NAÏVE BAYES CLASSIFIER AND FEATURE DISCRETIZATION

Kristjan Pilt, PhD

BAYES EQUATION SIMPLIFICATIONS

- Bayes' equation: $P(c_n|X) = \frac{P(c_n)\cdot P(X|c_n)}{\sum_{k=1}^{K} P(c_k)\cdot P(X|c_k)}$
- Bayes' equation can be extended to become a naïve Bayes classifier with **two simplifications**:
  1. To use the conditional **independence assumption**. Each feature (attribute) is conditionally independent of every other feature under given a class label c_n.
     - For example ball mass and its density are **dependent**
     - For example ball colour and size are **independent**
     - Therefore the calculation of $P(X|c_n)$ for number of features M can be carried out as follows:

$$P(x_1, x_2, \ldots, x_M|c_n) = P(x_1|c_n) \cdot P(x_2|c_n) \cdot \ldots \cdot P(x_M|c_n) = \prod_{m=1}^{M} P(x_m|c_n)$$

  2. To **ignore the denominator** $\sum_{k=1}^{K} P(c_k) \cdot P(X|c_k)$, because it appears in the denominator of $P(c_n|X)$ for all values of n. Removing the denominator will have no impact on the relative probability scores and will simplify calculations.

TAL TECH

NAÏVE BAYES CLASSIFIER

- Naïve Bayes classifier:

$$P(\mathrm{c}_n|\mathrm{X}) \propto P(c_n) \cdot \prod_{m=1}^{M} P(x_m|c_n)$$

m – number of feature
n – number of class

- $\propto$ - means proportional!
- The calculated probability can't be taken as a true probability of $P(c_n|X)$. For all n values the calculated probabilities are proportional with eachother and can be **compared**.

- The Naïve Bayes **classifier is trained** using training set by computing the $P(c_n)$ (*prior probability or initial guess*) and $P(x_m|c_n)$ for all possible m and n values (m=1...M and n=1...N).
- In case of **classification**, for each record in the testing set, the Naïve Bayes classifier assigns the classifier label c_n that **maximizes** $P(c_n) \cdot \prod_{m=1}^{M} P(x_m|c_n)$.

TAL TECH

NAÏVE BAYES CLASSIFIER EXAMPLE NR. 1

- Let the **spam filter** contain 3 keywords: 'lottery', 'win', and 'treasure'. The spam filter has to divide all e-mails into 'ham' or 'spam' class using Naive Bayes' classifier.
- Therefore the features are the three keywords (x_1 corresponds to 'lottery', etc.) and the value of each feature can be 1 or 0 (called as **Boolean features** as they can have values 0 or 1).
- The occurence of keywords is assumed to be **independent**.
- Overview of the representative training set:

	Number of e-mails	**Number of e-mails with keywords**					
		lottery (x_1)		win (x_2)		treasure (x_3)	
	K_n	$x_1 = 0$	$x_1 = 1$	$x_2 = 0$	$x_2 = 1$	$x_3 = 0$	$x_3 = 1$
ham (c_1)	990	986	4	915	75	974	16
spam (c_2)	10	2	8	3	7	5	5

TAL TECH

TRAINING OF NAÏVE BAYES CLASSIFIER IN EXAMPLE NR. 1

- Overview of the training set:

	Number of e-mails	Number of e-mails with keywords					
		lottery (x_1)		win (x_2)		treasure (x_3)	
	K_n	$x_1 = 0$	$x_1 = 1$	$x_2 = 0$	$x_2 = 1$	$x_3 = 0$	$x_3 = 1$
ham (c_1)	990	986	4	915	75	974	16
spam (c_2)	10	2	8	3	7	5	5

- As follows the conditional probabilities $P(x_m|c_n)$ and prior probabilities $P(c_n)$ (*initial guess*) have to be calculated:
- $P(x_m = 1|c_n) = \frac{\sum_{k=1}^{K_n} 1_{x_{k,m}=1}}{K_n}$, where k is the number of e-mail (data point) among the total number of e-mails K_n in the class c_n. m-th feature value of e-mail (data point) k is $x_{k,m}$.
- The same applies for x_m=0: $P(x_m = 0|c_n) = \frac{\sum_{k=1}^{K_n} 1_{x_{k,m}=0}}{K_n}$

$$P(x_1 = 1|c_1) = \frac{4}{990}$$

The probability that the e-mail is ham, in case the 'lottery' **is found** from the e-mail.

$$P(x_1 = 0|c_1) = \frac{986}{990}$$

The probability that the e-mail is ham, in case the 'lottery' **is not found** from the e-mail.

TAL TECH

TRAINING OF NAÏVE BAYES CLASSIFIER IN EXAMPLE NR. 1

- Overview of the training set:

	Number of e-mails	Number of e-mails with keywords					
		lottery (x_1)		win (x_2)		treasure (x_3)	
	K_n	$x_1 = 0$	$x_1 = 1$	$x_2 = 0$	$x_2 = 1$	$x_3 = 0$	$x_3 = 1$
ham (c_1)	990	986	4	915	75	974	16
spam (c_2)	10	2	8	3	7	5	5

- The prior probabilities $P(c_n)$ are calculated as follows:

$$P(c_n) = \frac{K_n}{\sum_{n=1}^{N} K_n}$$

where N is the total number of classes.

Therefore: $P(c_1) = \frac{990}{1000}$ $P(c_2) = \frac{10}{1000}$

TAL TECH

TRAINED NAÏVE BAYES CLASSIFIER IN EXAMPLE NR. 1

- Overview of the trained Naïve Bayes Classifier:

		lottery (x_1)		win (x_2)		treasure (x_3)	
	$P(c_n)$	$P(x_1=0\|c_n)$	$P(x_1=1\|c_n)$	$P(x_2=0\|c_n)$	$P(x_2=1\|c_n)$	$P(x_3=0\|c_n)$	$P(x_3=1\|c_n)$
ham (c_1)	0.9900	0.9960	0.0040	0.9242	0.0758	0.9838	0.0162
spam (c_2)	0.0100	0.2000	0.8000	0.3000	0.7000	0.5000	0.5000

- **The e-mail arrived** with words 'win' and 'treasure'. Is it 'spam' or 'ham' e-mail?
- Naïve Bayes classifier assigns the classifier label c_n that **maximizes** $P(c_n) \cdot \prod_{m=1}^{M} P(x_m|c_n)$.

