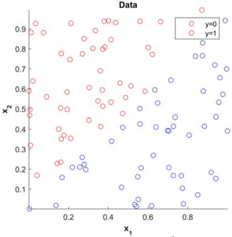
Optimize

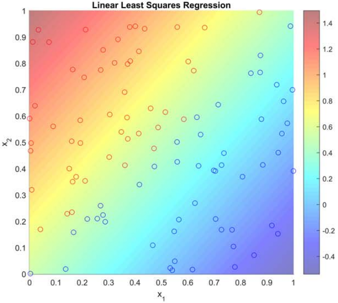
Matlab: glmfit
Python: sklearn.linear_model
R: glm

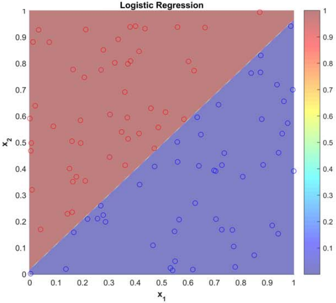
$$g_{\text{identity}}(z) = z$$

$$g_{\text{logistic}}(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

Linear vs. Logistic Regression for a classification problem







$f(x)=0.5270-1.0639x_1+0.9684x_2$

$f(x)=1/[1+exp(-(-41.6-2295x_1+2422x_2))]$