Machine Learning and Pervasive Computing

Stephan Sigg

Georg-August-University Goettingen, Computer Networks

27.04.2015

Overview and Structure

13.04.2015	Organisation
13.04.2015	Introduction
20.04.2015	Rule-based learning
27.04.2015	Decision Trees
04.05.2015	A simple Supervised learning algorithm
11.05.2015	–
18.05.2015	Excursion: Avoiding local optima with random search
25.05.2015	–
01.06.2015	k-Nearest Neighbour methods
08.06.2015	High dimensional data
15.06.2015	Artificial Neural Networks
22.06.2015	Probabilistic models
29.06.2015	Topic models
06.07.2015	Unsupervised learning
13.07.2015	Anomaly detection, Online learning, Recom. systems

Outline

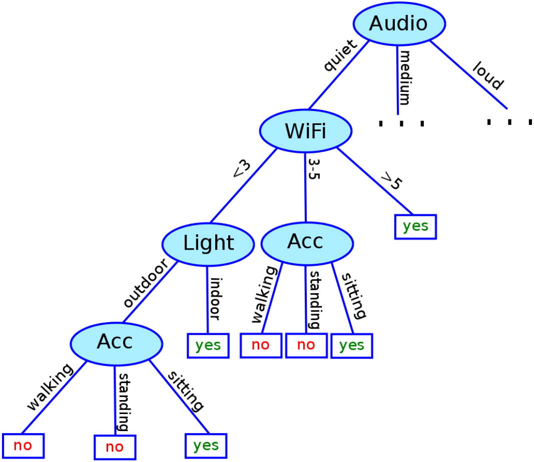
Decision Tree

C4.5

Confidence on a prediction

Decision tree

A decision tree is a tree that divides the examples from a dataset according to the features and classes observed for them



Decision tree

How to generate such decision tree?

Decision tree

How to generate such decision tree?

First select a feature to split on and place it at the root node.

Then repeat this procedure for all child nodes

Decision tree

How to generate such decision tree?

First select a feature to split on and place it at the root node.

Then repeat this procedure for all child nodes

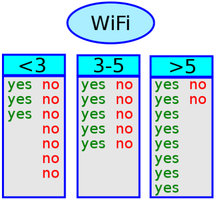
How to determine the feature to split on?

Decision tree

WiFi	yes	no	Accelerometer	yes	no	Audio	yes	no	Light	yes	no	At work yes	At work no
<3 APs	3	7	walking	4	8	quiet	8	5	outdoor	4	7	16	14
[3, 5]	5	5	standing	1	4	medium	6	3	indoor	12	7		
>5 APs	8	2	sitting	11	2	loud	2	6					

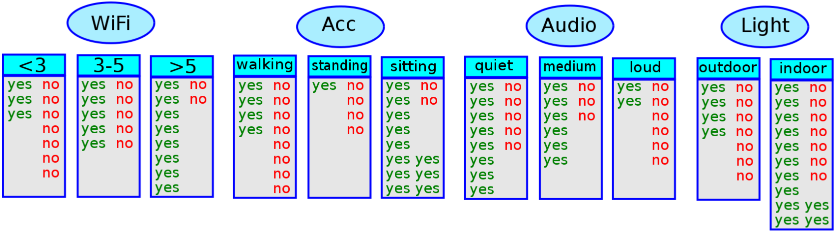
Decision tree

WiFi			Accelerometer			Audio			Light			At work	
	yes	no		yes	no		yes	no		yes	no	yes	no
<3 APs	3	7	walking	4	8	quiet	8	5	outdoor	4	7	16	14
[3, 5]	5	5	standing	1	4	medium	6	3	indoor	12	7		
>5 APs	8	2	sitting	11	2	loud	2	6					



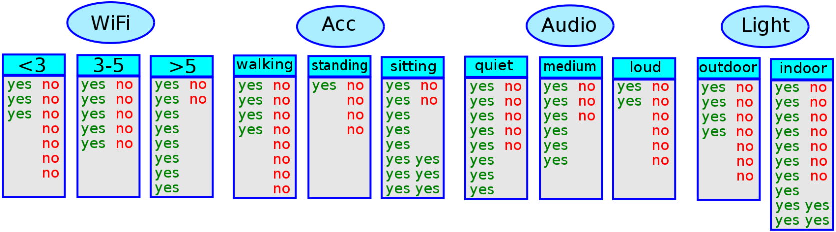
Decision tree

WiFi			Accelerometer			Audio			Light			At work	
	yes	no		yes	no		yes	no		yes	no	yes	no
<3 APs	3	7	walking	4	8	quiet	8	5	outdoor	4	7	16	14
[3, 5]	5	5	standing	1	4	medium	6	3	indoor	12	7		
>5 APs	8	2	sitting	11	2	loud	2	6					



