02450: Introduction to Machine Learning and Data Mining

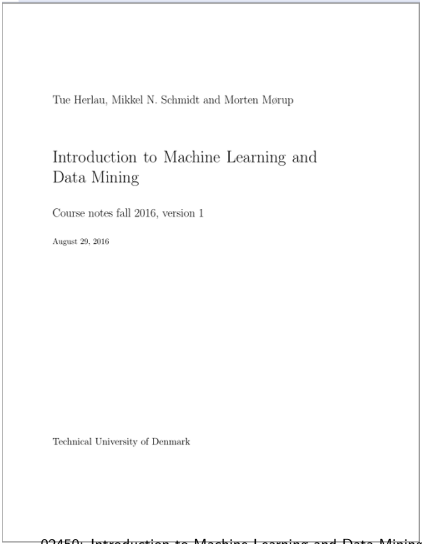
Nearest Neighbor, Bayes and Naive Bayes

DTU Compute
Department of Applied Mathematics and Computer Science

Reading Material

Reading material:
C10, C11
Feedback Groups of the day:

- Mariana Mesquita da Cunha, Daniel Molina
- Anders Holmgaard Opstrup, Huayu Zheng, Gu Jinshan
- Anders Verner Nielsen, Simon Nexø Jensen, Casper Thorø Vium Pedersen
- Zivile Vajegaite, Quoc Tien AU, Federico Romano
- Sai Tejaa Chintaluri, Guillem Anton Aguilà Calbet
- Pau Oliver, Laurens Devos, Yevgen Zainchkovskyy
- Morten Telling, Tobias Lindstrøm, Marcus Pagh


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: C10, C11
8. Artificial Neural Networks and Bias/Variance
   25 October: C12, C13
9. AUC and ensemble methods
   1 November: C14, C15

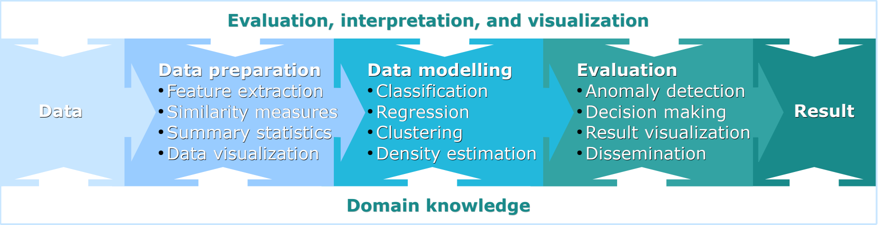
Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    8 November: C16 **(Project 2 due before 13:00)**
11. Mixture models and association mining
    15 November: C17, C18
12. Density estimation and anomaly detection
    22 November: C19

Recap

13. Recap and discussion of the exam
    29 November: C1-C19 **(Project 3 due before 13:00)**

Data modeling framework

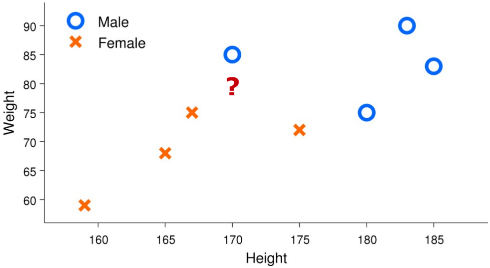


After today you should be able to:

Explain how K-Nearest Neighbors can be used to classify data
Account for the assumptions made in Naïve Bayes
Apply Bayes theorem to obtain the class posterior likelihood
Understand how to interpret Bayesian Belief Networks.

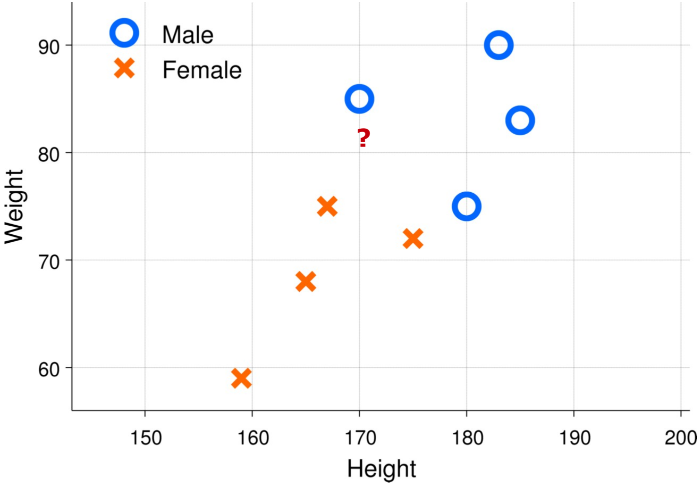
Classify gender based on height and weight

	Height	Weight	Gender
1	183	90	Male
2	180	75	Male
3	170	85	Male
4	185	83	Male
5	159	59	Female
6	167	75	Female
7	165	68	Female
8	175	72	Female
9	171	82	?



