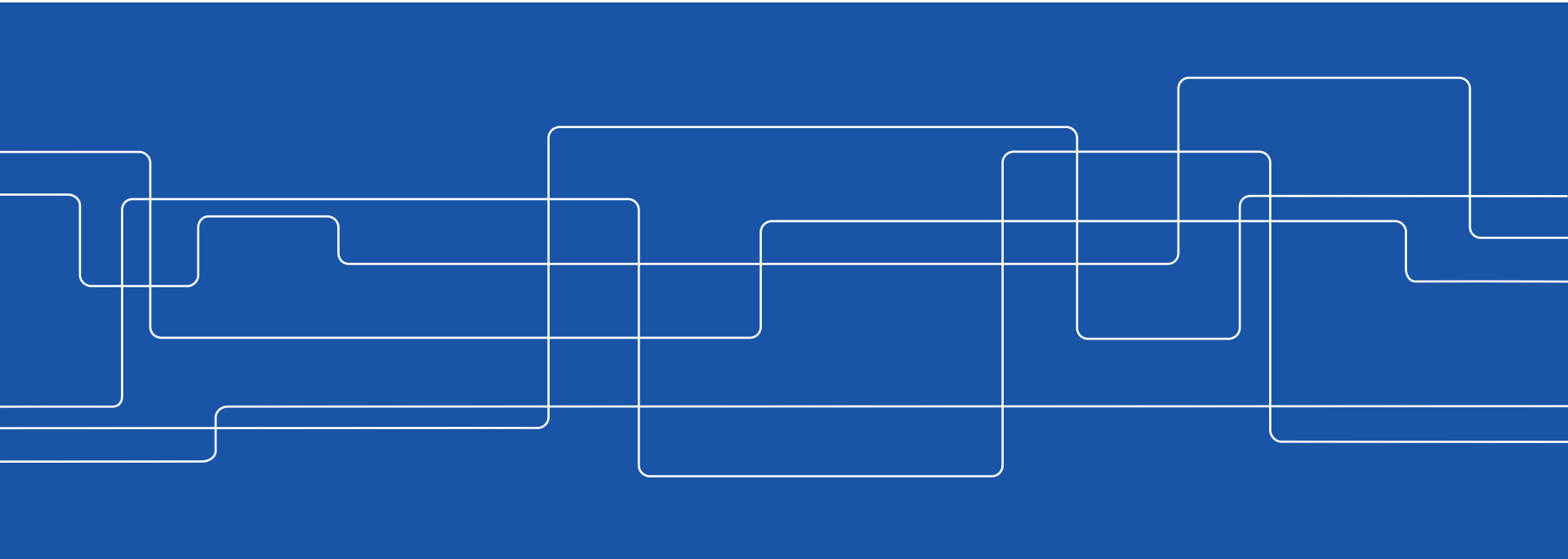
KTH ROYAL INSTITUTE
OF TECHNOLOGY

Lecture 12
Introduction to Machine Learning



On the Shoulder of Giants

Much of the material in this slide set is based upon:

”Automated Learning techniques in Power Systems” by L. Wehenkel, Université Liege

”Probability based learning” Josephine Sullivan, KTH

”Entropy and Information Gain” by F.Aiolli, University of Padova

Contents

Power Systems Analysis - traditionally

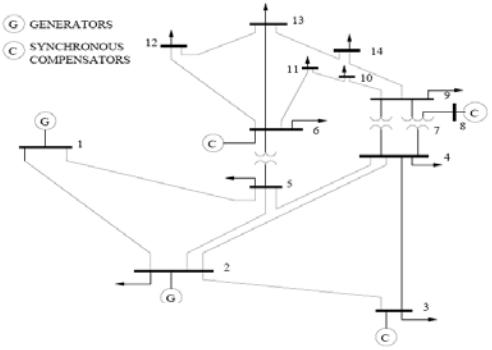
Power system analysis, control and operation is dependent on models

$$0 = -P_i + \sum_{k=1}^{N} |V_i||V_k|(G_{ik}\cos\theta_{ik} + B_{ik}\sin\theta_{ik})$$

$$0 = -Q_i + \sum_{k=1}^{N} |V_i||V_k|(G_{ik}\sin\theta_{ik} - B_{ik}\cos\theta_{ik})$$

Using the models, analytical and numerical analysis provides decision support for e.g.

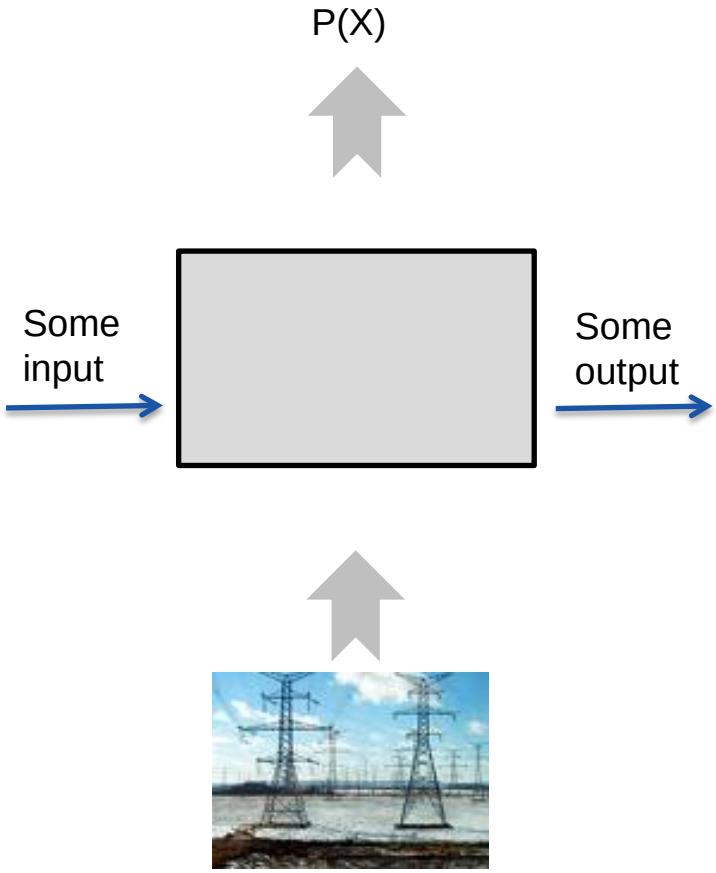
- Security
- Stability
- Optimal power flow
- Contingency analysis
- Expansion planning
- Market clearing