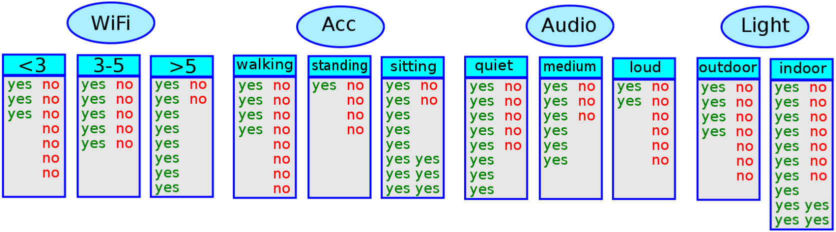
Which one is the best choice?

Decision tree



We are interested in the gain in information when a particular choice is taken

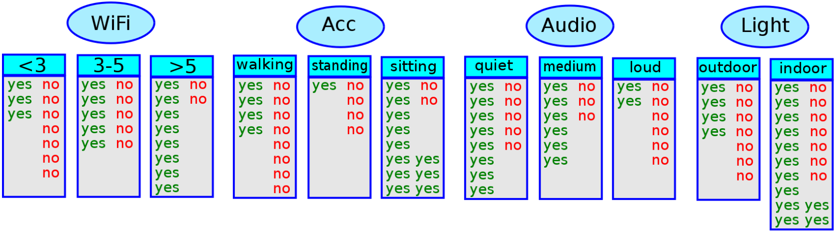
Decision tree



We are interested in the gain in information when a particular choice is taken

The decision tree should then decide for the split that promises maximum information gain.

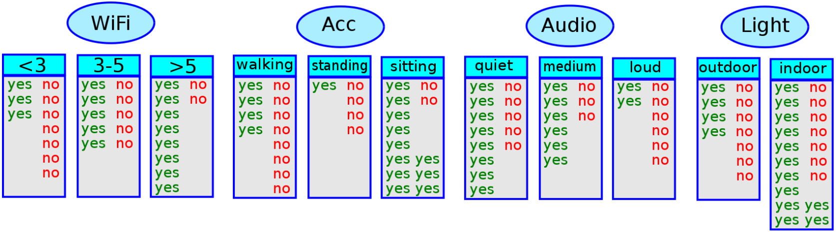
Decision tree



This can be estimated by the entropy of a value:

$$\mathcal{E}(p_1, p_2, \ldots, p_n) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \cdots - p_n \log_2 p_n$$

Decision tree

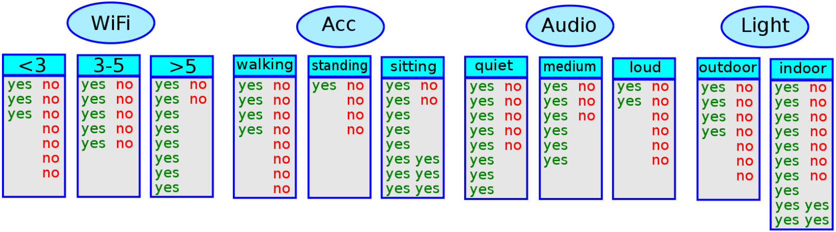


$$\mathcal{E}(p_1, p_2, \ldots, p_n) = \quad -p_1 \log_2 p_1 - p_2 \log_2 p_2 \cdots - p_n \log_2 p_n$$

WiFi information value:

$$\mathcal{E}\left(\frac{3}{10}, \frac{7}{10}\right)\frac{10}{30} + \mathcal{E}\left(\frac{5}{10}, \frac{5}{10}\right)\frac{10}{30} + \mathcal{E}\left(\frac{8}{10}, \frac{2}{10}\right)\frac{10}{30} =$$