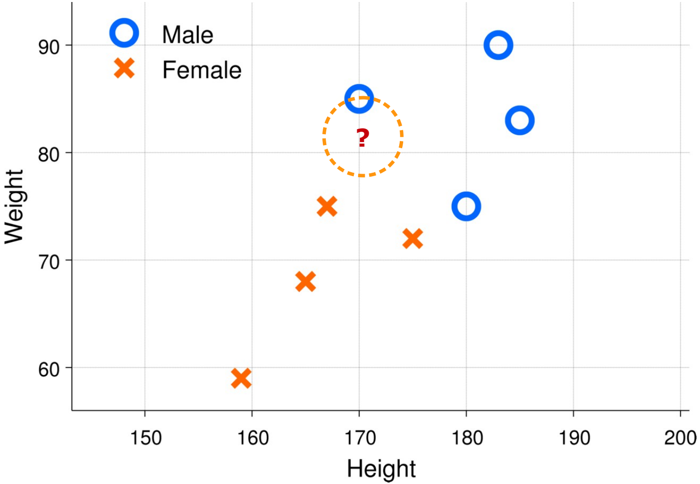
Nearest neighbor classifier

- 1 nearest neighbor



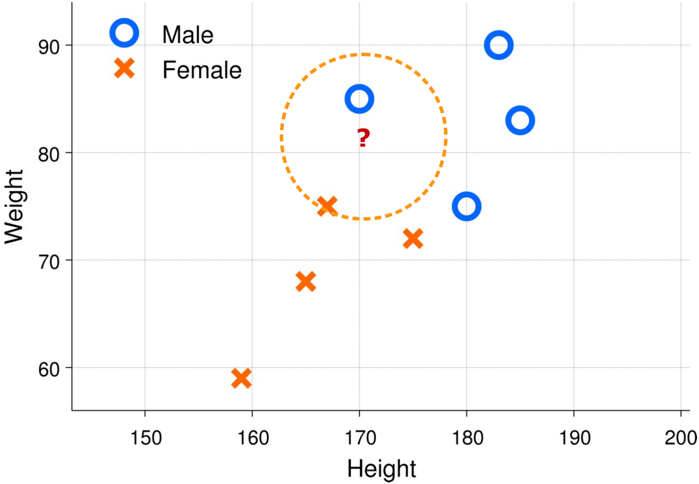
Nearest neighbor classifier

- 1 nearest neighbor



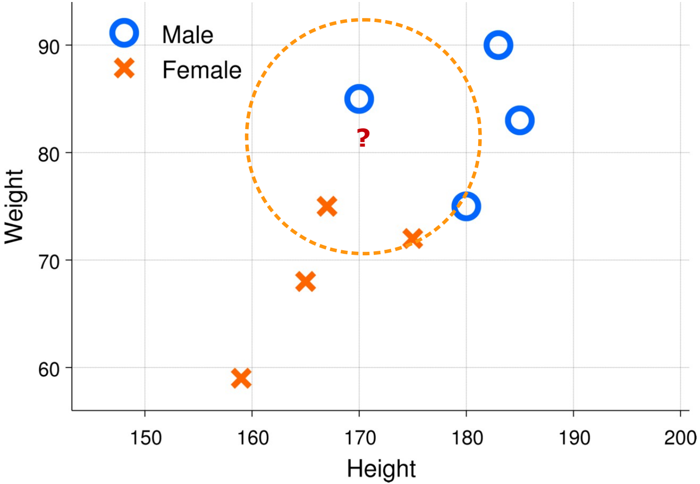
Nearest neighbor classifier

- 2 nearest neighbors



Nearest neighbor classifier

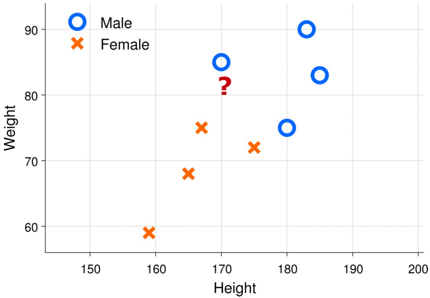
- 3 nearest neighbors



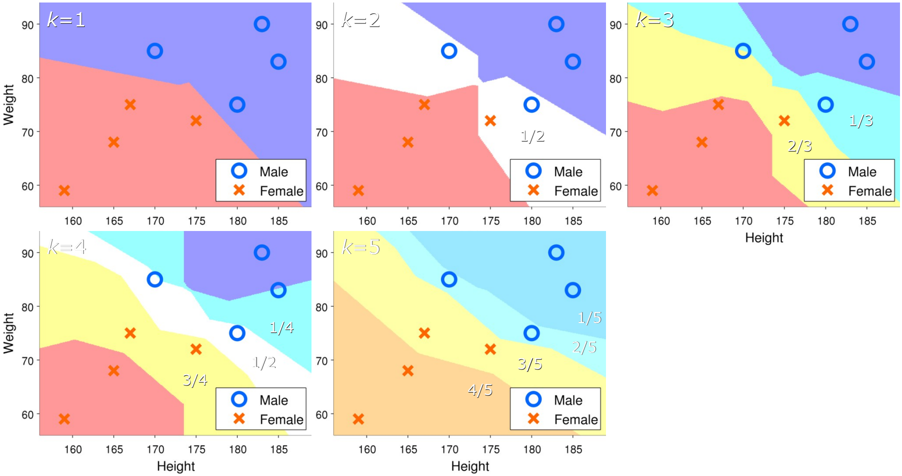
Nearest neighbor classifier

- Choose
  - The number of neighbors, *k*
  - A distance measure

1. Compute distance to all other data objects
2. Find the *k* nearest data objects
3. Classify according to majority of neighbors