- x_1=0 (word 'lottery' wasn't found), x_2=1 ('win' was found), x_3=1 ('treasure' was found).
- c_1: $P(c_1) \cdot \prod_{m=1}^{M} P(x_m|c_1) = 0.99 \cdot 0.996 \cdot 0.0758 \cdot 0.0162 = 0.0012$
- c_2: $P(c_2) \cdot \prod_{m=1}^{M} P(x_m|c_2) = 0.01 \cdot 0.2 \cdot 0.7 \cdot 0.5 = 0.0007$

TAL TECH

SMALL NUMBERS

- The multiplication of small number may lead to values that become very small in magnitude (close to zero).
- This is the problem of **numerical underflow**, caused by multiplying several probability values that are close to zero.
- **Numerical underflow** is a condition in a computer program where the result of a calculation is a number of smaller absolute value than the computer can actually represent in memory on its central processing unit (CPU).
- In order to avoid that is to compute the **logarithm of the products**:

$$\log_a(b \cdot c) = \log_a b + \log_a c$$

- Therefore, $P(\mathrm{c}_n|\mathrm{X}) \propto \log P(c_n) + \sum_{m=1}^{M} \log P(x_m|c_n)$

- c_1:$\log P(c_1) + \sum_{m=1}^{M} \log P(x_m|c_1) = -0.0044 + (-0.0018) + (-1.1206) + (-1.7915) = -2.9183$
- c_2:$\log P(c_2) + \sum_{m=1}^{M} \log P(x_m|c_2) = -2 + (-0.6990) + (-0.1549) + (-0.3010) = -3.1549$

TAL TECH

NAÏVE BAYES CLASSIFIER EXAMPLE NR. 2

- Consider the case of a bank that wants to market its term products (deposit products) to the appropriate customers. Given the demographics of clients and their reactions to previous campaign phone calls, the bank's goal is to predict which clients would subscribe to a term deposit.
- The dataset includes 2000 instances with following categorical variables: 1. job, 2. marital status, 3. education level, 4. result of previous marketing campaign contact, 5. was the client actually subscribed to the term deposit.
- Number 1 to 4 are the features and number 5 is the outcome.
- Each feature (x_m) has different number of possible values (e.g. *marital* has: divorced, married, and single) .

job (x_1)	subscribed		marital (x_2)	subscribed		education (x_3)	subscribed		poutcome (x_4)	subscribed	
	yes	no		yes	no		yes	no		yes	no
blue-collar	45	390	divorced	41	187	primary	32	303	failure	68	142
management	51	372	married	115	1086	secondary	68	942	other	44	35
technician	34	305	single	55	516	tertiary	55	509	success	32	26
admin.	35	200				unknown	56	35	unknown	67	1586
services	17	151									
retired	2	90									
other	27	281									

TAL TECH

TASK

- **Train classifier** and **estimate**, whether new client will likely to subscribe to the term deposit?
- A new client who has a career in management, is married, holds a secondary degree, and whose outcome of the previous marketing campaign contact was a success.
- Is this client likely to subscribe to the term deposit?

job (x_1)	**subscribed** $P(x_1\|c_1)$ $P(x_1\|c_2)$	**marital (x_2)**	**subscribed** $P(x_2\|c_1)$ $P(x_2\|c_2)$	**education (x_3)**	**subscribed** $P(x_3\|c_1)$ $P(x_3\|c_2)$	**poutcome (x_4)**	**subscribed** $P(x_4\|c_1)$ $P(x_4\|c_2)$	$P(c_1)$	$P(c_2)$
blue-collar		divorced		primary		failure			
management		married		secondary		other			
technician		single		tertiary		success			
admin.				unknown		unknown			
services									
retired									
other									

TAL TECH

TRAINING OF NAÏVE BAYES CLASSIFIER IN EXAMPLE NR. 2

job (x_1)	subscribed yes	subscribed no	marital (x_2)	subscribed yes	subscribed no	education (x_3)	subscribed yes	subscribed no	poutcome (x_4)	subscribed yes	subscribed no
blue-collar	45	390	divorced	41	187	primary	32	303	failure	68	142
management	51	372	married	115	1086	secondary	68	942	other	44	35
technician	34	305	single	55	516	tertiary	55	509	success	32	26
admin.	35	200				unknown	56	35	unknown	67	1586
services	17	151									
retired	2	90									
other	27	281									

- Calculation of $P(x_m|c_n) = \frac{\sum_{k=1}^{K_n} 1_{x_{k,m}}}{K_n}$, for each class and each feature x_m possible value.
- Calculation of $P(c_n) = \frac{K_n}{\sum_{n=1}^{N} K_n}$

  K_n - total number of objects in the class c_n.

  N - total number of classes

TAL TECH

TRAINED NAÏVE BAYES CLASSIFIER IN EXAMPLE NR. 2

- Overview of the trained Naïve Bayes Classifier:

job	subscribed		marital	subscribed		education	subscribed		poutcome	subscribed			
	$P(x_1\|c_1)$	$P(x_1\|c_2)$		$P(x_2\|c_1)$	$P(x_2\|c_2)$		$P(x_3\|c_1)$	$P(x_3\|c_2)$		$P(x_4\|c_1)$	$P(x_4\|c_2)$	$P(c_1)$	$P(c_2)$
blue-collar	0.213	0.218	divorced	0.194	0.105	primary	0.152	0.169	failure	0.322	0.079	0.106	0.895
management	0.242	0.208	married	0.545	0.607	secondary	0.322	0.527	other	0.209	0.02		
technician	0.161	0.17	single	0.261	0.288	tertiary	0.261	0.285	success	0.152	0.015		
admin.	0.166	0.112				unknown	0.265	0.02	unknown	0.318	0.887		
services	0.081	0.084											
retired	0.009	0.05											
other	0.128	0.157											

- A client who has a career in management, is married, holds a secondary degree, and whose outcome of the previous marketing campaign contact was a success. Is this client likely to subscribe to the term deposit?
- c_1: $P(c_1) \cdot \prod_{m=1}^{M} P(x_m|c_1) = 0.106 \cdot 0.242 \cdot 0.545 \cdot 0.322 \cdot 0.152 = 0.00068$
- c_2: $P(c_2) \cdot \prod_{m=1}^{M} P(x_m|c_2) = 0.895 \cdot 0.208 \cdot 0.607 \cdot 0.527 \cdot 0.015 = 0.00089$

TAL TECH

SMOOTHING

- If one of the feature values does not appear with one (or more) of the class within the training set, the corresponding probability $P(x_m|c_n)$ will equal zero.