Decision tree



$$\mathcal{E}(p_1, p_2, \ldots, p_n) = \quad -p_1 \log_2 p_1 - p_2 \log_2 p_2 \cdots - p_n \log_2 p_n$$

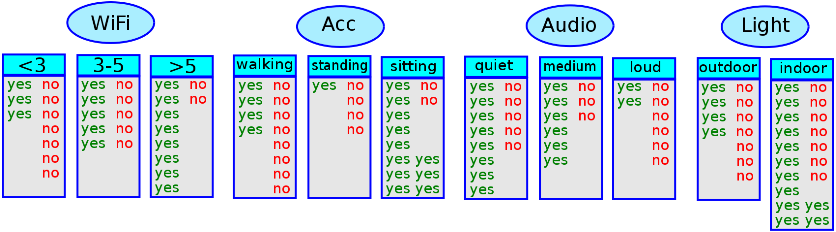
WiFi information value:

$$\mathcal{E}\left(\frac{3}{10}, \frac{7}{10}\right)\frac{10}{30} + \mathcal{E}\left(\frac{5}{10}, \frac{5}{10}\right)\frac{10}{30} + \mathcal{E}\left(\frac{8}{10}, \frac{2}{10}\right)\frac{10}{30} = \quad \left(-\tfrac{3}{10}\log_2 \tfrac{3}{10} - \tfrac{7}{10}\log_2 \tfrac{7}{10}\right)\cdot \tfrac{10}{30}$$

$$+\left(-\tfrac{5}{10}\log_2 \tfrac{5}{10} - \tfrac{5}{10}\log_2 \tfrac{5}{10}\right)\cdot \tfrac{10}{30}$$

$$+\left(-\tfrac{8}{10}\log_2 \tfrac{8}{10} - \tfrac{2}{10}\log_2 \tfrac{2}{10}\right)\cdot \tfrac{10}{30}$$

Decision tree

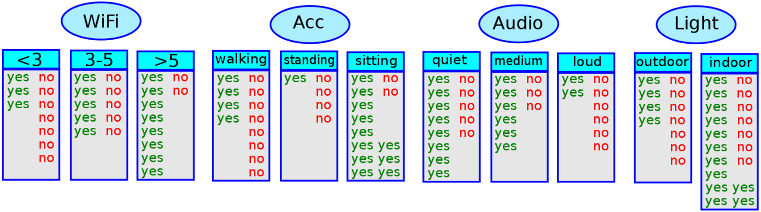


$$\mathcal{E}(p_1, p_2, \ldots, p_n) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \cdots - p_n \log_2 p_n$$

WiFi information value:

$$\mathcal{E}\left(\frac{3}{10}, \frac{7}{10}\right)\frac{10}{30} + \mathcal{E}\left(\frac{5}{10}, \frac{5}{10}\right)\frac{10}{30} + \mathcal{E}\left(\frac{8}{10}, \frac{2}{10}\right)\frac{10}{30} = \left(-\tfrac{3}{10}\log_2\tfrac{3}{10} - \tfrac{7}{10}\log_2\tfrac{7}{10}\right)\cdot\tfrac{10}{30}$$
$$+\left(-\tfrac{5}{10}\log_2\tfrac{5}{10} - \tfrac{5}{10}\log_2\tfrac{5}{10}\right)\cdot\tfrac{10}{30}$$
$$+\left(-\tfrac{8}{10}\log_2\tfrac{8}{10} - \tfrac{2}{10}\log_2\tfrac{2}{10}\right)\cdot\tfrac{10}{30}$$
$$\approx 0.868$$

Decision tree



Information value:

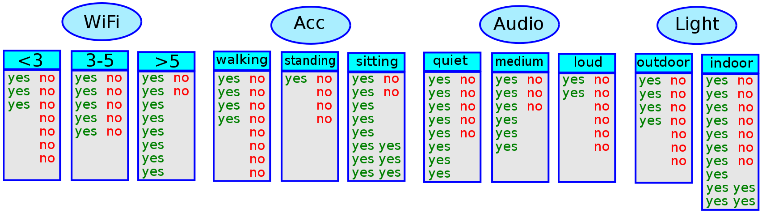
WiFi: $\approx$ 0.868

Acc: $\approx$ ...

Audio: $\approx$ ...