Power Systems Analysis – An automated learning approach

Understanding states in the power system is established through observation of inputs and outputs without regard to the physical electrotechnical relations between the states.

Adding knowledge about the electrotechnical rules means adding heuristics to the learning.



Given a set of examples (the learning set (LS)) of associated input/output pairs, derive a general rule representing the underlying input/output relationship, which may be used to explain the observed pairs and/or predict output values for any new unseen input.

Why automated (or machine) learning

Historically it developed as computers became more powerful (1990s and on). Reached a peak late 1990s.

Provides complementary perspectives to systems approach:

- Computational efficiency.

  Creating models of 10000+ bus systems including load and weather forecasts is challenging using system theoretical approach

- Interpretability

  The correlation of variables can more directly be adressed enabling understanding of phenomena across subject fields

- Management of uncertainties

  Errors in sensors and parameters can be disregarded

What is the ”learning” in learning

Machine learning depends on having data about the system (power system) to be analysed/modeled. Such data should suitably be information about some state of the system.

Typical example of data is time-series of measurements

A trivial example is in weather forecasting, where time-series of temperature, wind speed and humidity from various locations across the country are obviously useful to create forecasts for the weather the next day.

Based on the input ”*Learning*” data set, models are created and thereafter the models are fed with real data, and their ability to forecast results is evaluated.

Classes of methods for learning

In **Supervised** learning a set of input data and output data is provided, and with the help of these two datasets the model of the system is created.

For this introductory course, our focus is here With a short look at unsupervised learning

In **Unsupervised** learning, no ideal model is anticipated, but instead the analysis of the states is done in order to identify possible correlations bewteen datapoints.

In **Reinforced** learning, the model in the system can be gradually refined through means of a utility function, that tells the system that a certain ouput is more suitable than another.

Classification vs Regression

Two forms of Supervised learning

Classification: The input data is number of switch operations a circuitbreaker has performed and tthe output is a notification whether the switch needs maintenance or not. ”Boolean”

Regression: Given the wind speed in a incoming weather front, the output is the anticipated production in a set of wind turbines. ”Floating point”

Supervised learning - a preview

In the scope of this course, we will be studying three forms of supervised learning.

- Decision Trees

  Overview and practical work on exercise session.

- Artificial Neural Networks

  Overview only, no practical work.

- Statistical methods – k-Nearest Neighbour

  Overview and practical work on exercise session. Also included in Project Assignment

  kNN algorithm can also be used for unsupervised clustering.

Steps of developing a learning model

Representation

Which are the inputs and outputs of relevance to the model – what is it that we want the model to be able to tell us?

Attribute selection

Reduction to the minimal set of useful attributes that will be used in creation of the model, making sure to remove non-relevant parameters.

Model selection

Create, or find, the suitable model of a certain structure (Decision tree, ANN, kNN, etc.) depending on the chosen set of input and output parameters.

Interpretation and validation of the model to make sure that the developed model provides output consistent with prior knowledge and physical limitations.

Once at this stage, the model can be used to predict new outputs based on alternative inputs – all within the constraints of its applicability as determined from the above process.

Contents

Some further notation

The universe $\mathbf{U}$ is the set of all possible objects o

Objects have attributes denoted with $a(o)$

With $a(X)$, where $\mathbf{X}$ is a subset of $\mathbf{U}$, we mean all possible values that $a(.)$ can take in $\mathbf{X}$

For any subset V of $\mathbf{a(U)}$ we denote by $a^{-1}(V)$ the set of objects that exist in U and whose attrbute values are in V

A database (yet a definition of the term)

A database is a subset of **objects** described by a certain number of **attributes** providing information about each object.

In a power system application, a database is for instance a number of simulated scenarios of varying load and generation profiles.

- Candidate Attributes are those that are used to develop (learn) the rule – also called Learning Set (LS)
- Selected Attributes are those that are used in the rule to validate it (also called Test Set)

Classification problems

Classification problems are such that we want to assign each object in the database to a specific class $\mathbf{C}$.

