STRUCTURE AND COMPONENTS OF MACHINE LEARNING

Kristjan Pilt, PhD

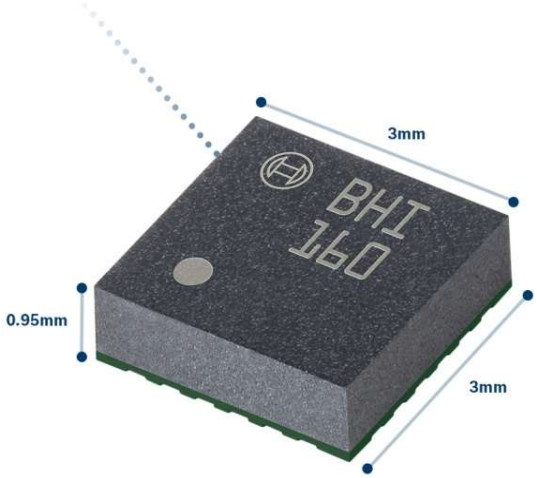
BACKGROUND

- Keywords: arterial stiffness, physiological measurements, signal processing, biomedical optics, photoplethysmography

- Machine learning experience: activity recognition, fall detection

Bosch Sensortec GmbH BOSCH

Activity recognition



Main features

Sensor Hub
Enables the full Android sensor stack

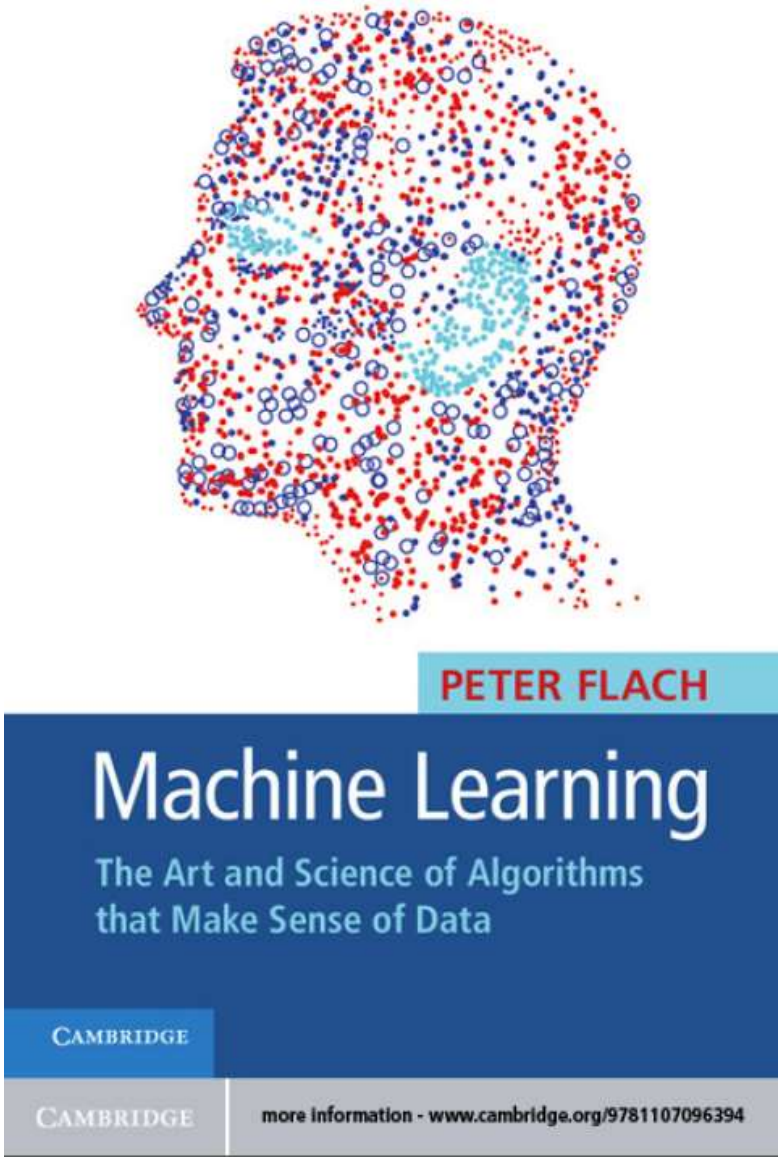
Accelerometer
Detects linear motion and gravitational forces

Gyroscope
Measures the rate of rotation in space (roll, pitch, yaw)

Software
Intelligently fuses raw data from multiple sensors

TAL TECH

LITERATURE



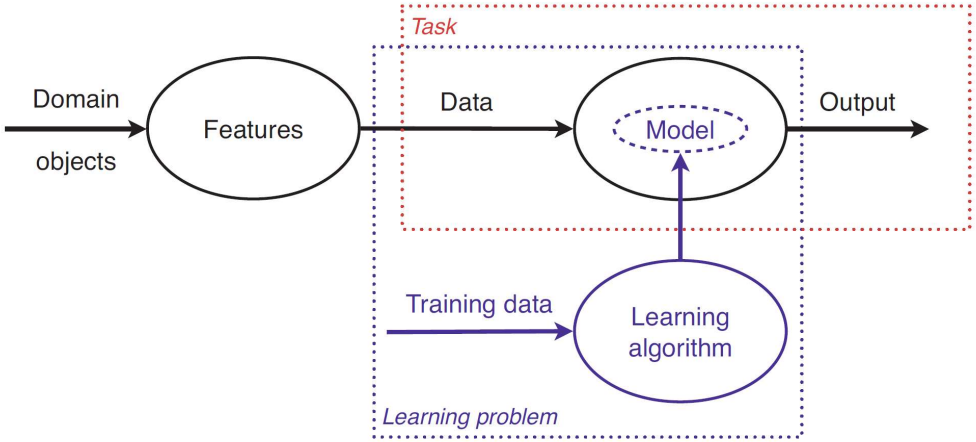
TAL TECH

OVERVIEW OF THE COMPONENTS

The ingredients of machine learning:

- **Tasks** – the problems that can be solved with machine learning
- **Models** – the output of machine learning
- **Features** – the workhorses of machine learning

- Aim of the machine learning:
- To find right features for compilation of right model to solve right task.
- A task requires an appropriate mapping – a model – from data described by features to outputs.