- E.g. In training set of e-mails the word ‘money’ is detected in ‘spam’, but not in ‘ham’ e-mails.

- Therefore, the multipilcation $P(c_n) \cdot \prod_{m=1}^{M} P(x_m|c_n)$ will become zero in case of classification procedure, regardless of how large some of the conditional probabilities are.

- A **smoothing technique** assigns a small nonzero probability to rare events not included in some of the classes of the training dataset.

TAL TECH

LAPLACE SMOOTHING (ADD-ONE)

- **Laplace smoothing** (or add-one) technique that pretends to see every outcome once more than it actually appears.

- In the calculation of conditional probabilities **one is added** to the counts.

- E.g. From 150 clients 50 subscribe to the term deposit and their marital status:

marital (x_1)	subscribed yes	subscribed no
divorced	0	11
married	32	61
single	18	28

marital	subscribed $P(x_1 \mid c_1)$	subscribed $P(x_1 \mid c_2)$
divorced	0.000	0.110
married	0.640	0.610
single	0.360	0.280

One is added to counts

marital (x_1)	subscribed yes	subscribed no
divorced	1	12
married	33	62
single	19	29

marital	subscribed $P(x_1 \mid c_1)$	subscribed $P(x_1 \mid c_2)$
divorced	0.019	0.117
married	0.623	0.602
single	0.358	0.282

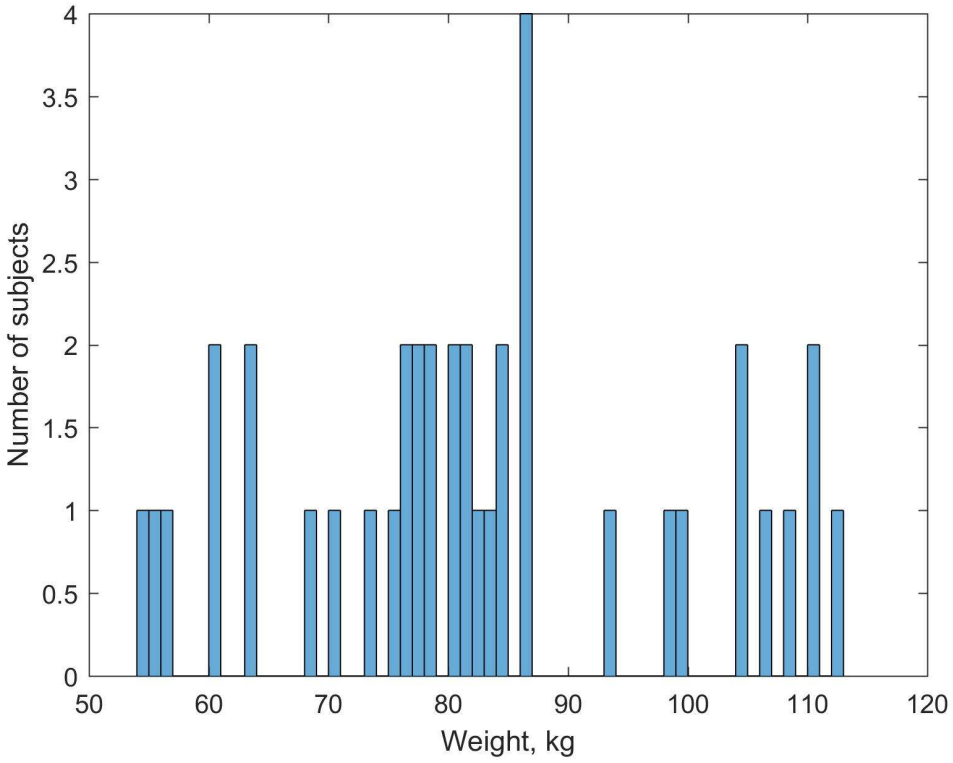
TAL TECH

FEATURES OF NAÏVE BAYES CLASSIFIER

- Weight of 39 healthy subjects between the age of 21 and 71 was measured and the results were given with full kilograms. E.g. Instead of 95.2kg, the weight measurement result is given 95kg.

- Weight is used as one feature to discriminate healthy and unhealthy subjects using Naïve Bayes classifier.

- **How should we calculate conditional probability for 90kg?**



TAL TECH

CONTINUOUS FEATURES

- Naïve Bayes in its simple form is used only with categorical features. It means that the features are discrete.

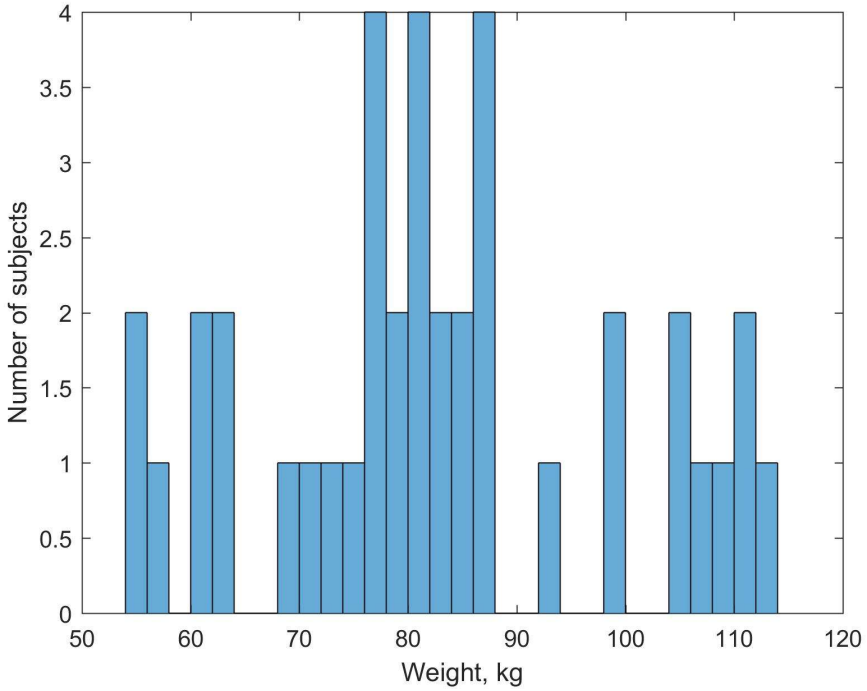
- There are different options how to use continuous features with Naïve Bayes classifier:

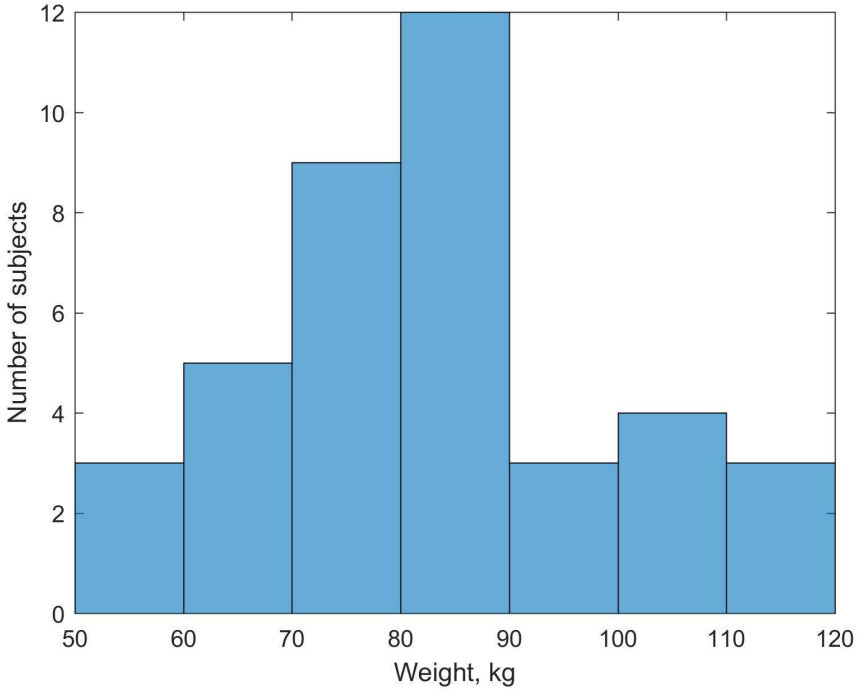
  - Discretization
    - Supervised
    - Unsupervised
  - Continuous feature modelling with probability density function

TAL TECH

UNSUPERVISED DISCRETIZATION

- The measured weight can be graphically represented with different number and width of bins, so it fills the whole scale of possible weights.





TAL TECH

SCHEMES OF UNSUPERVISED DISCRETIZATION

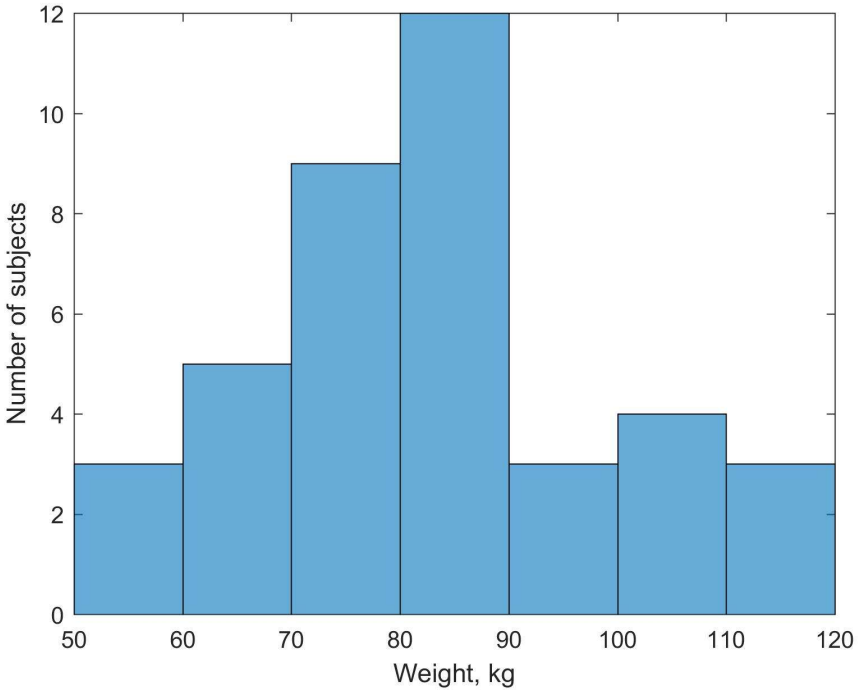
- There are different ways how to decide the number of bins:

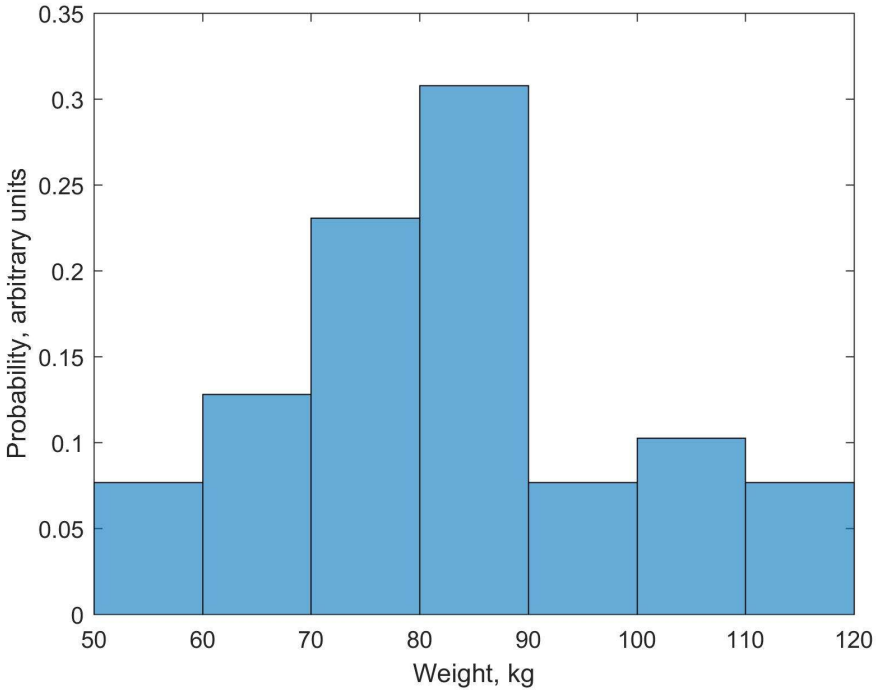
  - To choose the bins so that each bin has approximately the same number of instances: this is referred to as **equal-frequency discretization**.
    - More generally the bin boundaries are quantiles
    - E.g. If we choose two bins then this method coincides with thresholding on the median
    - E.g. If we choose four bins then this method coincides with thresholding on the quartile
  - **Equal-width discretization** chooses the bin boundaries so that each interval has the same width.
    - The width can be established by dividing the feature range by the number of bins if the feature has upper and lower limits
    - The bin boundaries at an integer number of standard deviations above and below the mean
  - To treat feature discretization as a univariate **clustering** problem.
    - E.g. In order to generate K bins we can uniformly sample K initial bin centres and run K-means until convergence.