Nearest neighbor decision surface

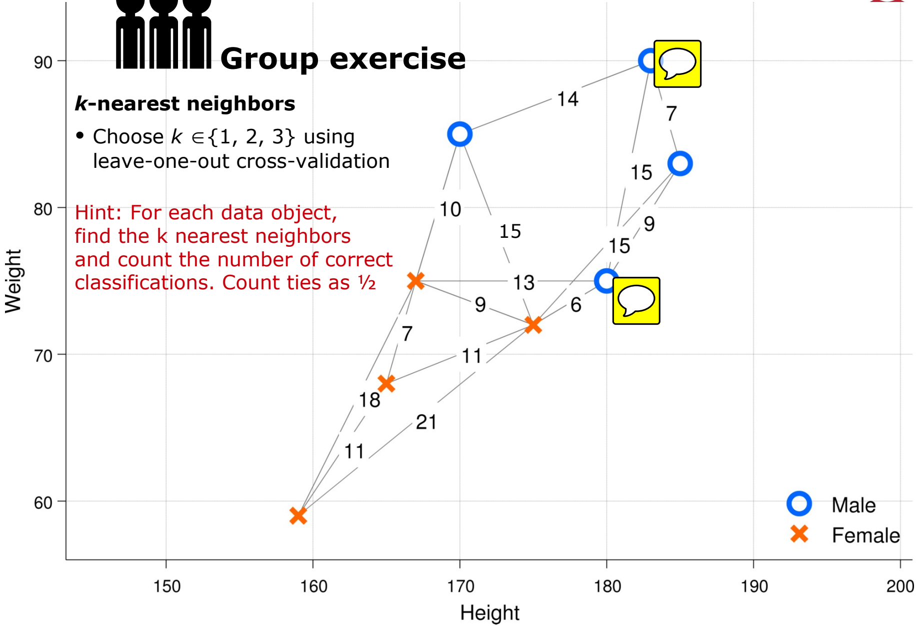


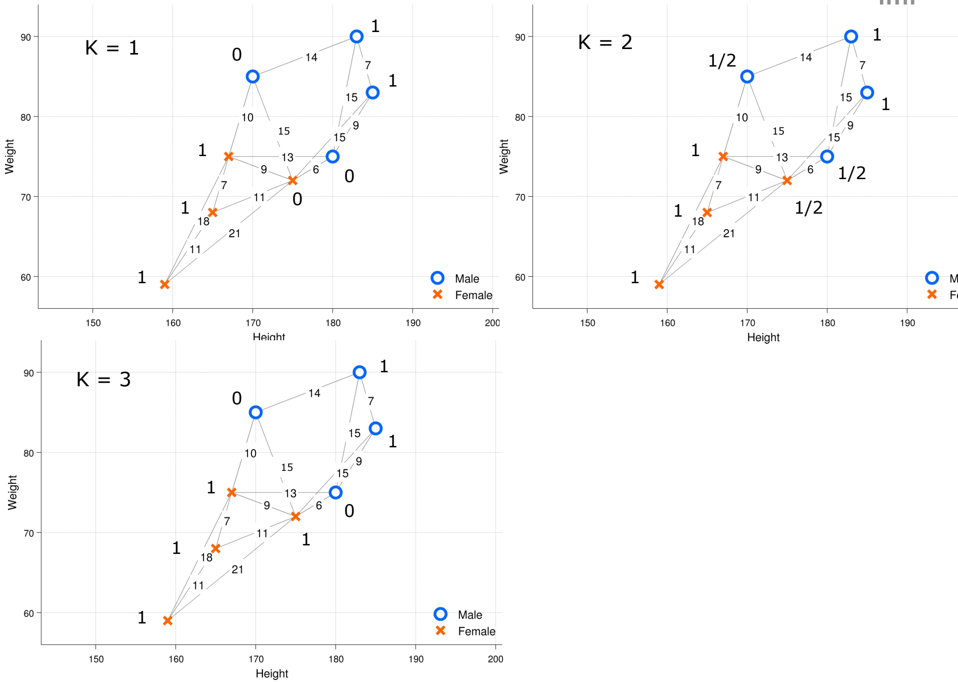
Group exercise

***k*-nearest neighbors**

- Choose $k \in \{1, 2, 3\}$ using leave-one-out cross-validation

Hint: For each data object, find the k nearest neighbors and count the number of correct classifications. Count ties as ½

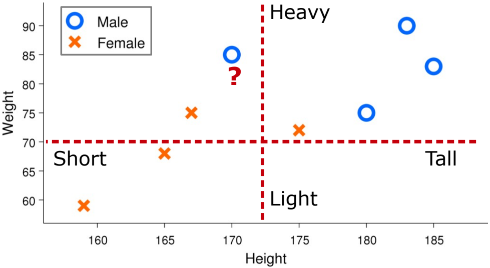




K = 1
K = 2
K = 3
0
1
1/2
Weight
Height
Male
Female
90
80
70
60
150
160
170
180
190
200