TAL TECH

EXAMPLE

The spam filter SpamAssassin – open-source spam filter:

```
-0.1 RCVD_IN_MXRATE_WL       RBL: MXRate recommends allowing
                             [123.45.6.789 listed in sub.mxrate.net]
 0.6 HTML_IMAGE_RATIO_02     BODY: HTML has a low ratio of text to image area
 1.2 TVD_FW_GRAPHIC_NAME_MID BODY: TVD_FW_GRAPHIC_NAME_MID
 0.0 HTML_MESSAGE            BODY: HTML included in message
 0.6 HTML_FONx_FACE_BAD      BODY: HTML font face is not a word
 1.4 SARE_GIF_ATTACH         FULL: Email has a inline gif
 0.1 BOUNCE_MESSAGE          MTA bounce message
 0.1 ANY_BOUNCE_MESSAGE      Message is some kind of bounce message
 1.4 AWL                     AWL: From: address is in the auto white-list
```

Adds a ‘junk’ flag and a summary report to the e-mail’s headers if the score is 5 or more.

TAL TECH

EXAMPLE

Another example of e-mail from „European Conference on Machine Learning (ECML) and the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)“

Domain name was: www.ecmlpkdd2008.org

```
2.5 URI_NOVOWEL             URI: URI hostname has long non-vowel sequence
3.1 FROM_DOMAIN_NOVOWEL     From: domain has series of non-vowel letters
```

The importance of a SpamAssassin test and test weights can be different for different users.

Machine learning is an excellent way of creating software that adapts to the user.

TAL TECH

EXAMPLE

Let's have two tests (x_1 and x_2)

Four training e-mails.

E-mail	x_1	x_2	Spam?	$4x_1 + 4x_2$
1	1	1	1	8
2	0	0	0	0
3	1	0	0	4
4	0	1	0	4

The e-mail is spam, if the score is 5 or more.

Just a mathematical problem?!

„Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience."

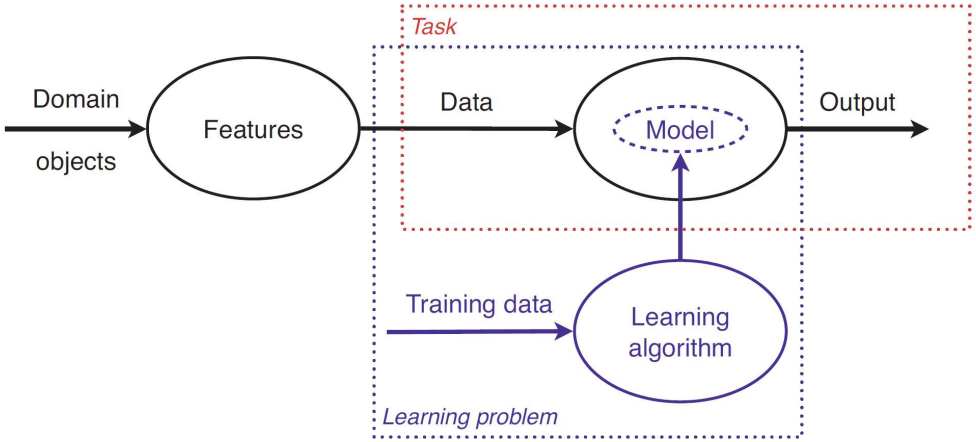
TAL TECH

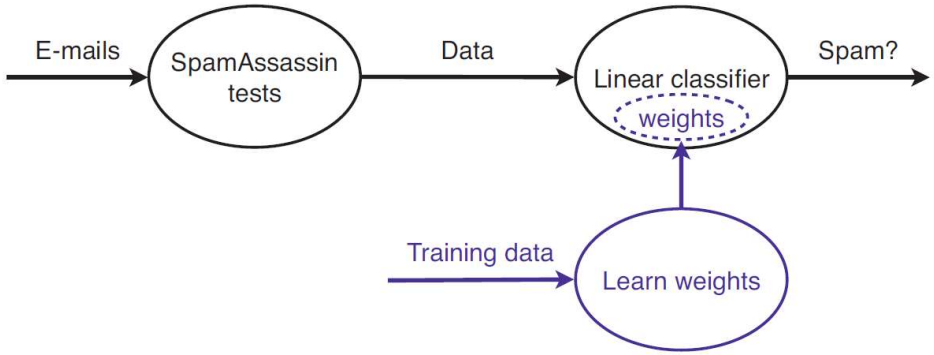
OVERVIEW OF THE COMPONENTS

The ingredients of machine learning:

- **Tasks** – the problems that can be solved with machine learning
- **Models** – the output of machine learning
- **Features** – the workhorses of machine learning

- Aim of the machine learning:
- To find right features for compilation of right model to solve right task.
- A task requires an appropriate mapping – a model – from data described by features to outputs.





TAL TECH

TASKS

- The most common machine learning tasks are predictive – predicting a target variable from features.
    - Binary and multi-class classification: categorical target (*E.g. distinguish between spam and ham e-mails*)
    - Regression: numerical target (*E.g. assessment of an incoming e-mail's urgency on a sliding scale*)
    - Clustering: hidden target (*E.g. Clustering similar e-mails into groups*)
- Descriptive tasks are concerned with exploting underlying structure in the data to discover associations.

TAL TECH

BINARY CLASSIFICATION TASK

- It constitutes a **binary classification task**, which is easily the most common task in machine learning.
- One obvious variation is to consider classification problems with more than two classes.
    - First task is to distinguish between spam and ham
    - The second task is, among ham e-mails, to distinguish between work-related and private ones

- It is often beneficial to view multi-class classification as a machine learning task in its own right.

- The notion of a decision boundary is less obvious when there are more than two classes.

TAL TECH

REGRESSION TASK

- Sometimes it is more natural to abandon the notion of discrete classes altogether and instead predict a real number - e.g. an assessment of an incoming e-mail’s urgency on a sliding scale (Example).

- This **task is called regression**, and essentially involves learning a real-valued function from training examples labelled with true function values.

- Constructing a function (e.g. the function value depends linearly on some numerical features) which minimizes the difference between the predicted and true function values.

TAL TECH

CLUSTERING TASK

- Classification and regression assume the availability of a training set
- Can we learn to distinguish classes without a labelled training set?
- The task of grouping data without prior information on the groups is called **clustering**.
- Learning from unlabelled data is called **unsupervised learning**.