Light: $\approx$ ...

Decision tree



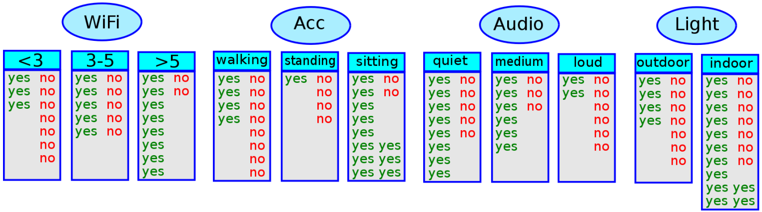
Information value:

WiFi:	$\approx$	0.868
Acc:	$\approx$	0.756
Audio:	$\approx$	0.884
Light:	$\approx$	0.948

Information gain:

Initial information value (working [yes/no]): 0.997

Decision tree



Information value:

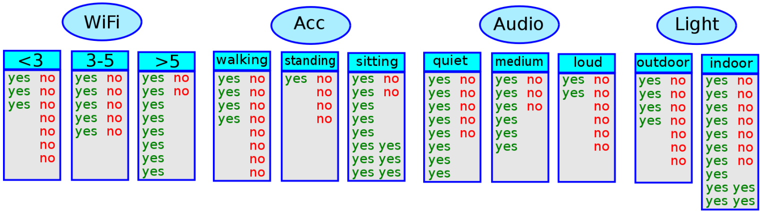
WiFi:	$\approx$	0.868
Acc:	$\approx$	0.756
Audio:	$\approx$	0.884
Light:	$\approx$	0.948

Information gain:

WiFi:	$\approx$	0.129
Acc:	$\approx$	0.241
Audio:	$\approx$	0.113
Light:	$\approx$	0.049

Initial information value (working [yes/no]): 0.997

Decision tree



Information value:

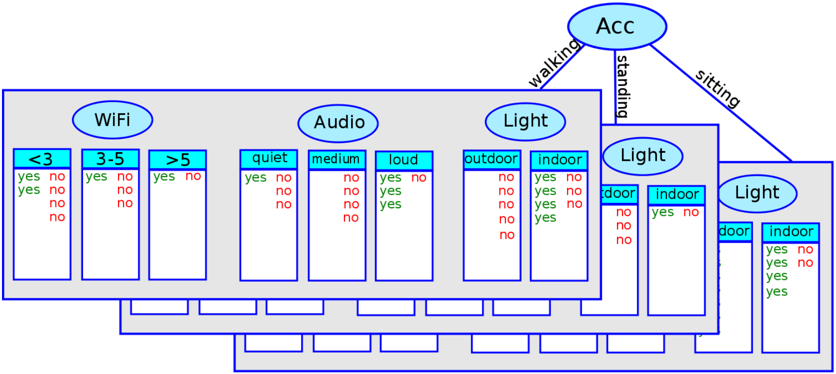
WiFi:	$\approx$	0.868
Acc:	$\approx$	**0.756**
Audio:	$\approx$	0.884
Light:	$\approx$	0.948

Information gain:

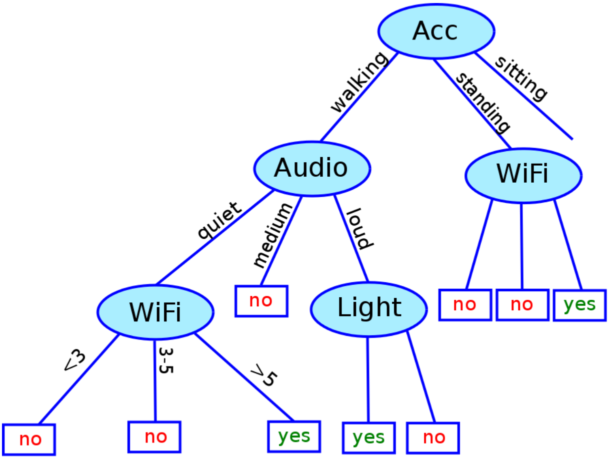
WiFi:	$\approx$	0.129
Acc:	$\approx$	**0.241**
Audio:	$\approx$	0.113
Light:	$\approx$	0.049