TAL TECH

UNSUPERVISED DISCRETIZATION – EQUAL-WIDTH

- The probabilities for each bin can be calculated and used in classification.

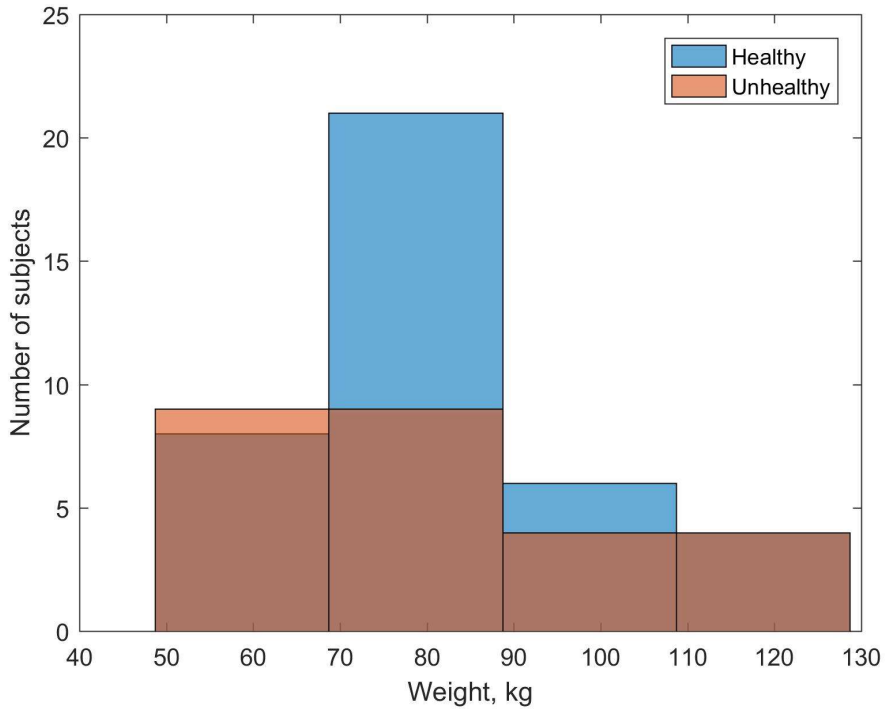


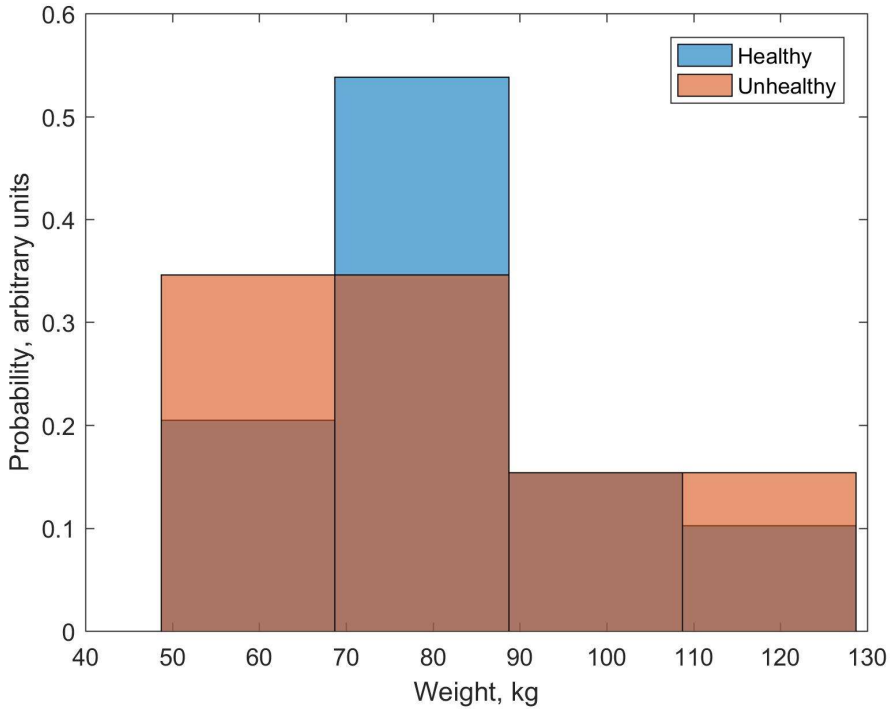


TAL TECH

UNSUPERVISED DISCRETIZATION – EQUAL-WIDTH

- The weights of 26 unhealthy subjects are added.

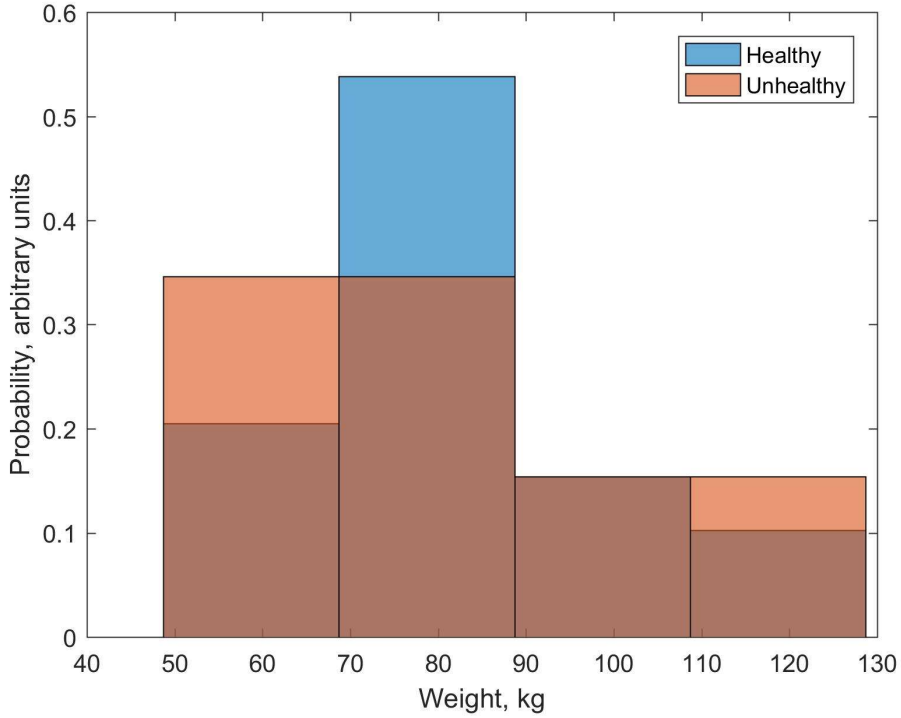




TAL TECH

UNSUPERVISED DISCRETIZATION – EQUAL-WIDTH

- The weights of unhealthy subjects are added.