TAL TECH

CLUSTERING

- Assessing the similarity between instances (the things we're trying to cluster, e.g., e-mails) and putting similar instances in the same cluster and 'dissimilar' instances in different clusters.

- E.g. the similarity of e-mails would be measured in terms of the words they have in common.

- One e-mail contains 42 (different) words and another contains 112 words, and the two e-mails have 23 words in common, then their similarity (**Jaccard coefficient**) would be:

$$23/(42+112-23) = 23/130 = 0.18$$

- The average similarity of an e-mail to the other e-mails in its group is much larger than the average similarity to e-mails from other groups.

TAL TECH

ASSOCIATION RULES

- There are many other patterns that can be learned from data in an unsupervised way.
- Association rules are a kind of pattern that are popular in marketing applications - ‘Customers Who Bought This Item Also Bought’
- Data mining algorithms that zoom in on items that frequently occur together.
- There exist many other types of associations that can be learned and exploited, such as correlations between real-valued variables.

TAL TECH

LOOKING FOR STRUCTURE

- Hidden variable in here is the number of film genres
- Methods for discovering hidden variables come into their own when the number of values of the hidden variable is much smaller than the number of rows and columns of the original matrix.
- Matrix decomposition can often reveal useful hidden structure.

Subject nr.	The Shawshank Redemption	The Usual Suspects	The Godfather	The Big Lebowski
1	1	0	1	0
2	0	2	2	2
3	0	0	0	1
4	1	2	3	2
5	1	0	1	1
6	0	2	2	3
Average	0.5	1	1.5	1.5

=

Subject nr.	Drama	Crime	Comedy
1	1	0	0
2	0	1	0
3	0	0	1
4	1	1	0
5	1	0	1
6	0	1	1

X

Genre	People's preferences		
	Drama	Crime	Comedy
Drama	1	0	0
Crime	0	2	0
Comedy	0	0	1

X

Genre	The Shawshank Redemption	The Usual Suspects	The Godfather	The Big Lebowski
Drama	1	0	1	0
Crime	0	1	1	1
Comedy	0	0	0	1

TAL TECH

SUMMARIZING TERMINOLOGY

- Distinction between **supervised learning** from labelled data and **unsupervised learning** from unlabelled data.
- Distinction between whether the model output involves the target variable (**predictive model**) or not (**descriptive model**)

	Predictive model	**Descriptive model**
Supervised learning	classification, regression	subgroup discovery
Unsupervised learning	predictive clustering	descriptive clustering, association rule discovery

TAL TECH

- **Semi-supervised learning** – labelling small amount of data => training model => classifying unlabeled data => using most confident predictions = > refining model

EVALUATING PERFORMANCE ON TASK

- An important thing to keep in mind with all these machine learning problems is that they don't have a 'correct' answer!

- If you sort the entries in your address book alphabetically on last name, there is only one correct result.
  - Different algorithms, different speeds, amount of data to handle
  - We never compare such algorithms with respect to the correctness of the result

- The things are different with machine learning.

- We need to have some idea of how well an **algorithm is expected to perform** on new data, not in terms of runtime or memory usage – although this can be an issue too – but **in terms of classification performance** (if our task is a classification task).

TAL TECH

PERFORMANCE EVALUATION

- Performance of spam filter (example).
  - Count the number of correctly classified e-mails, both spam and ham, and divide that by the total number of examples to get a proportion which is called the **accuracy of the classifier**.
  - It is unclear whether the **overfitting** occured
  - 90% of the data for training, and the remaining 10% as a test set.
  - If overfitting occurs, the test set performance will be considerably lower than the training set performance.
  - Lucky or unlucky result

- **Cross-validation** – train-test split is repeated and the accuracy is evaluated.

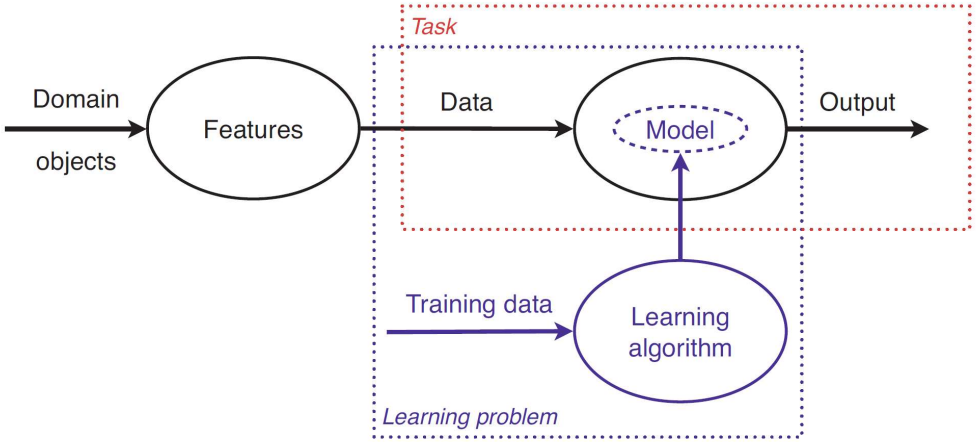
- Cross-validation can also be applied to other **supervised learning** problems, but **unsupervised learning** methods typically need to be evaluated differently.

TAL TECH

MODELS

The ingredients of machine learning:

- **Tasks** – the problems that can be solved with machine learning
- **Models** – the output of machine learning
- **Features** – the workhorses of machine learning



TAL TECH

MODELS

Models learned from the data, in order to solve a given task.

Machine learning models can be distinguished according to their main intuition:

- **Geometric models** use intuitions from geometry such as separating (hyper-)planes, linear transformations and distance metrics.
- **Probabilistic models** view learning as a process of reducing uncertainty, modelled by means of probability distributions.
- **Logical models** are defined in terms of easily interpretable logical expressions.