Initial information value (working [yes/no]): 0.997

Decision tree



Decision tree



Outline

Decision tree – C4.5

Improved decision tree implementation: C4.5

- Dealing with numeric values
- Missing values
- Noisy data

C4.5 – Dealing with numeric values

Nominal feature values

For nominal features, the decision tree splits on every possible value. Therefore, the information content of this feature is 0 after such branch has been conducted

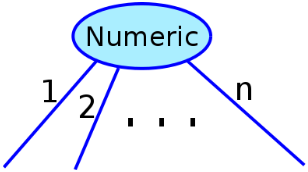
C4.5 – Dealing with numeric values

Nominal feature values

For nominal features, the decision tree splits on every possible value. Therefore, the information content of this feature is 0 after such branch has been conducted

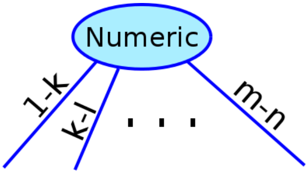
Numeric feature values

For numeric feature values, splitting on each possible value would lead to a very wide tree of small depth.



C4.5 – Dealing with numeric values

For numeric values, the tree is split into several intervals.



C4.5 – Missing values

Missing values in a data set

Missing values are a common/prominent event in real-world data sets

- participants in a survey refuse to answer
- malfunctioning sensor nodes
- Biology: plants or animals might die before all variables have been measured
- ...

Most machine learning schemes make the implicit assumption that there is no significance in the fact that a certain value is missing.

C4.5 – Missing values

Missing values in a data set

Missing values are a common/prominent event in real-world data sets

- participants in a survey refuse to answer
- malfunctioning sensor nodes
- Biology: plants or animals might die before all variables have been measured
- ...

Most machine learning schemes make the implicit assumption that there is no significance in the fact that a certain value is missing.

The absence of data might already hold valuable information!

C4.5 – Missing values

The absence of data might already hold valuable information!

[1] Witten et al., Data Mining, Morgan Kaufmann, 2011

C4.5 – Missing values

The absence of data might already hold valuable information!

Example

People analyzing medical databases have noticed that cases may, in some circumstances, be diagnosable simply from the tests that a doctor decides to make – regardless of the outcome of the tests[1]

[1] Witten et al., Data Mining, Morgan Kaufmann, 2011

C4.5 – Missing values

Possible solution

Considering whether the sets of samples with values have significant difference in their final outcome when compared to the sets of samples with missing values

C4.5 – Noisy data

Fully expanded decision trees often contain unnecessary structure that should be simplified before deployment

C4.5 – Noisy data

Fully expanded decision trees often contain unnecessary structure that should be simplified before deployment

Pruning

Prepruning Trying to decide through the tree-building process when to stop developing subtrees

- Might speed up tree creation phase
- Difficult to spot dependencies between features at this stage (features might be meaningful together but not on their own)

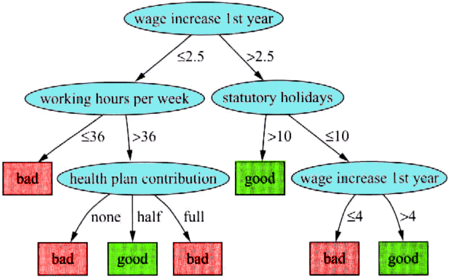
Postpruning Simplification of the decision tree after the tree has been created

C4.5 – Noisy data

Postpruning – subtree replacement

Select some subtrees and replace them with single leaves

- Will reduce accuracy on the training set
- May increase accuracy on independently chosen test set (reduction of noise)