Each object in U can be assigned to one of mutually excluding classes $\mathbf{C}$, where $\mathbf{c(o)}$ denotes the class of $\mathbf{o}$.

$$C \triangleq \{c_1, \ldots, c_m\}$$

This can formally be specified as:

$$C = \{C_1, \ldots, C_m\} \quad : \quad C_i \triangleq \{o \in U | c(o) = c_i\}.$$

Learning Sets and Test sets

In a classification problem the Learning set is

$$\boldsymbol{LS} \triangleq \{(\boldsymbol{v}^1, c^1), (\boldsymbol{v}^2, c^2), \ldots, (\boldsymbol{v}^N, c^N)\},$$

where

$$\boldsymbol{v}^k = (v_1^k, v_2^k, \ldots, v_n^k)^T = \boldsymbol{a}(o_k)$$

And

$$c^k = c(o_k)$$

Decision rules for classification

A Decision function, is a function $d(.)$ on a object such that

$$d(a_1(o), \ldots, a_n(o)) \text{ or simply } d(o) \quad : \quad \boldsymbol{U} \longmapsto \boldsymbol{C}.$$

A decision rule will create partitions of the universe U (separate all the objects in different subsets D_i)

$$\boldsymbol{D}_i \triangleq d^{-1}(c_i) = \{o \in \boldsymbol{U} \mid d(o) = c_i\} \quad (i = 1, \ldots, m).$$

The Hypothesis space is the space of all possible decision rules d that can operate on objects o in U

Regression problems

Regression problems takes as input objects o and provide as output a real number $y(U)$ (not a class).

$$r(a_1(o), \ldots, a_n(o)) \text{ or simply } r(o) \quad : \quad U \longmapsto y(U).$$

Probabilities – some recap and intro

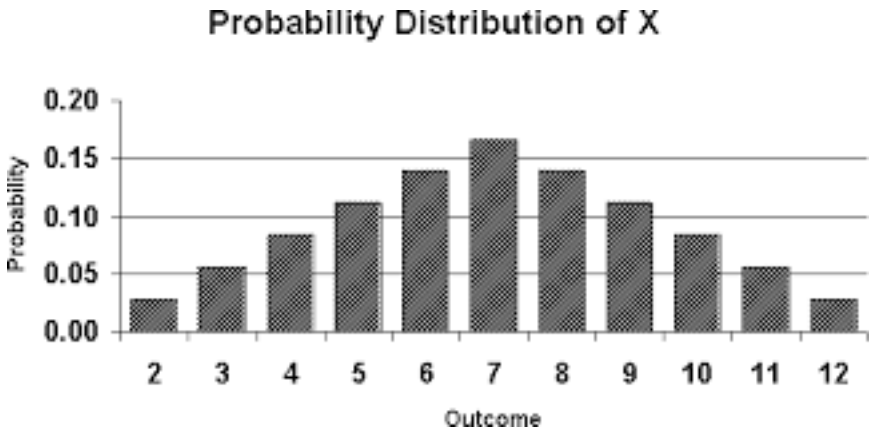
A random variable $\mathbf{X}$ denotes a quantity that is uncertain, for example:

- Rolling a die, flipping a coin
- Measuring temperature

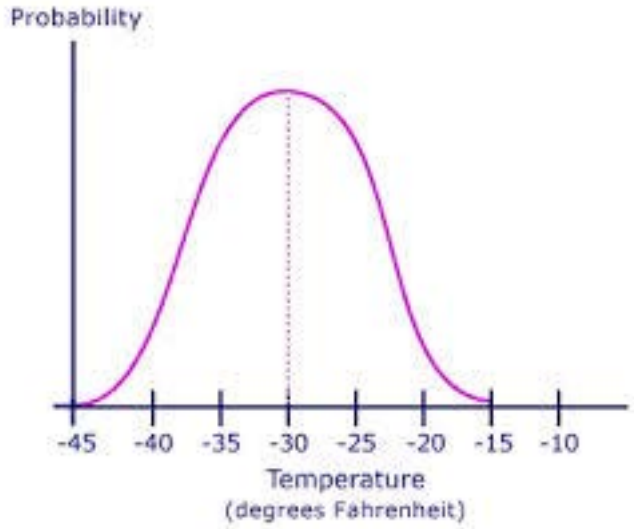
The probability distribution $\mathbf{P(x)}$ of a random variable describes that the variable will have different values when observed, and also which values we are more likely to get.

Discrete vs Continuous

Random Variables can be discrete or continuous



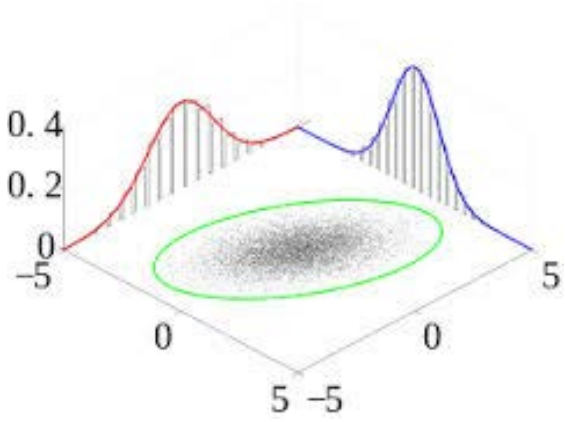
A boolean PDF has one of two values



Joint Probability Distributions

Assume that you have two random variables x and y

If you observe multiple paired instances of x and y, this is described in the joint probability density function $P(x,y)$.