Weight, kg (x_1)	Healthy $P(x_1\|c_1)$	Unhealthy $P(x_1\|c_2)$
48.7-68.7	0.2051	0.3462
68.7-88.7	0.5385	0.3462
88.7-108.7	0.1538	0.1538
108.7-128.7	0.1026	0.1538

TAL TECH

CONTINUOUS FEATURE MODELLING WITH FUNCTION

- The distribution of data can be modelled using some probability density function, e.g. Normal (Gaussian) distribution:

$$P(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{\frac{-(x-\mu)^2}{2 \cdot \sigma^2}}$$

- Anderson–Darling test can be used to detect whether the data set has normal distribution. (MATLAB function ‘adtest’)

- For normal probability density function the mean (μ) and standard deviation (σ) value of dataset has to be calculated.

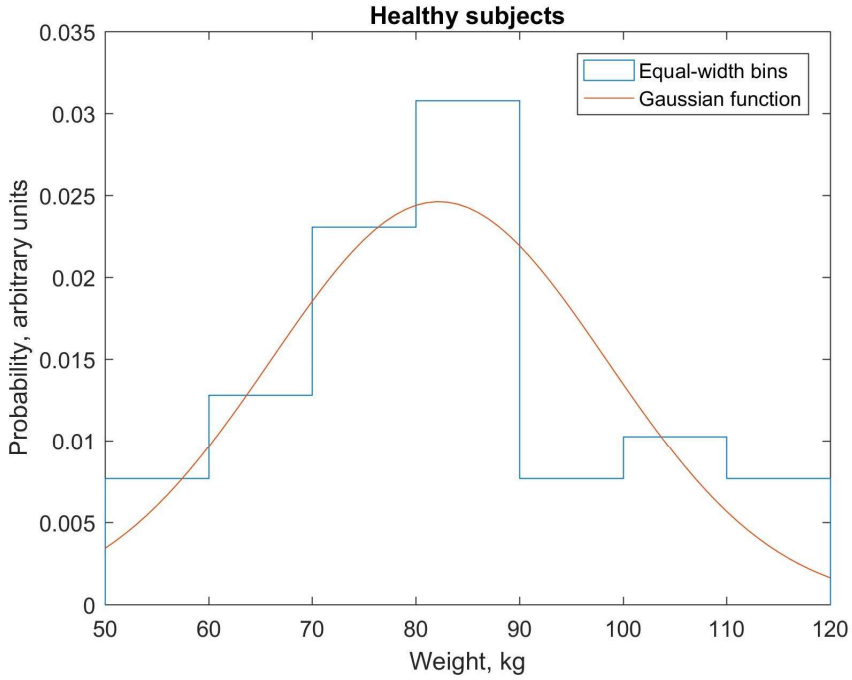
TAL TECH

EXAMPLE OF CONTINUOUS FEATURE MODELLING WITH FUNCTION

- The weight datasets of healthy and unhealthy subjects was tested using Anderson–Darling test and *P* values were recieved 0.1147 and 0.5710, respectively.

- In case the $p < 0.05$ then the dataset is not falling under normal distribution.

- Therefore, the normal probability density distribution function can be applied.



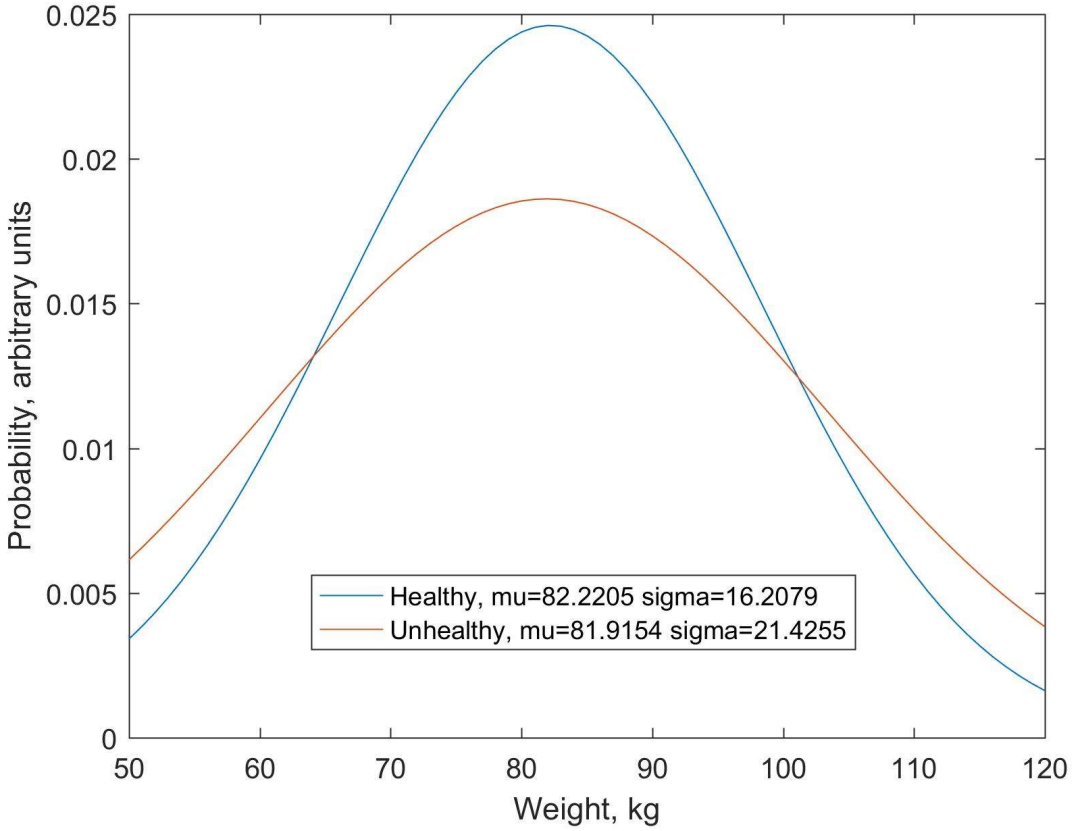
EXAMPLE OF CONTINUOUS FEATURE MODELLING WITH FUNCTION