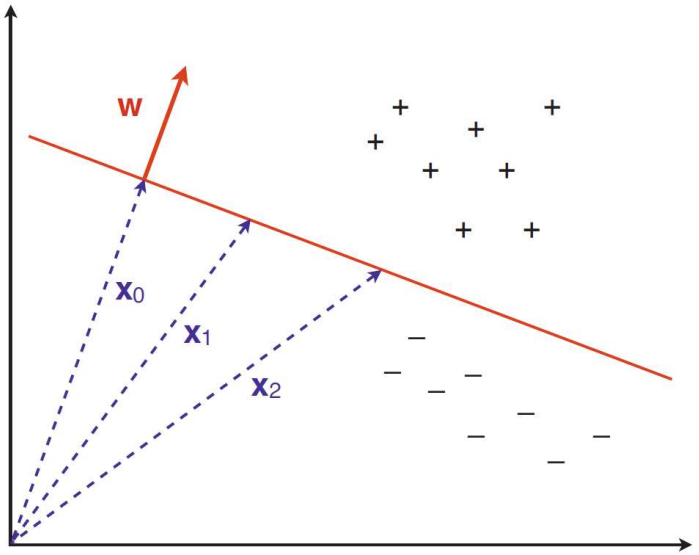
Alternatively, they can be characterised by their mode of operation:

- **Grouping models** divide the instance space into segments; in each segment a very simple (e.g., constant) model is learned.
- **Grading models** learning a single, global model over the instance space.

TAL TECH

GEOMETRIC MODELS

- **Instance space** is the set of all possible or describable instances.
- If all features are numerical, then we can use each feature as a coordinate in a Cartesian coordinate system.
- A **geometric model** is constructed directly in instance space, using geometric concepts such as lines, (hyper-) planes and distances.
- For instance, the given linear classifier is a geometric classifier.
- A Cartesian instance space has as many coordinates as there are features, which can be tens, hundreds, thousands, or even more.



TAL TECH

BASIC LINEAR CLASSIFIER

- The data is linearly separable if linear decision boundary separates the two classes.

- A linear decision boundary is defined by the equation:

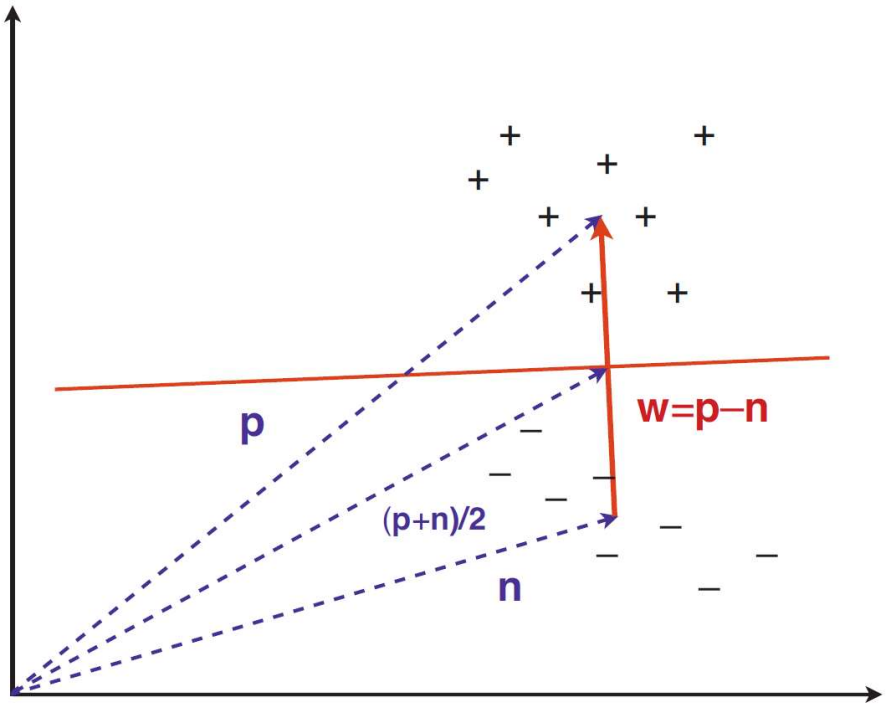
$$\mathbf{w} \cdot \mathbf{x} = t,$$

where **w** is a vector perpendicular to the decision boundary, **x** points to an arbitrary point on the decision boundary, and *t* is the decision threshold.

- A good way to think of the vector **w** is as pointing from the 'centre of mass' of the negative examples, **n**, to the centre of mass of the positives **p**.

- The 'centre of mass' are calculated:

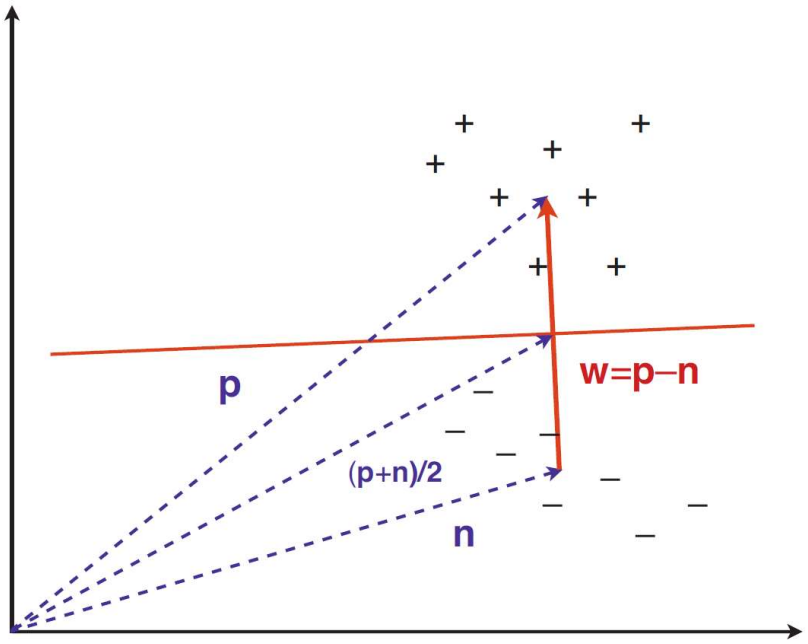
$$\mathbf{p} = \frac{1}{k} \cdot \sum_{x \epsilon} \mathbf{x} \qquad \mathbf{n} = \frac{1}{k} \cdot \sum_{x \epsilon n} \mathbf{x}$$