By summing (integrating) of one of the variables (e.g. y) you get the PDF for the other (e.g. x)

$$P(x) = \int_y P(x,y)\,dy$$

Conditional Probability

For a joint probability distribution $P(x,y)$ the conditional probability denoted $P(X|Y)$ is the probability of getting the value x, when we know $y=y^*$

$$P(x \mid y = y^*) = \frac{P(x, y = y^*)}{P(y = y^*)} = \frac{P(x, y = y^*)}{\int_x P(x, y = y^*)\, dx}$$

Classification probability

To simplify notation, we say that

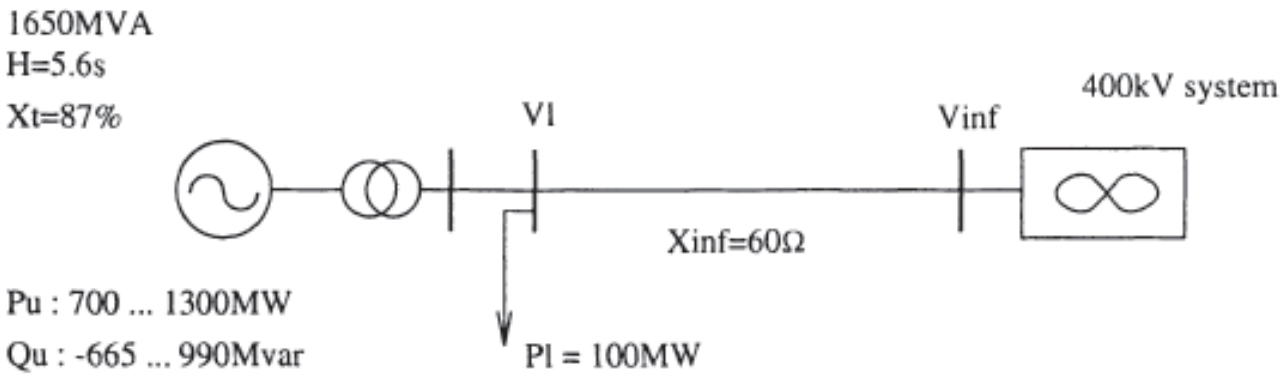
$$P^{i}(\boldsymbol{X}) \triangleq P(\boldsymbol{X}|\boldsymbol{C}_{i}).$$

Denotes the probability P_i of finding X in a given class C_i

Remember, Classes are usually few (1-3), like safe – unsafe.

Contents

Example from Automatic Learning techniques in Power Systems



One Machine Infinite Bus (OMIB) system

- Assuming a fault close to the Generator will be cleared within 155 ms by protection relays
- We need to identify situations in which this clearing time is sufficient and when it is not
- Under certain loading situations, 155 ms may be too slow.

Source: Automatic Learning techniques in Power Systems, L. Wehenkel

OMIB – further information

In the OMIB system the following parameters influence security

- Amount of active and reactive power of the generator (
- Amount of load nearby the generator (PI)
- Voltage magnitudes at the load bus and at the infinite bus Short-circuit reactance Xinf, representing the effect of variable topology in the large system represented by the infinite bus.

In the example, Voltages at generator and Infinite bus are assumed similar and constant for simplicity

Source: Automatic Learning techniques in Power Systems, L. Wehenkel

Our database of objects with attributes

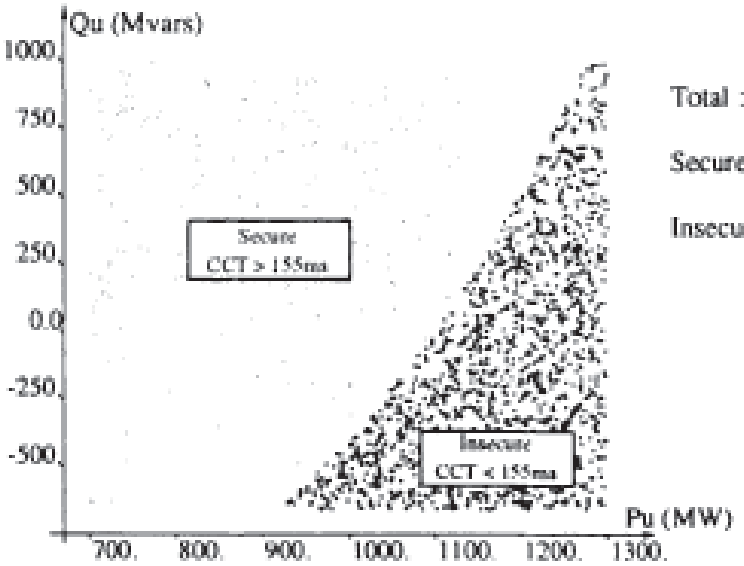
In a simulator, we randomly sample values for P_u and Q_u creating a database with 5000 samples (objects) and for each object we have a set of attributes $(P_u, Q_u, V_1, P_1, V_{inf}, X_{inf}, CCT)$ as per below.

Table 1.1. Sample of OMIB operating states

State Nb	Pu (MW)	Qu (MVAr)	Vl (p.u.)	Pl (MW)	Vinf (p.u.)	Xinf (Ω)	CCT (s)
1	876.0	-193.7	1.05	-100	1.05	60	0.236
2	1110.9	-423.2	1.05	-100	1.05	60	0.112
3	980.1	79.7	1.05	-100	1.05	60	0.210
4	974.1	217.1	1.05	-100	1.05	60	0.224
5	927.2	-618.5	1.05	-100	1.05	60	0.158
...	...	...	...	...	...	...	...
2276	1090.4	-31.3	1.05	-100	1.05	60	0.157
...	...	...	...	...	...	...	...
4984	1090.2	-20.0	1.05	-100	1.05	60	0.158
...	...	...	...	...	...	...	...