14
7
15
10
9
13
6
11
18
21

Bayesian classifiers

	Height	Weight	Gender
1	Tall	Heavy	Male
2	Tall	Heavy	Male
3	Short	Heavy	Male
4	Tall	Heavy	Male
5	Short	Light	Female
6	Short	Heavy	Female
7	Short	Light	Female
8	Tall	Light	Female
9	Short	Heavy	L[illegible]



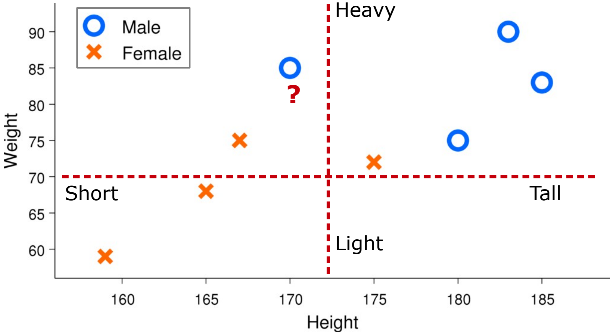
Bayesian classifiers

- What is the probability that **?** is male

$$p(\text{Gender} = \text{Male}|\text{Height} = \text{Short}, \text{Weight} = \text{Heavy})$$

- Shorthand notation:

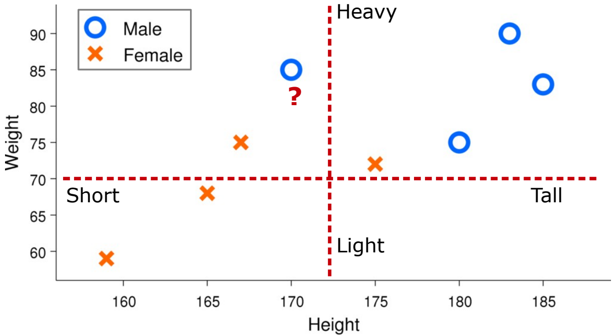
$$p(G = m|H = s, W = h) = p(m|s, h)$$



Bayesian classifiers

- Bayes rule

$$p(m|s,h) = \frac{p(s,h|m)p(m)}{\sum_{G\in\{m,f\}} p(s,h|G)p(G)} = \frac{\frac{1}{4}\cdot\frac{4}{8}}{\frac{1}{4}\cdot\frac{4}{8}+\frac{1}{4}\cdot\frac{4}{8}} = \frac{1}{2}$$



Bayesian classifiers

- **Contingency table**
  - All combinations of attribute values
  - Huge table

	Height	Weight	Gender
1	Tall	Heavy	Male
2	Tall	Heavy	Male
3	Short	Heavy	Male
4	Tall	Heavy	Male
5	Short	Light	Female
6	Short	Heavy	Female
7	Short	Light	Female
8	Tall	Light	Female

Gender	Height	Weight	Fraction
Male	Short	Light	0/4
		Heavy	1/4
	Tall	Light	0/4
		Heavy	3/4
Female	Short	Light	2/4
		Heavy	1/4
	Tall	Light	0/4
		Heavy	1/4

Bayesian classifiers

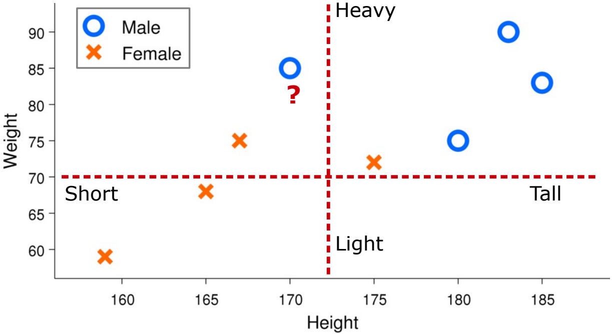
- Naïve Bayes assumption
  - Conditional probabilities of attributes are independent

$$p(\text{Height}, \text{Weight}|\text{Gender}) = p(\text{Height}|\text{Gender}) \times p(\text{Weight}|\text{Gender})$$

Bayesian classifiers

- **Naïve Bayes classifier**

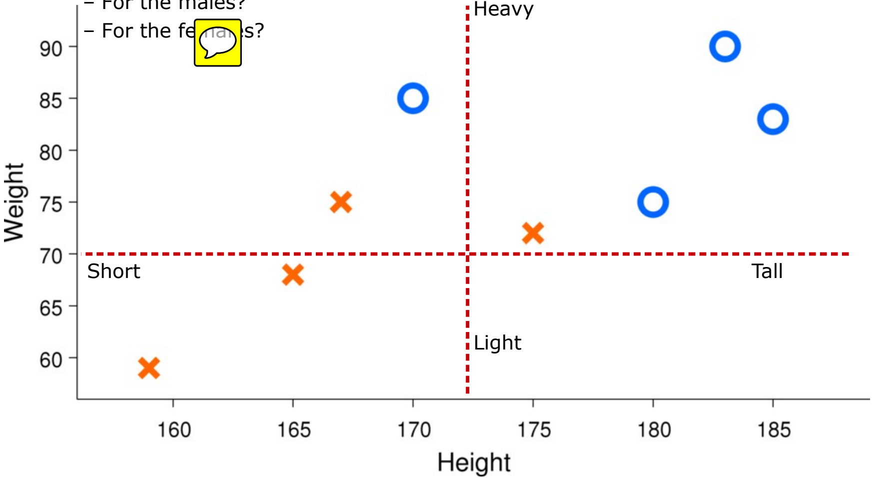
$$p(m|s,h) = \frac{p(s|m)p(h|m)p(m)}{\sum_{G\in\{m,f\}} p(s|G)p(h|G)p(G)} = \frac{\frac{1}{4}\cdot\frac{4}{4}\cdot\frac{4}{8}}{\frac{1}{4}\cdot\frac{4}{4}\cdot\frac{4}{8}+\frac{3}{4}\cdot\frac{2}{4}\cdot\frac{4}{8}} = \frac{2}{5}$$



Naïve Bayes assumption
$$p(\text{Height}, \text{Weight}|\text{Gender}) = p(\text{Height}|\text{Gender}) \times p(\text{Weight}|\text{Gender})$$

Group exercise

- Does the naïve Bayes assumption hold empirically
  - For the males?
  - For the females?