TAL TECH

BASIC LINEAR CLASSIFIER

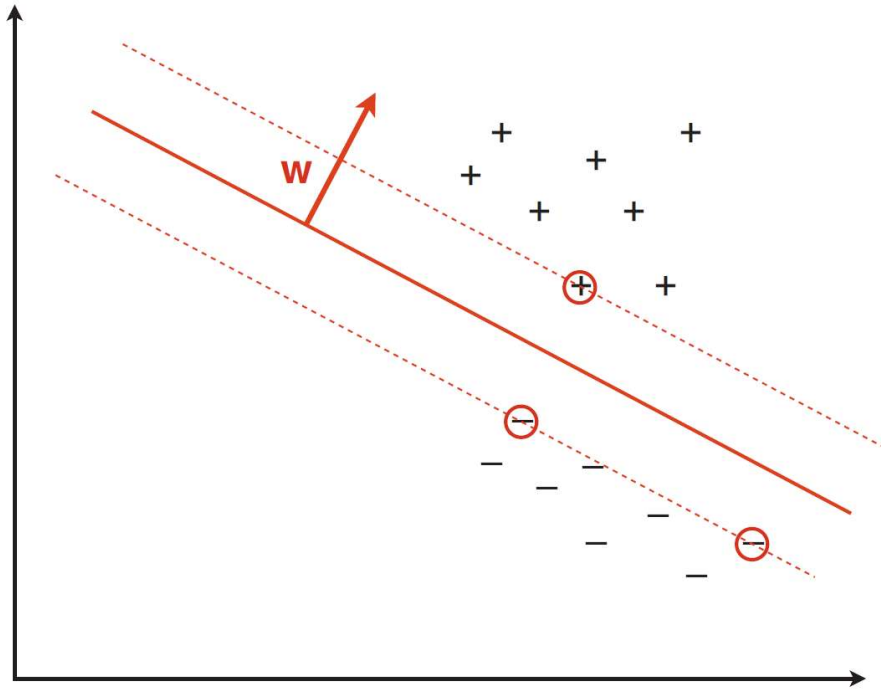
- **w** is proportional (or equal) to **p**−**n**.
- By setting the decision threshold appropriately, we can intersect the line from **n** to **p** half-way.
- This is called the **basic linear classifier**.
- Because data is usually noisy, linear separability doesn't occur very often in practice, unless the data is very sparse, as in text classification.



TAL TECH

BASIC LINEAR CLASSIFIER

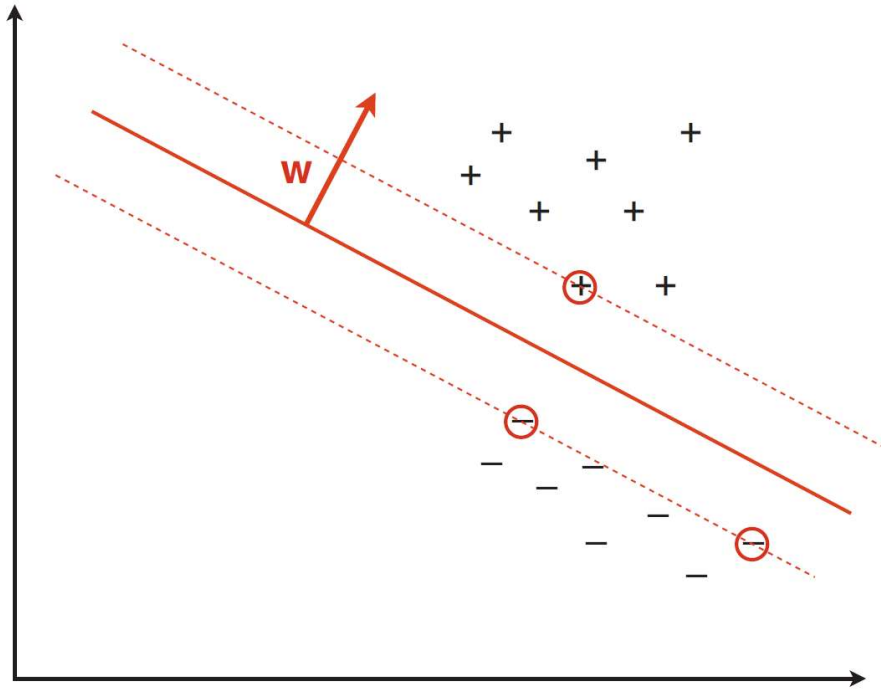
- Which of the infinitely many decision boundaries should we choose?
- Large margin classifiers, where the *margin* of a linear classifier is the distance between the decision boundary and the closest instance.
- **Support vector machines** are a powerful kind of linear classifier that find a decision boundary whose margin is as large as possible.



TAL TECH

NEAREST-NEIGHBOUR CLASSIFIER

- If the **distance** between two instances (x and y) is small then the instances are similar in terms of their feature values, and so nearby instances would be expected to receive the same classification or belong to the same cluster.
- Euclidean distance: $d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$
- To classify a new instance, we retrieve from memory the most similar training instance (i.e., the training instance with smallest Euclidean distance from the instance to be classified), and simply assign that training instance's class.
- This classifier is known as the **nearest-neighbor classifier**.



TAL TECH

VARIATIONS OF NEAREST-NEIGHBOUR CLASSIFIER

- We can retrieve the *k* most similar training instances and take a vote (k-nearest neighbor)
- We can weight each neighbor's vote inversely to its distance
- etc.
- Predictions are local in the sense that they are based on only a few training instances, rather than being derived from a global model built from the entire data set.

TAL TECH

K-MEANS CLUSTERING

- Suppose we want to cluster our data into K clusters, and we have an initial guess of how the data should be clustered.

1. The means of each initial cluster is calculated
2. Each point is reassigned to the nearest cluster mean.
3. Unless the initial guess was a lucky one, this will have changed some of the mean values of clusters, so we repeat steps 1 and 2 until no change occurs.

- How to construct the initial guess of the clusters?

- This is usually done randomly: either by randomly partitioning the data set into K ‘clusters’ or by randomly guessing K ‘cluster centres’.

TAL TECH

PROBABILISTIC MODELS

- Let X denote the variables we know about, e.g., our instance's feature values; and let Y denote the target variables we're interested in, e.g., the instance's class.

- The key question in machine learning is how to model the relationship (f) between X and Y .

- The statistician's approach is to assume that there is some underlying random process that generates the values for these variables, according to a well-defined but unknown probability distribution. We want to use the data to find out more about this distribution.

TAL TECH

PROBABILISTIC MODELS

- How to use the defined probability distribution?
- Y could indicate whether the e-mail is spam, and X could indicate whether the e-mail contains the words ‘Viagra’ and ‘lottery’.
- The probability of interest is then P(Y |Viagra, lottery) <- conditional probability
- For a particular e-mail: P(Y |Viagra =1, lottery = 0)
- This is called a **posterior probability** because it is used after the features X are observed.