Source: Automatic Learning techniques in Power Systems, L. Wehenkel

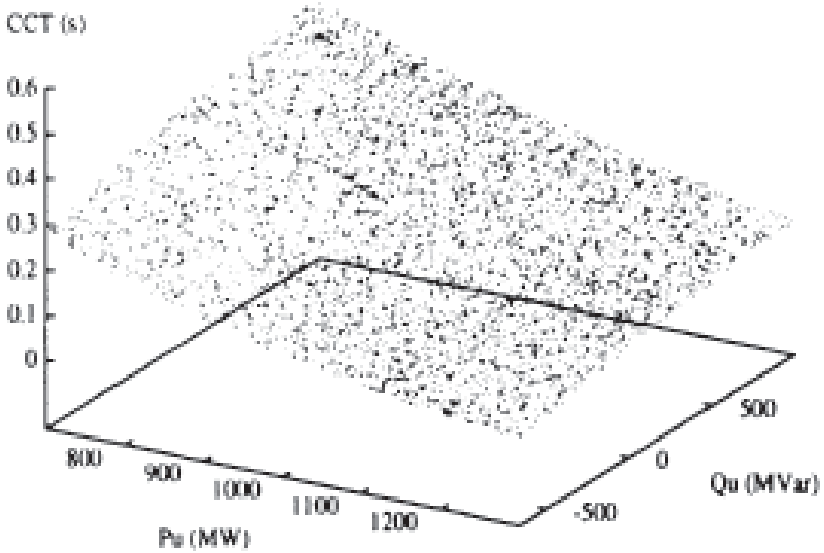
Plot of database content



Total : 5000 states

Secure : 3510 states

Insecure : 1490 states



Source: Automatic Learning techniques in Power Systems, L. Wehenkel

Our task

By using the sampled datapoints (5000 total) our task is now to determine a model that maps a the input paramters to a classification of the system being Secure or Unsecure.

To start with, we select from the samples

1. A Learning set (LS) which are the objects (with attributes) that we will use to create the model
2. A Test set (TS) which are the objects (with attributes) we will use to validate the model

Contents

How to measure information content

Entropy $\mathbf{H}$ is a measure of *Unpredictability*.

Defined as:

$$H = -\sum p_i \log p_i$$

Where

p_i is the probability of event i

Some examples

1. Flipping a coin
2. Rolling a 6 sided die
3. Rolling a loaded 6 sided die

Entropy in a Dataset

$$H_C(\boldsymbol{X}) \triangleq -\sum_{i=1,\ldots,m} P(\boldsymbol{C}_i|\boldsymbol{X}) \log P(\boldsymbol{C}_i|\boldsymbol{X}),$$

The classification Entropy, is the entropy related to the probability of a value x belonging to a class Ci

Or simply put, how difficult is it to guess which partition of U, i.e. Class Ci that an object o belongs to.

An example of classification entropy

Color	Size	Shape	Eadible?
Yellow	Small	Round	Yes
Yellow	Small	Round	No
Green	Small	Irregular	Yes
Green	Large	Irregular	No
Yellow	Large	Round	Yes
Yellow	Small	Round	Yes
Yellow	Small	Round	Yes
Yellow	Small	Round	Yes
Green	Small	Round	No
Yellow	Large	Round	No
Yellow	Large	Round	Yes
Yellow	Large	Round	No
Yellow	Large	Round	No
Yellow	Large	Round	No
Yellow	Small	Irregular	Yes
Yellow	Large	Irregular	Yes

Source: F. Aiolli - Sistemi Informativi University of Padova

Entropy example

Entropy for the example data set is calculated as:

$$I(all\_data) = -\left[\left(\frac{9}{16}\right)\log_2\left(\frac{9}{16}\right) + \left(\frac{7}{16}\right)\log_2\left(\frac{7}{16}\right)\right]$$

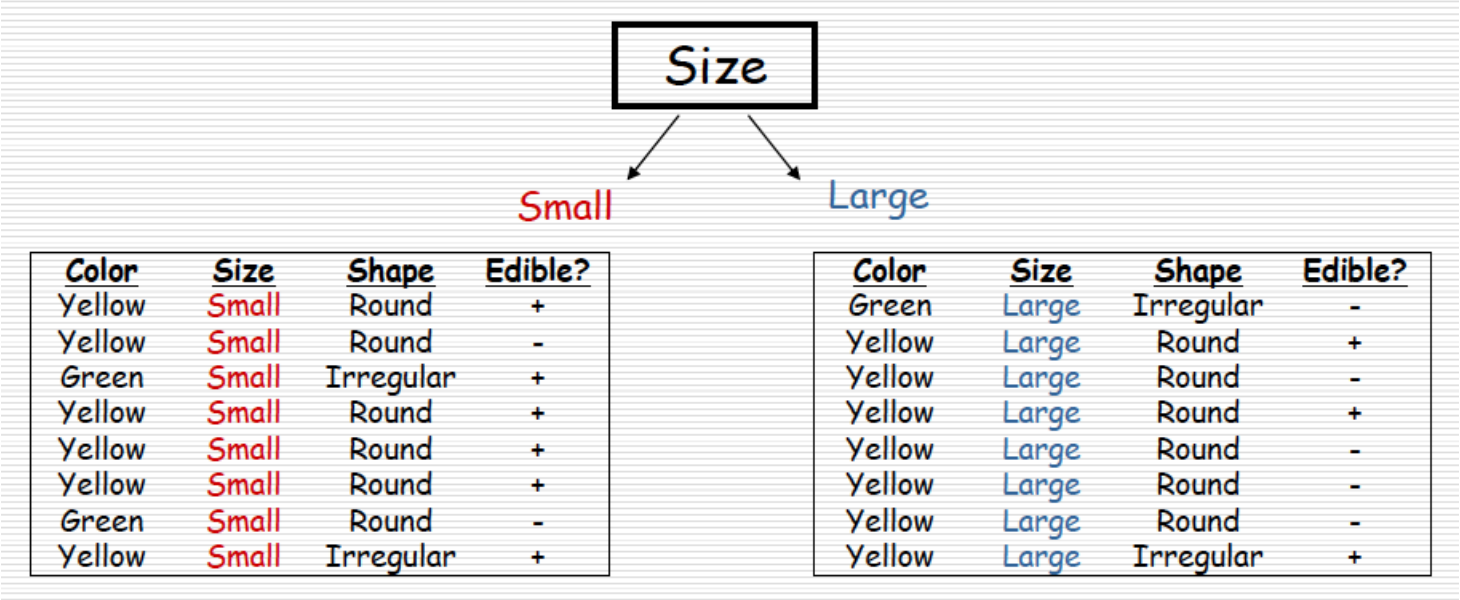
Giving: 0,9836

Is this reasonable?