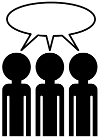


Bayesian classifiers

- **Naïve Bayes contingency table**
  - Only counts for each attribute
  - Small table

	Height	Weight	Gender
1	Tall	Heavy	Male
2	Tall	Heavy	Male
3	Short	Heavy	Male
4	Tall	Heavy	Male
5	Short	Light	Female
6	Short	Heavy	Female
7	Short	Light	Female
8	Tall	Light	Female

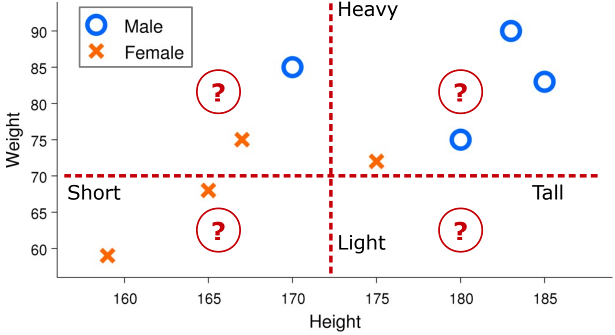
Gender	Attribute	Fraction
Male	Height=Short	1/4
	Weight=Light	0/4
Female	Height=Short	3/4
	Weight=Light	2/4



Group exercise

Bayes classifiers

- Classify (compute the posterior probability of G=m) for the four **?** using
  - Bayes classifier $$p(m|s,h) = \frac{p(s,h|m)p(m)}{\sum_{G\in\{m,f\}} p(s,h|G)p(G)} = \frac{\frac{1}{4}\cdot\frac{4}{8}}{\frac{1}{4}\cdot\frac{4}{8}+\frac{1}{4}\cdot\frac{4}{8}} = \frac{1}{2}$$
  - Naïve Bayes classifier $$p(m|s,h) = \frac{p(s|m)p(h|m)p(m)}{\sum_{G\in\{m,f\}} p(s|G)p(h|G)p(G)} = \frac{\frac{1}{4}\cdot\frac{4}{4}\cdot\frac{4}{8}}{\frac{1}{4}\cdot\frac{4}{4}\cdot\frac{4}{8}+\frac{3}{4}\cdot\frac{2}{4}\cdot\frac{4}{8}} = \frac{2}{5}$$



Gender	Attribute	Fraction
Male	Height = Short	1/4
	Weight = Light	0/4
Female	Height = Short	3/4
	Weight = Light	2/4

Gender	Height	Weight	Fraction
Male	Short	Light	0/4
		Heavy	1/4
	Tall	Light	0/4
		Heavy	3/4
Female	Short	Light	2/4
		Heavy	1/4
	Tall	Light	0/4
		Heavy	1/4

Robust estimation

- Probability of y given x for discrete variables

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

Number of objects having value *y and x*

Total number of objects that have value x

  - Not defined when *n*=0

- M-estimate

$$p(y|x) = \frac{n_c + m_c}{n + m}$$

Pseudo observations of objects having value y and x

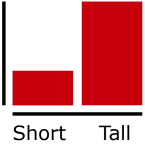
Equivalent pseudo-sample size of objects having value x

  - If no objects take value *x* the probability will be $\frac{m_c}{m}$
  - Corresponds to putting *m* extra objects into the data set

Bayesian classifiers

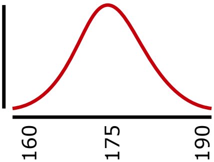
$$p(\text{Height}|\text{Gender} = \text{Male})$$

- Handling continuous attributes
  - Two way split (x<a)
    - Converts into binary attribute
      (We have used this in the previous example)



  - Discretize into a number of bins
    - Converts into discrete ordinal attribute

  - Probability density estimation
    - Assume attribute follows a Normal distribution
    - Use data to compute parameters
      (mean and variance)

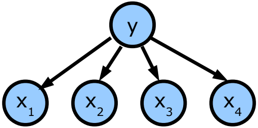


Baysian Belief Networks (BBN)