TAL TECH

DECISION RULE

Viagra	lottery	$P(Y = \text{spam}\|\text{Viagra}, \text{lottery})$	$P(Y = \text{ham}\|\text{Viagra}, \text{lottery})$
0	0	0.31	**0.69**
0	1	**0.65**	0.35
1	0	**0.80**	0.20
1	1	0.40	**0.60**

- To classify a new e-mail we determine whether the words 'Viagra' and 'lottery' occur in it, look up the corresponding probability P(Y = spam|Viagra, lottery), and predict spam if this probability exceeds 0.5 and ham otherwise.

- Such a recipe to predict a value of Y on the basis of the values of X and the posterior distribution P(Y |X) is called a **decision rule**.

- Even though this example table is small, it will grow unfeasibly large very quickly (with *n* Boolean variables 2^n cases have to be distinguished)

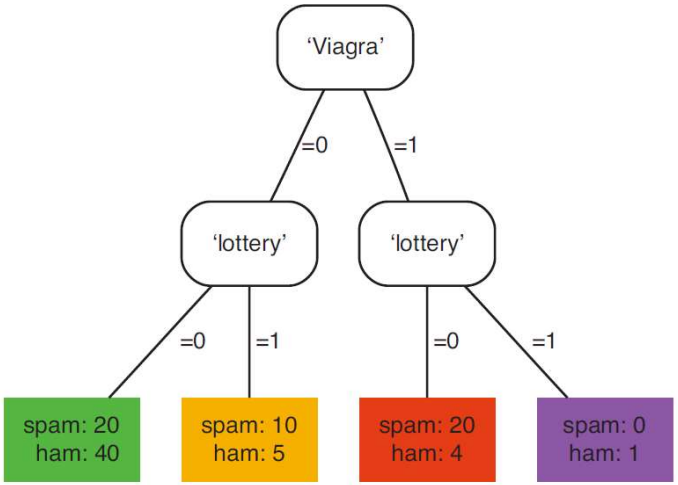
TAL TECH

LOGICAL MODELS

- Logical models are more algorithmic in nature.
- This type models can be easily translated into rules that are understandable by humans, such as:

  if Viagra = 1 then Class = Y = spam

- Can be organised in a tree structure, which can be called as a **feature tree**.

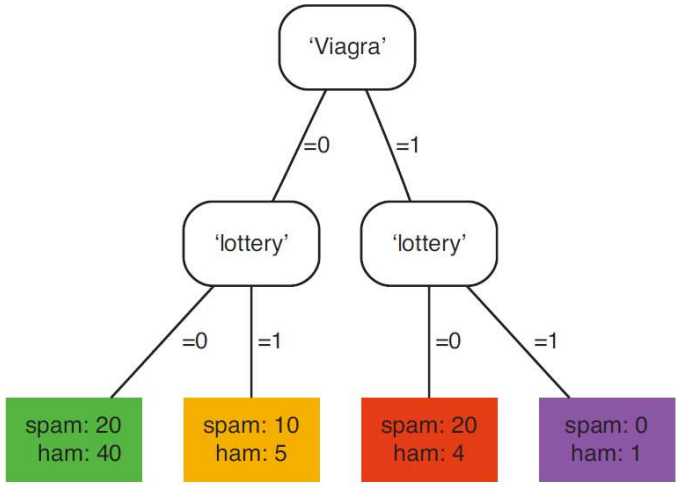


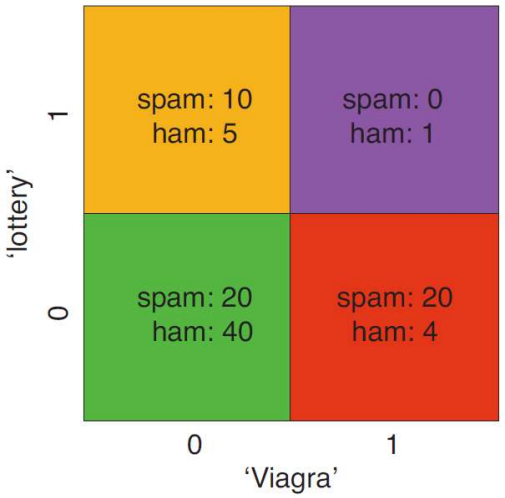
DECISION TREE

- The leaves of the tree therefore correspond to rectangular areas **in the instance space** (or hyperrectangles, more generally) which we will call **instance space segments**, or segments for short.

- Depending on the task we are solving, we can then label the leaves with a class, a probability, a real value, etc.

- Feature trees whose leaves are labelled with classes are commonly called **decision trees**.