Information Gain

The reduction in Entropy achieved by partitioning the dataset differently.

Lets separate for instance per the attribute Size.



Color	Size	Shape	Edible?
Yellow	Small	Round	+
Yellow	Small	Round	-
Green	Small	Irregular	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Green	Small	Round	-
Yellow	Small	Irregular	+

Color	Size	Shape	Edible?
Green	Large	Irregular	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Irregular	+

Source: F. Aiolli - Sistemi Informativi University of Padova

Information Gain calculation

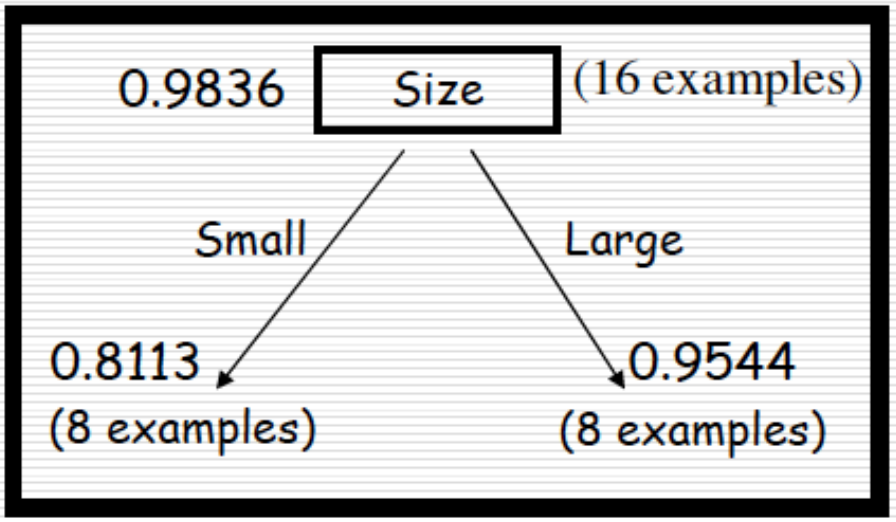
The two partitions has their own entropy value.

We can calculate for each possible attribute its expected entropy. This is the degree to which the entropy would change if partitioned based on this attribute.

To determine resulting entropy, you add the entropies of the two partitions, weighted by the proportion of examples from the parent node that ended up in that partition.

$$G(S,A) = I(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v)$$

Hence for the example



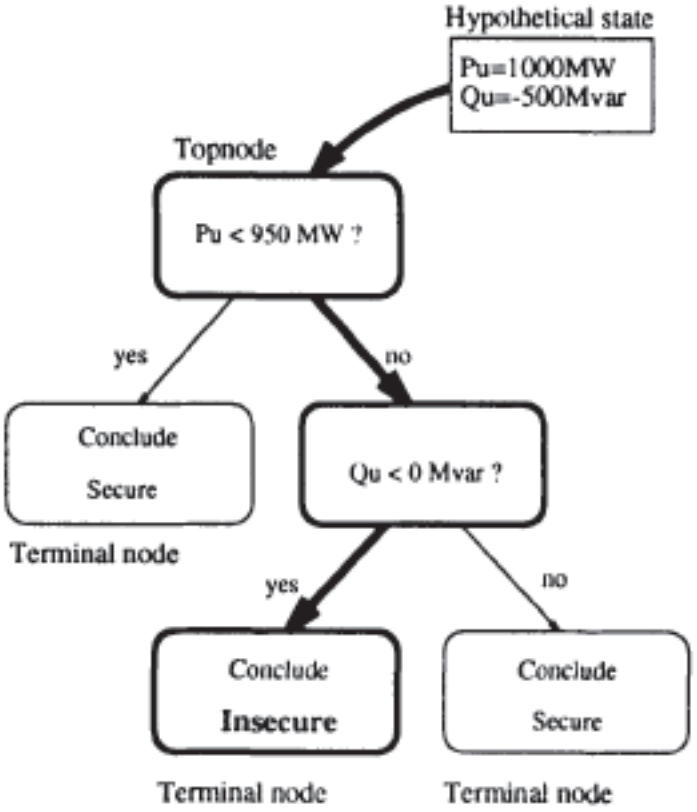
Entropy of left child is 0.8113
I(size=small) = 0.8113

Entropy of right child is 0.9544
I(size=large) = 0.9544

$$I(S_{Size}) = (8/16)*.8113 + (8/16)*.9544 = .8828$$

Contents

Back to our Power System example



Perhaps we can partition our dataset according to some attribute?