- Independence assumption may not hold for some attributes (use BBN)

Naïve Bayes

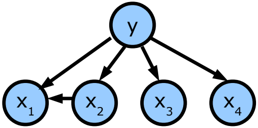
$p(\mathbf{x}|y)=p(x_1|y)p(x_2|y)p(x_3|y)p(x_4|y)$



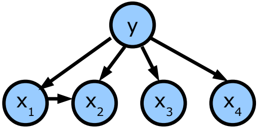
$p(x_1|y)$ $p(x_2|y)$ $p(x_3|y)$ $p(x_4|y)$

When x_1 and x_2 are not independent given y

$$p(\mathbf{x}|y) = p(x_1,x_2|y)p(x_3|y)p(x_4|y)$$
$$= p(x_1|x_2,y)p(x_2|y)p(x_3|y)p(x_4|y)$$



$$= p(x_2|x_1,y)p(x_1|y)p(x_3|y)p(x_4|y)$$



Remember basic rules of probability

– Sum rule

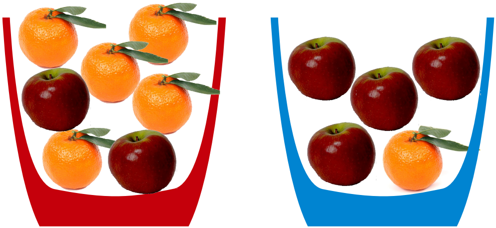
$$p(x) = \sum_y p(x, y)$$

– Product rule

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

– Bayes' rule

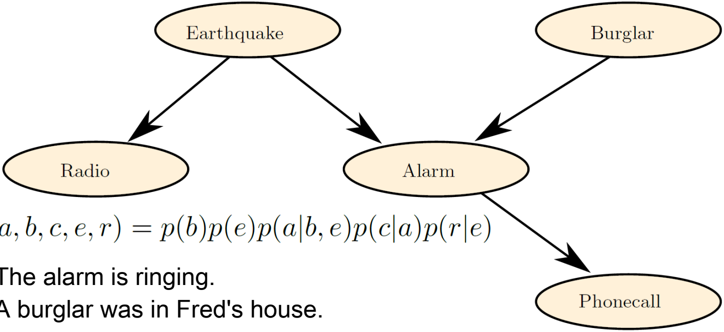
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$



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

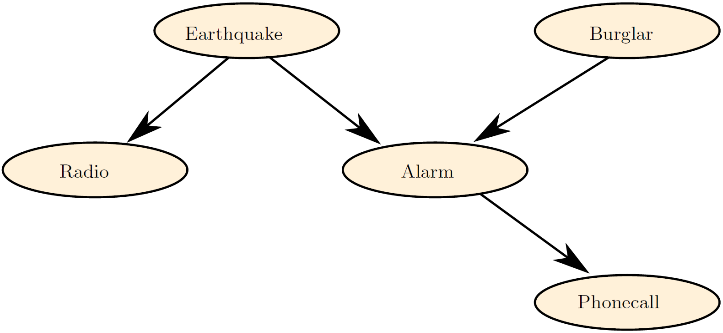
Exampel taken from:
Information Theory, Inference, and Learning Algorithms, by David J. C. MacKay (chapter 21)
http://www.inference.phy.cam.ac.uk/itprnn/book.pdf, originally proposed by Judea Pearl 1988.

"Fred lives in Los Angeles and commutes 60 miles to work. Whilst at work, he receives a phone-call from his neighbour saying that Freds burglar alarm is ringing. What is the probability that there was a burglar in his house today? While driving home to investigate, Fred hears on the radio that there was a small earthquake that day near his home. Oh, he says, feeling relieved, it was probably the earthquake that set of the alarm. What is the probability that there was a burglar in his house?"



$$p(a,b,c,e,r) = p(b)p(e)p(a|b,e)p(c|a)p(r|e)$$

a : The alarm is ringing.
b : A burglar was in Fred's house.
c : Fred received a phone-call reporting the alarm.
e : A small earthquake took place today near Fred's house.
r : The radio report of the earthquake is heard by Fred.