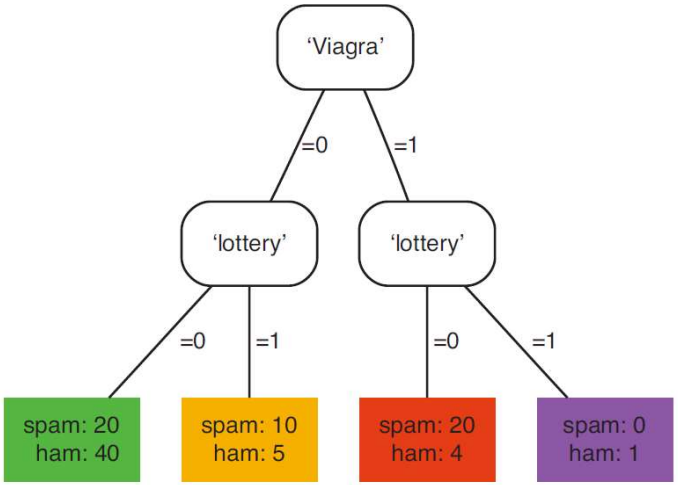


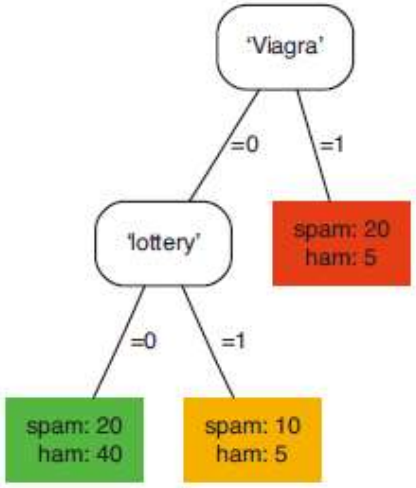


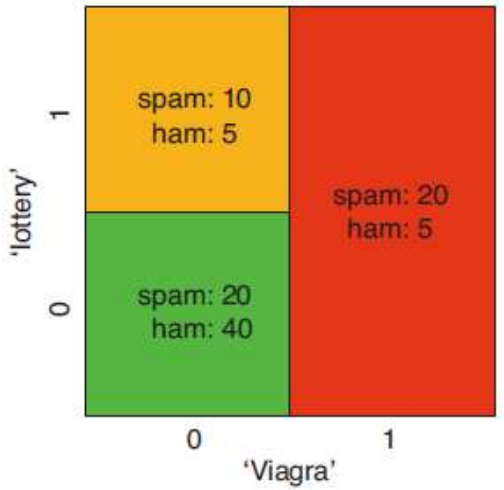
PRUNING

- Since the right most leaf covers only a single example, there is a danger that this tree is overfitting the data and that the previous tree is a better model.

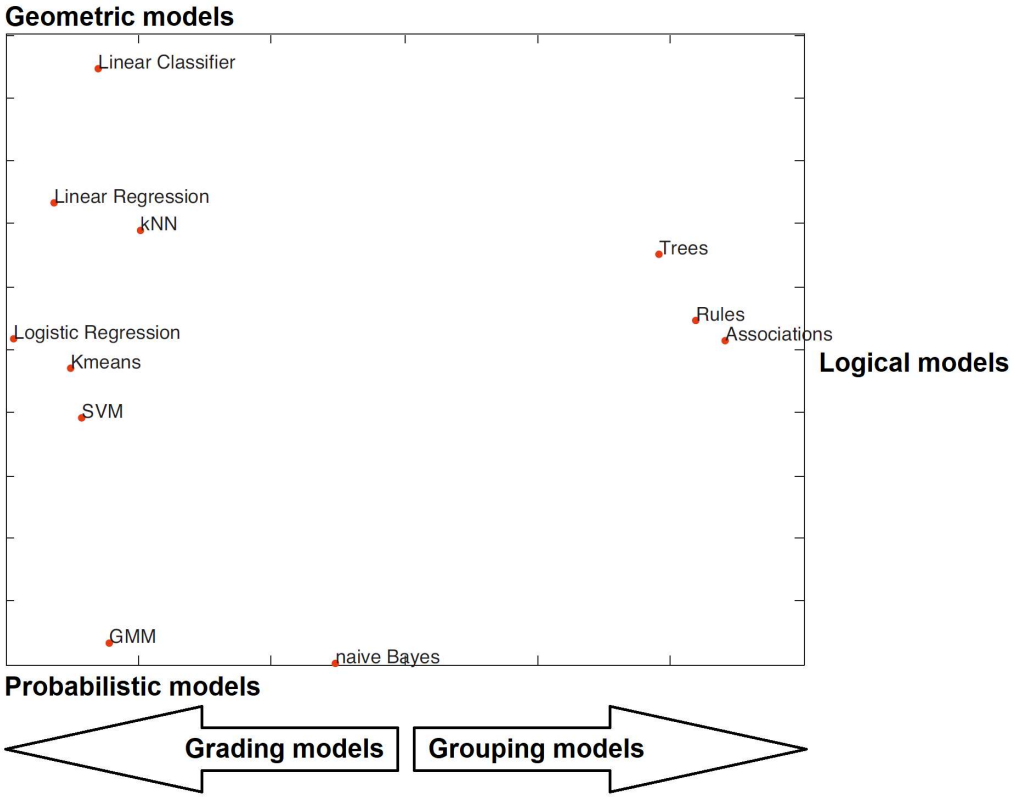
- Decision tree learners often employ **pruning** techniques which delete splits such as these.