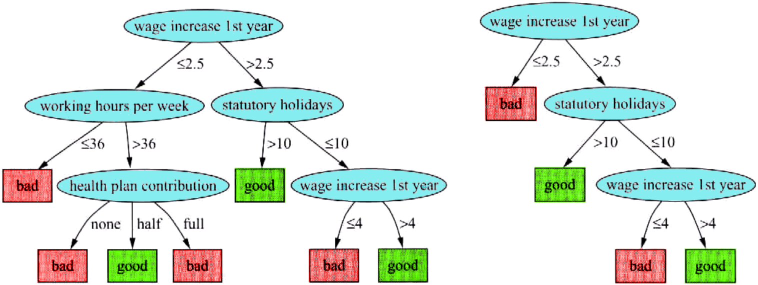


C4.5 – Noisy data

Postpruning – subtree replacement

Select some subtrees and replace them with single leaves

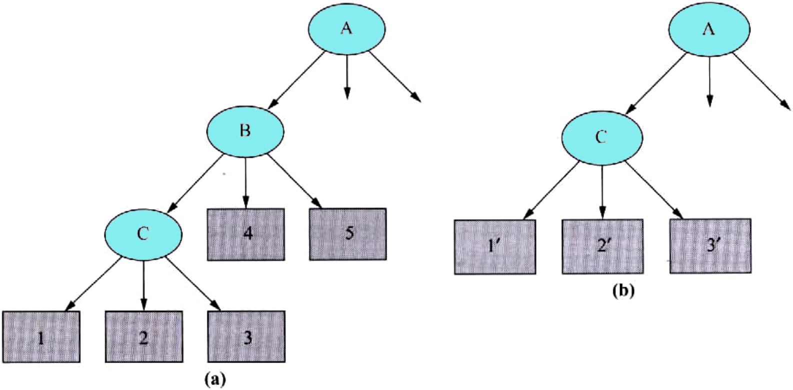
- Will reduce accuracy on the training set
- May increase accuracy on independently chosen test set (reduction of noise)



C4.5 – Noisy data

Postpruning – subtree raising

Complete subtree is raised one level and samples at the nodes of the subtree have to be recalculated



C4.5 – Estimating error rates

When should we raise or replace subtrees?

C4.5 – Estimating error rates

When should we raise or replace subtrees?

Estimating error rates

Estimation of error rates at internal nodes and leaf nodes.

C4.5 – Estimating error rates

When should we raise or replace subtrees?

Estimating error rates

Estimation of error rates at internal nodes and leaf nodes.

Assumption: Label of node is chosen as the majority vote from all its leaves

C4.5 – Estimating error rates

When should we raise or replace subtrees?

Estimating error rates

Estimation of error rates at internal nodes and leaf nodes.

Assumption: Label of node is chosen as the majority vote from all its leaves

- Will lead to a certain number of errors E
- … out of the total number of instances N

C4.5 – Estimating error rates

When should we raise or replace subtrees?

Estimating error rates

Estimation of error rates at internal nodes and leaf nodes.

Assumption: Label of node is chosen as the majority vote from all its leaves

- Will lead to a certain number of errors E
- ... out of the total number of instances N
- Assume:
  1. True probability of error at that node is q
  2. N instances are generated by Bernoulli process with parameter q and errors E

C4.5 – Estimating error rates

When should we raise or replace subtrees?

Bernoulli process

A Bernoulli process is a repeated coin flipping, possibly with an unfair coin

C4.5 – Estimating error rates

Estimating error rates – Calculating the success probability

Given a confidence c (C4.5 uses 25%), we find a confidence limit z (for $c = 25\% \rightarrow z = 0.69$) such that

$$\mathcal{P}\left[\frac{q' - q}{\sqrt{\frac{q(1-q)}{N}}} > z\right] = c$$

(with the observed error rate $q' = \frac{E}{N}$)

C4.5 – Estimating error rates

Estimating error rates – Calculating the success probability

Given a confidence c (C4.5 uses 25%), we find a confidence limit z (for $c = 25\% \rightarrow z = 0.69$) such that

$$\mathcal{P}\left[\frac{q' - q}{\sqrt{\frac{q(1-q)}{N}}} > z\right] = c$$

(with the observed error rate $q' = \frac{E}{N}$)

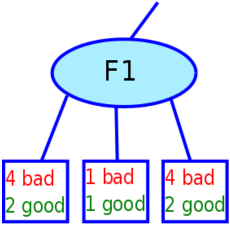
- This leads to an upper confidence limit for q which we can use to estimate a pessimistic error rate e

$$e = \frac{q' + \frac{z^2}{2N} + z\sqrt{\frac{q'}{N} - \frac{q'^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}$$

C4.5 – Estimating error rates

Example

Lower left leaf $(E = 2, N = 6)$ Utilising the formula for e, we obtain
$q' = 0.33$ and $e = 0.47$