- Let's assume that the feature of the classifier is weight of the subject.
- Then the conditional probability for healthy and unhealthy classes are calculated using the normal probability density function with determined variables μ and σ from training set.

$$P(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{\frac{-(x-\mu)^2}{2 \cdot \sigma^2}}$$

- The conditional probabilities for subject with 80kg are:
- $P(x_1=80|c_1) = 0.024$
- $P(x_1=80|c_2) = 0.019$

Healthy $P(x_1\|c_1)$		Unhealthy $P(x_1\|c_2)$	
μ	σ	μ	σ
82.22	16.2	81.92	21.43



EXAMPLE OF MIXED FEATURE NAÏVE BAYES CLASSIFIER

Weight, kg	Healthy $P(x_1\|c_1)$		Unhealthy $P(x_1\|c_2)$		Smoking	Healthy	Unhealthy		
(x_1)	μ	σ	μ	σ	(x_2)	$P(x_2\|c_1)$	$P(x_2\|c_2)$	$P(c_1)$	$P(c_2)$
	82.22	16.2	81.92	21.43	yes	0.234	0.567	0.6	0.4
					no	0.766	0.433		

- The Naïve Bayes classifier was trained based on training data from 65 subjects (39 healthy, 26 unhealthy), which represents some population. Unhealthy subjects were determined with diagnosis of any disease. The classifier has two inputs: weight of subject and smoking status.
- The test subject has weight of 95kg and smokes. Is he probably healthy or unhealthy?

- c_1: $P(c_1) \cdot \prod_{m=1}^{M} P(x_m|c_1) = 0.6 \cdot 0.0180 \cdot 0.234 = 0.002527$
- c_2: $P(c_2) \cdot \prod_{m=1}^{M} P(x_m|c_2) = 0.4 \cdot 0.0155 \cdot 0.567 = 0.003515$

TAL TECH

NAÏVE BAYES CLASSIFIER IN MATLAB

- The MATLAB function ‘fitcnb’ can be used to train Naive Bayes classifier:

- *Mdl* = fitcnb(*X, Y*);

  where *X* is the table of features (attributes or predictors) and *Y* is the class number (class labels)

- By default, the software models the distribution of features within each class using a Gaussian distribution.

- In case any of the feature is discreet then the Multivariate multinomial distribution (‘mvmn’) has to be used:

- *Mdl* = fitcnb(*X, Y, ‘DistributionNames’, ‘mvmn’)*;

TAL TECH

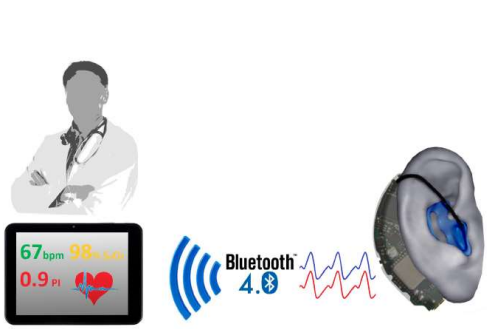
NAÏVE BAYES CLASSIFIER IN MATLAB

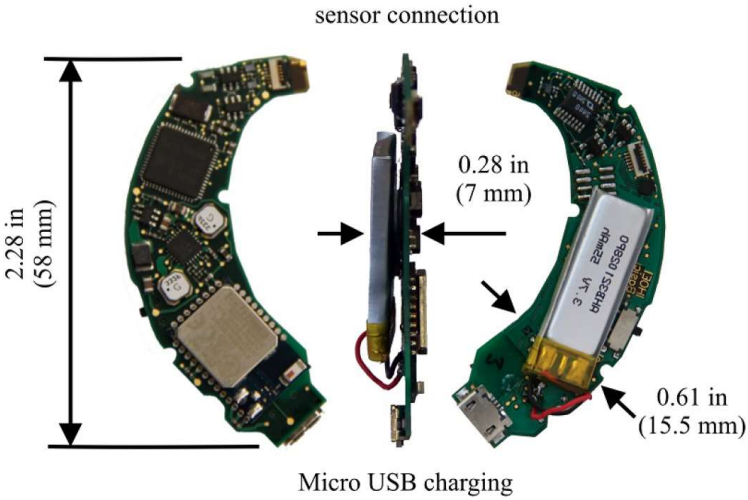
- In case the features in the table are continuous and discreet then the distribution has to be defined for every feature in the table *X*.
- *Mdl* = fitcnb(*X, Y, 'DistributionNames', {'mvmn', 'normal', 'normal', 'mvmn'}*);
- For the prediction of the class the ‘predict’ function can be used:
- *label* = predict(*Mdl*, *X_test*);

TAL TECH

NAÏVE BAYES CLASSIFIER EXAMPLE NR. 3

- Estimation of breathing cycle using in ear photoplethysmographic (PPG) signal.





B. Venema, V. Blazek and S. Leonhardt, "In-ear photoplethysmography for mobile cardiorespiratory monitoring and alarming",
2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2015, pp. 1-5, doi: 10.1109/BSN.2015.7299367.

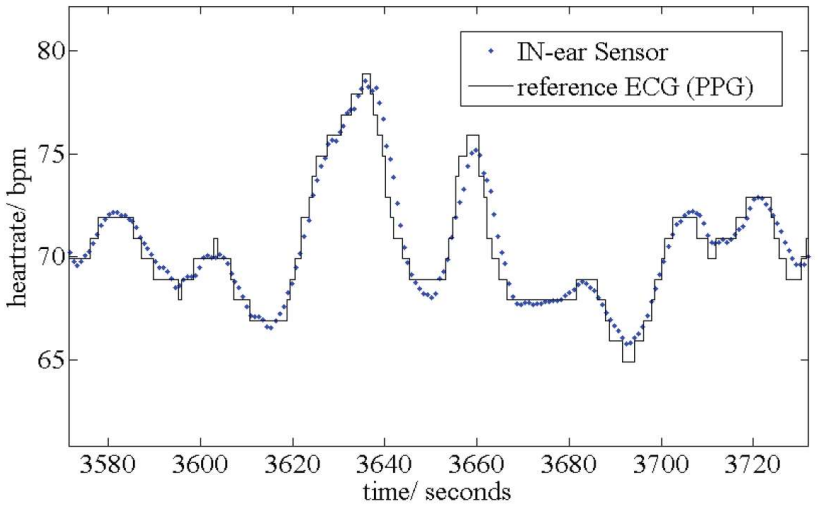
TAL TECH

NAÏVE BAYES CLASSIFIER EXAMPLE NR. 3

- The combination of PPG signal amplitude variation with heart rate variability (cardiorespiratory coupling) was used for breathing cycle estimation.