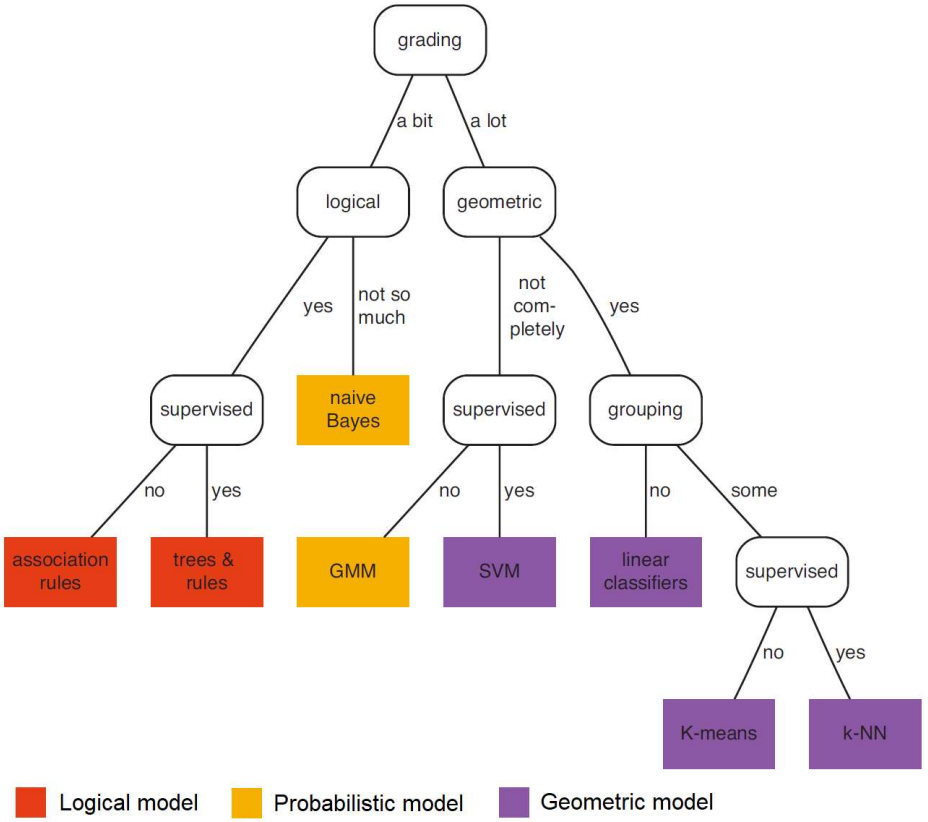






GROUPING AND GRADING





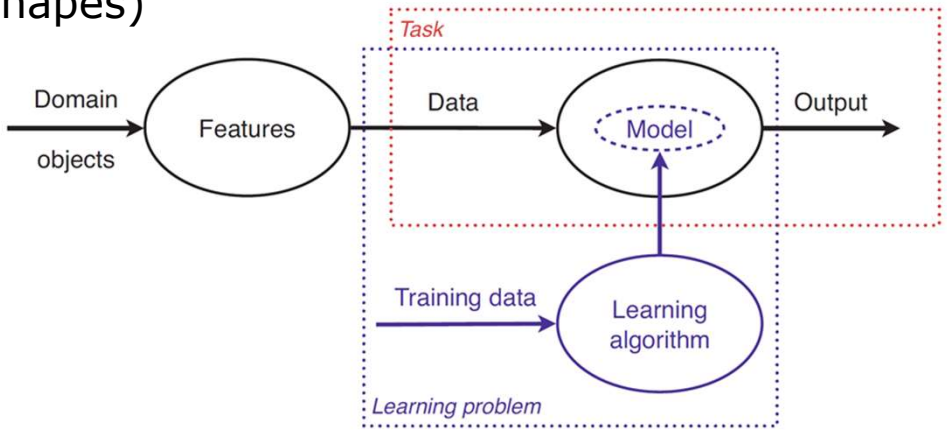
FEATURES

- A feature can be thought of as a kind of measurement that can be easily performed on any instance.

- Mathematically, features are functions that map from the instance space to some set of feature values called the domain of the feature.

- Since measurements are often numerical, the most common feature domain is the set of real numbers. Other typical feature domain include:
  - Set of integers (E.g. When the feature counts something)
  - Booleans (E.g. Feature is a statement that can be true or false)
  - Arbitrary finite sets (E.g. Set of colours, a set of shapes)

- E.g For classification of tumor cells the cell symmetry and cell radius are the features.



TAL TECH

TWO USES OF FEATURES

- In first case the features are **used for splits**.
- Very common use of features, particularly in logical models, is to zoom in on a particular area of the instance space.
- Let f be a feature counting the number of occurrences of the word ‘Viagra’ in an e-mail, and let x stand for an arbitrary e-mail, then the condition $f(x) = 0$ selects e-mails that don’t contain the word ‘Viagra’, $f(x) = 0$ or $f(x) > 0$ selects e-mails that do, $f(x) \geq 2$ selects e-mails that contain the word at least twice, etc.
- Such conditions are called **binary splits**, because they divide the instance space into two groups: those that satisfy the condition, and those that don’t.

TAL TECH

TWO USES OF FEATURES

- **Non-binary splits** are also possible

- If g is a feature that has the value:
  - 'tweet' for e-mails with up to 20 words,
  - 'short' for e-mails with 21 to 50 words,
  - 'medium' for e-mails with 51 to 200 words,
  - 'long' for e-mails with more than 200 words.

- then the expression g(x) represents a four-way split of the instance space.

TAL TECH

TWO USES OF FEATURES

- In second case the features are **used as predictors**.

- Linear classifier employs a decision rule of the form $\sum_{i=1}^{n} w_i \cdot x_i > t$

, where x_i is a numerical feature.

- The linearity of this decision rule means that each feature makes an independent contribution to the score of an instance.

- This contribution depends on the weight w_i:
  - if w_i is large and positive, a positive x_i increases the score;
  - if $w_i < 0$, a positive x_i decreases the score;
  - if $w_i \approx 0$, x_i influence is negligible.
- Thus, the feature makes a precise and measurable contribution to the final prediction.

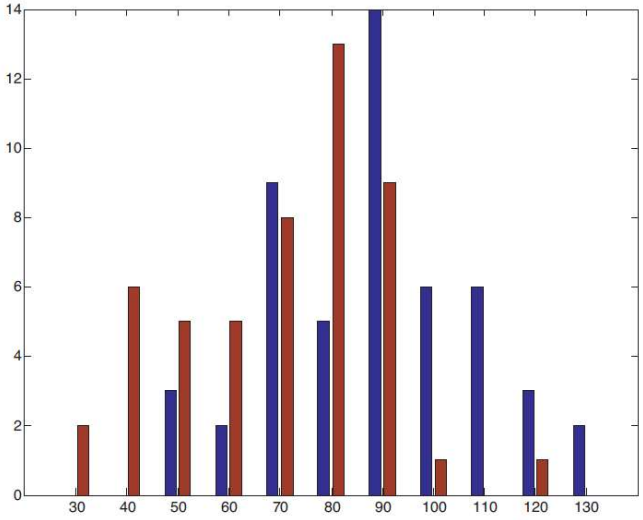
TAL TECH

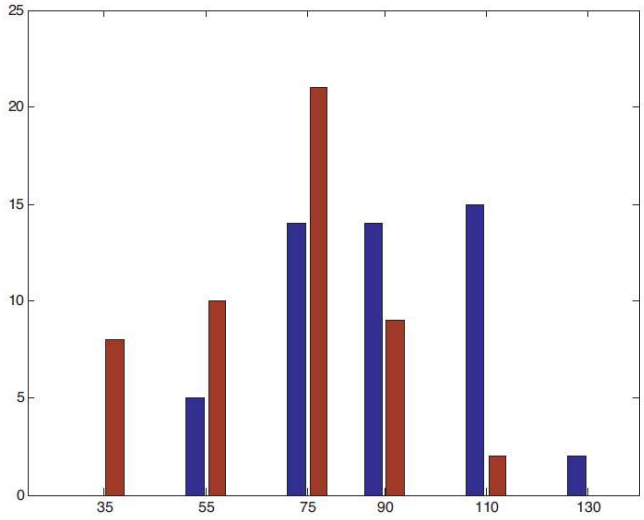
FEATURE CONSTRUCTION

- In the spam filter example, and text classification more generally, the messages or documents don't come with built-in features; rather, they need to be constructed by the developer of the machine learning application.

- This **feature construction** process is absolutely crucial for the success of a machine learning application.

- Real-valued features often contain unnecessary detail that can be removed by **discretisation** (e.g. body weight).





TAL TECH