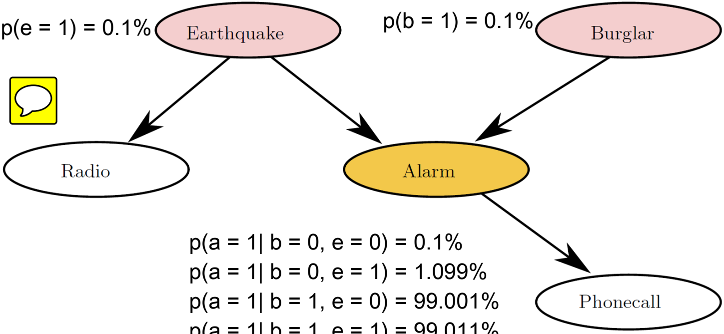


$$p(a=1) = \sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} \sum_{e\in\{0,1\}} \sum_{r\in\{0,1\}} p(a=1,b,c,e,r)$$

$$p(a=1) = \sum_{b\in\{0,1\}} \sum_{c\in\{0,1\}} \sum_{e\in\{0,1\}} p(b)p(e)p(a=1|b,e)p(c|a=1)p(r|e)$$

$$\sum_{b\in\{0,1\}} \sum_{e\in\{0,1\}} \left[p(b)p(e)p(a=1|b,e) \left(\sum_{c\in\{0,1\}} p(c|a=1) \sum_{r\in\{0,1\}} p(r|e) \right) \right]$$

$$= \sum_{b\in\{0,1\}} \sum_{e\in\{0,1\}} p(b)p(e)p(a=1|b,e)$$



$p(e = 1) = 0.1\%$

$p(b = 1) = 0.1\%$

$p(a = 1 | b = 0, e = 0) = 0.1\%$
$p(a = 1 | b = 0, e = 1) = 1.099\%$
$p(a = 1 | b = 1, e = 0) = 99.001\%$
$p(a = 1 | b = 1, e = 1) = 99.011\%$

What is $p(a=1)$?
What is $p(b=0|a=1)$?
What is $p(b=0|e=1,a=1)$?

Hints:

Sum rule: $p(x) = \sum_y p(x, y)$

Product rule: $p(x, y) = p(x|y)p(y)$

Bayes' rule: $p(x|y) = \dfrac{p(y|x)p(x)}{p(y)}$

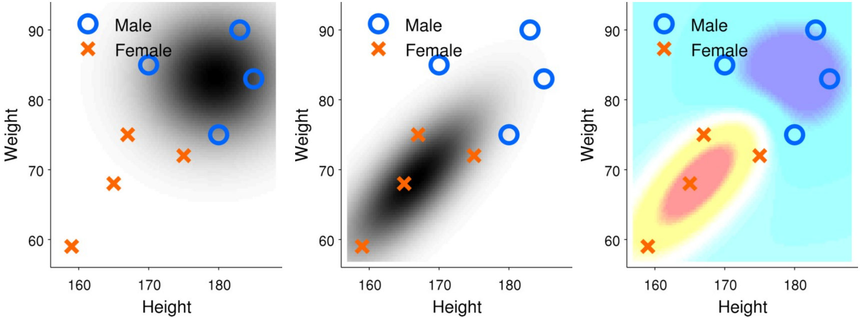
Bayesian classification by the multivariate normal distribution

Continuous density estimation

- Fit a Normal distribution to each class
  - Compute class mean and covariance
- Classify using Bayes rule as before

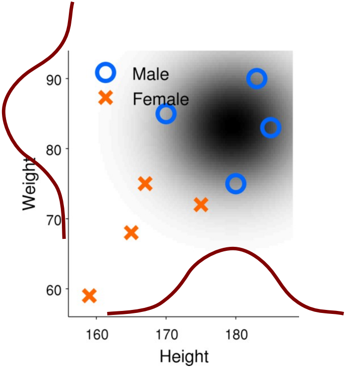
$$P(\boldsymbol{x}|y=c) = \frac{1}{(2\pi)^{M/2} det(\boldsymbol{\Sigma}_c)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c (\boldsymbol{x}-\boldsymbol{\mu}_c)\right)$$

$$P(y=c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y=c)P(y=c)}{\sum_{c'} P(\boldsymbol{x}|y=c')P(y=c')}$$



Group exercise

- What does the Naive Bayes assumption of independence of the attributes correspond to in terms of the parameters of the multivariate normal distribution?



Midterm practice test

The midterm practice test is used solely for you to test your knowledge and for me to see how well you have understood the covered material so far.

The test **does not** count towards your grade for this course.