Lets try $P_u < 950MW$

Equivalent If-Then rules :

Rule 1 : If (Pu < 950MW) then Conclude Secure

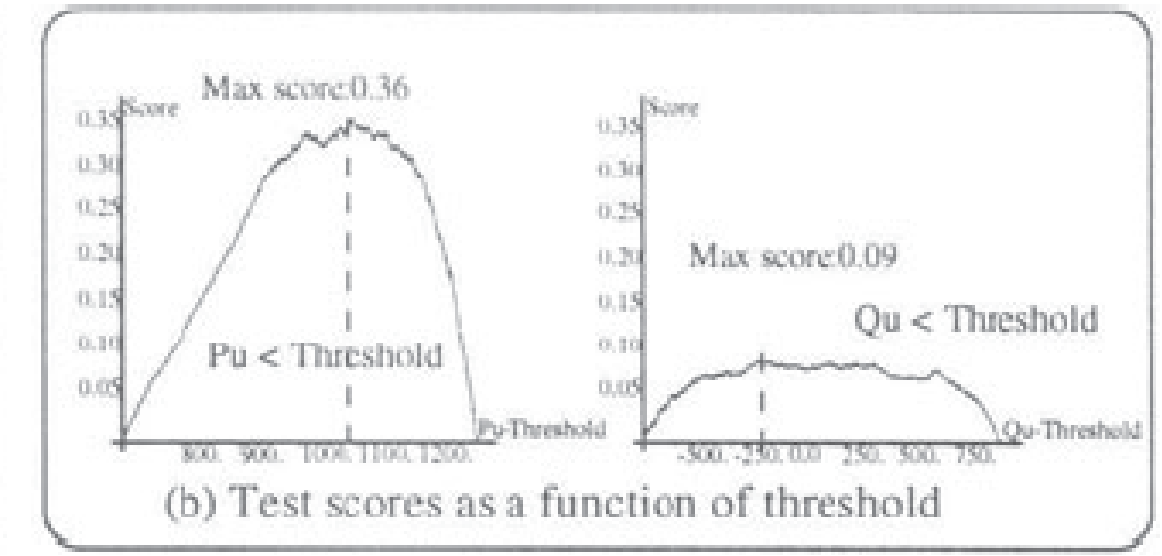
Rule 2 : If (Pu > 950MW) and (Qu < 0Mvar) then Conclude Insecure

Rule 3 : If (Pu > 950MW) and (Qu > 0Mvar) then Conclude Secure

Finding best partition.

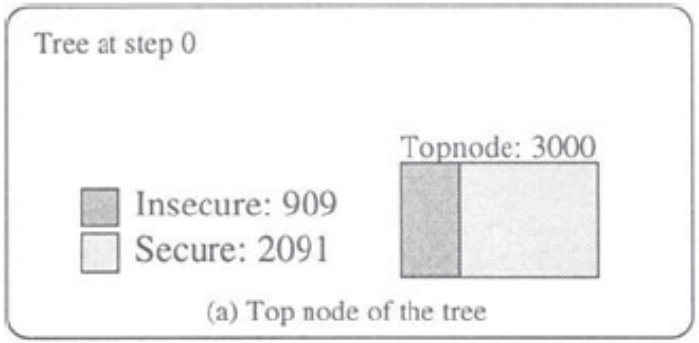
Starting with the candidate attributes (Pu and Qu) in our case

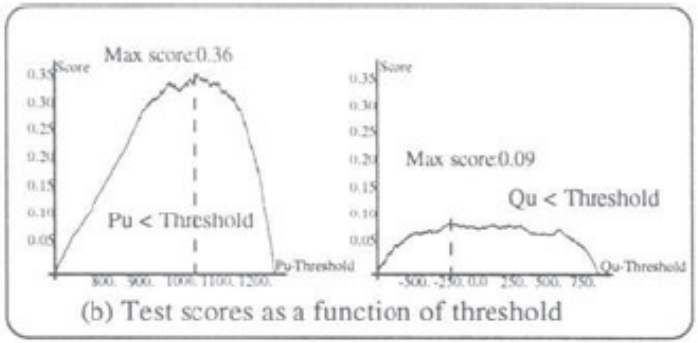
We check chich of the values for Pu and Qu that create the most valuable partition in terms of information gain.