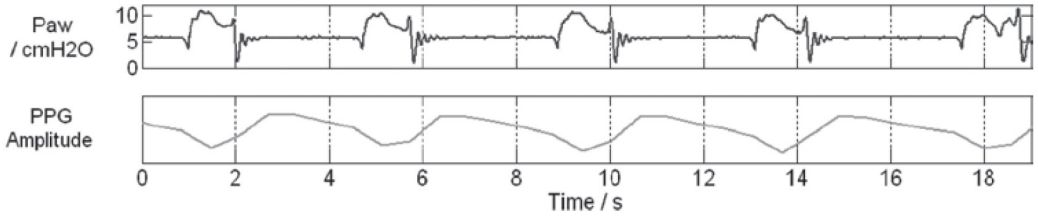
- The signals were separated into 0.25 second time sections.
- Three classes: breath-in, breath-out, nothing happens.
- For every section classification features were extracted from both signals:
  - Maximum value
  - Mean value
  - Standard deviation
  - Slope
- 1,827 respiration cycles from different subjects were used as training data.
- Naive Bayes classifier was trained.

Cardio respiratory coupling



B. Venema, V. Blazek and S. Leonhardt, "In-ear photoplethysmography for mobile cardiorespiratory monitoring and alarming", *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2015, pp. 1-5, doi: 10.1109/BSN.2015.7299367.

PPG signal amplitude variation



FINDING INDEPENDENT FEATURES

- Naive Bayes classifier assumes that the features are independent.

- Methods to test the feature independency:

  - **Pearson's correlation** – the linear correlation between two continuous variables
  - **Spearman's correlation** – the monotonic relationship between two continuous variables, making it more robust to nonlinear relationships
  - **Chi-square test of independence** – This is used when both variables are categorical. It assesses whether there is a significant association between the two variables.

TAL TECH

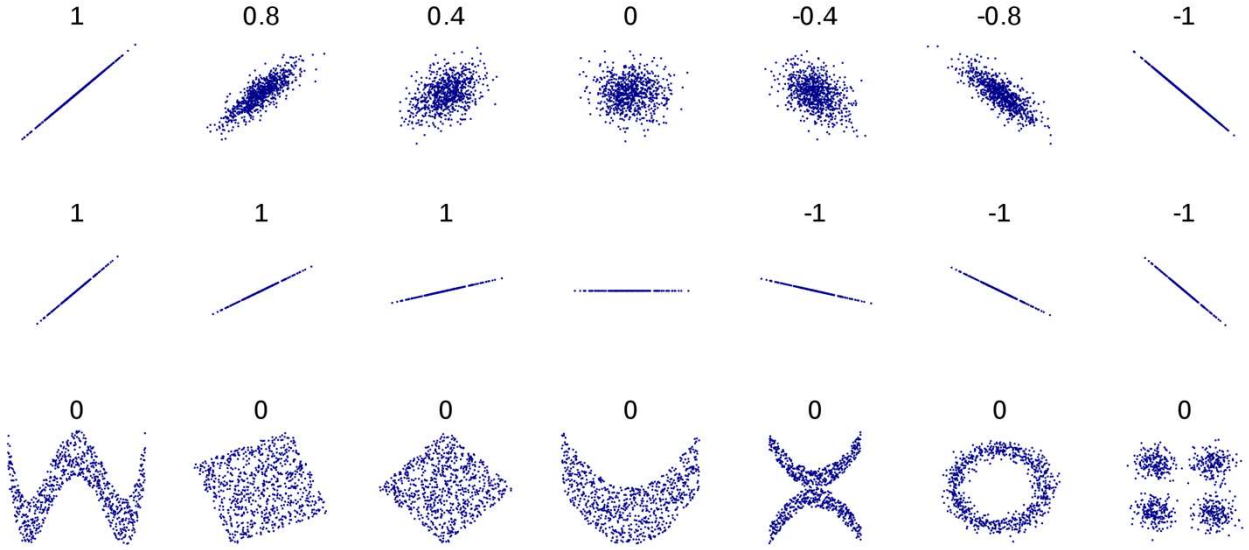
PEARSON'S CORRELATION

- Pearson's correlation coefficient, *r*, indicates how strong is the linear relationship between two continuous variables

Correlation Coefficient Value (*r*)	Direction and Strength of Correlation
-1	Perfectly negative
-0.8	Strongly negative
-0.5	Moderately negative
-0.2	Weakly negative
0	No association
0.2	Weakly positive
0.5	Moderately positive
0.8	Strongly positive
1	Perfectly positive

D Ratnasari *et al* 2016 *J. Phys.: Conf. Ser.* 694 012062

Pearson's correlation coefficients of different cases:



https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

PRACTICAL WORK ASSIGNMENT

- **Task:** The task is to train Naïve Bayes classifier based on training dataset and estimate the cardiovascular disease occurence of the subjects in testing dataset.

- Use MATLAB or Excel and process the dataset. Select the appropriate tools for the processing based on the lecture. Motivate shortly your selections (write as a comment to the script). Comment your code line by line. In case of Excel write short description in Word.
- **The result** is a MATLAB script or Excel chart, which includes the following sections:
  - Loading data
  - Conditioning of the features – finding the independent features using correlations (cross correlation table), feature smoothing, discretization or feature modelling where neccessary
  - Training of the Naïve Bayes classifier
  - Estimation of cardiovascular disease occurence of the subjects in testing dataset
- MATLAB has it's own Naïve Bayes classifier, however, in this task **you should write your own classifier or create Excel chart!**
- Nevertheless, you can check your result using MATLAB classifier
- Deadline: 16th of March 2025?

TAL TECH

Send the solution to: kristjan.pilt@taltech.ee