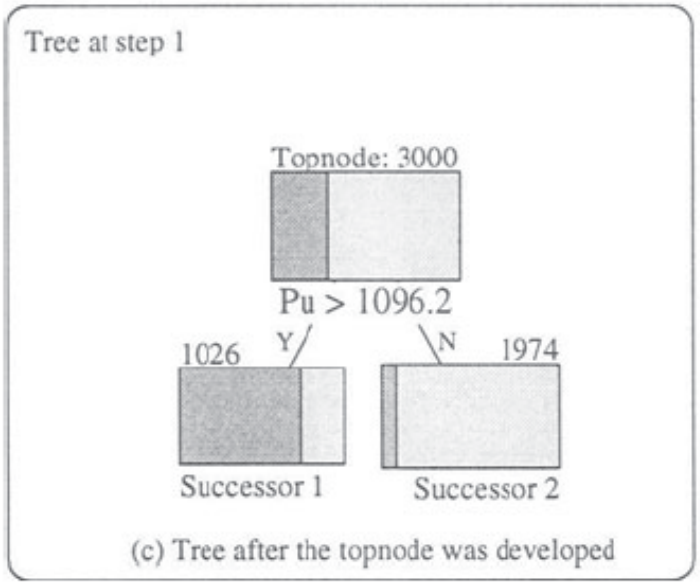


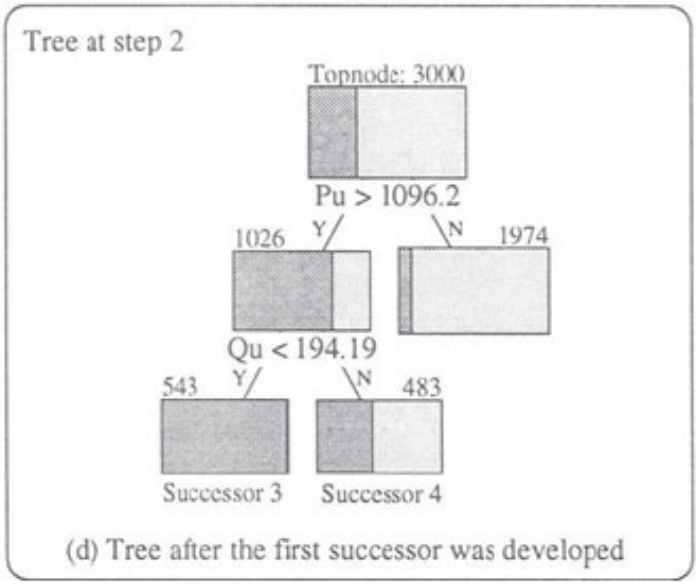
Pu > 1096,2 MW is the best partition

Gradual expansion of the Decision Tree

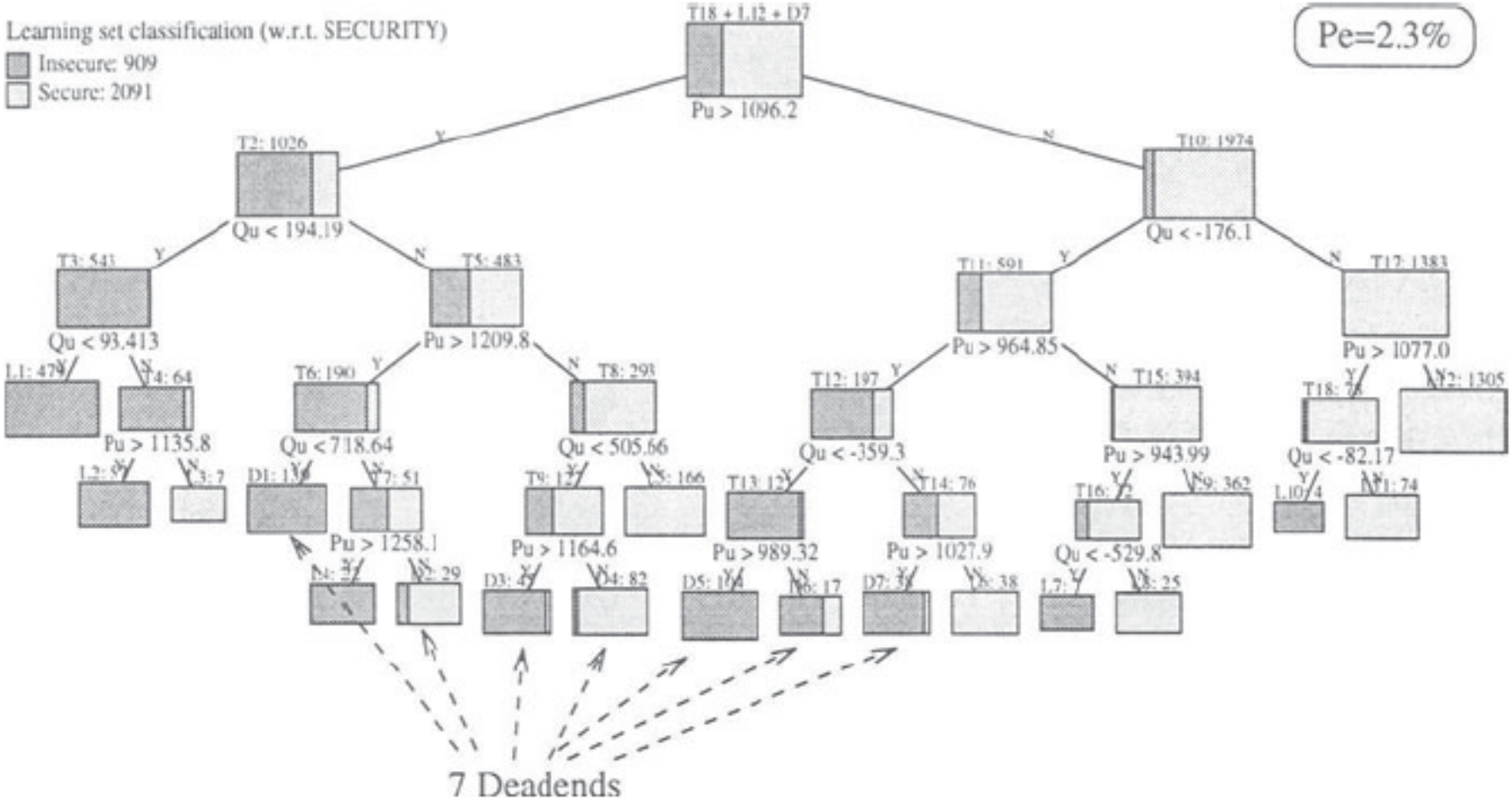








Complete Decision Tree



How to stop?

The splitting of data sets continues until either:

A perfect partition is reached – i.e. One which perfecly explains the content of the class – a *leaf*

One where no infomration is gained no matter how the data set is split. – a *deadend*.

Validation of the Decision Tree

By using the Test Set (2000 samples) we can calidate the Decision tree.

By testing for each Object in the Test Set, we determine if the Decision tree provides the right answer for the Object.

In this particular example, the probability o error can be determined to 2,3. I.e. Of the 2000 samples 46 were classifiifed to the wrong class.

Contents

Machine Learning vs Systems Theory

Some definitions

An illustrative example

Information content – entropy

Decision Trees