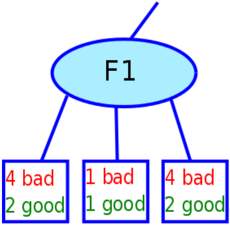


C4.5 – Estimating error rates

Example

Lower left leaf $(E = 2, N = 6)$ Utilising the formula for e, we obtain $q' = 0.33$ and $e = 0.47$

Center leaf$(E = 1, N = 2)$ $e = 0.72$



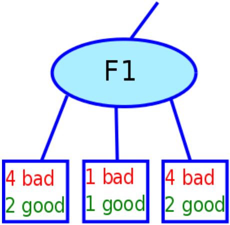
C4.5 – Estimating error rates

Example

Lower left leaf ($E = 2, N = 6$) Utilising the formula for e, we obtain $q' = 0.33$ and $e = 0.47$

Center leaf($E = 1, N = 2$) $e = 0.72$

Right leaf ($E = 2, N = 6$) $e = 0.47$



C4.5 – Estimating error rates

Example

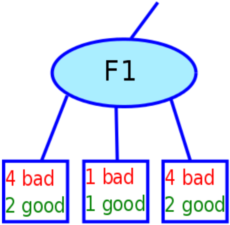
Lower left leaf ($E = 2, N = 6$) Utilising the formula for e, we obtain $q' = 0.33$ and $e = 0.47$

Center leaf($E = 1, N = 2$) $e = 0.72$

Right leaf ($E = 2, N = 6$) $e = 0.47$

Combine Eror estimates Utilising ratio 6:2:6 this leads to a combined error estimate of

$$\frac{0.47 \cdot 6}{14} + \frac{0.72 \cdot 2}{14} + \frac{0.47 \cdot 6}{14} \approx 0.51$$



C4.5 – Estimating error rates

Example

Lower left leaf $(E = 2, N = 6)$ Utilising the formula for e, we obtain $q' = 0.33$ and $e = 0.47$

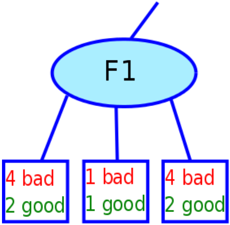
Center leaf$(E = 1, N = 2)$ $e = 0.72$

Right leaf $(E = 2, N = 6)$ $e = 0.47$

Combine Eror estimates Utilising ratio 6:2:6 this leads to a combined error estimate of

$$\frac{0.47 \cdot 6}{14} + \frac{0.72 \cdot 2}{14} + \frac{0.47 \cdot 6}{14} \approx 0.51$$

Error estimate for parent node $q' = \frac{5}{14} \rightarrow e = 0.46$
$0.46 < 0.51 \Rightarrow$ prune children away



C4.5 – Further heuristics employed

Postpruning – Confidence value $c = 25\%$

Postpruning – Split Threshold Candidate splits on a numeric feature are only considered when at least $\min(10\%, 25)$ of all training samples are cut off by the split

Prepruning with information gain Given S candidate splits on a certain numeric attribute, $\log_2 \frac{S}{N}$ is subtracted from the information gain

- in order to prevent overfitting
- When information gain is negative, tree-construction will stop

C4.5 – Remarks

Postpruning

- Postpruning in C4.5 is very fast and therefore popular
- However, the statistical assumptions are shaky
  - use of upper confidence limit
  - assumption of normal distribution for error rate calculation
  - use of statistics from the training set
- Often, the algorithm does not prune enough and a better performance can be achieved with a more compact decision tree

Outline

Confidence on a prediction

Assume we measure the error of a classifier on a test set and obtain a numerical error rate of q (a success rate of $p = (1 - q)$).

What can we say about the true success rate?

- It will be close to p,
- but how close? (within 5% or 10% ?)

This depends on the size of the test set

Naturally, we are more confident on the success probability p when it were based on a large number of values.

Confidence on a prediction

In statistics, a succession of independent events that either succeed or fail is called a Bernoulli process

Confidence on a prediction

In statistics, a succession of independent events that either succeed or fail is called a Bernoulli process

Assume that out of N events, S are successful.

Confidence on a prediction

In statistics, a succession of independent events that either succeed or fail is called a Bernoulli process

Assume that out of N events, S are successful.

Then we have an observed success rate of $p' = \frac{S}{N}$

Confidence on a prediction

In statistics, a succession of independent events that either succeed or fail is called a Bernoulli process

Assume that out of N events, S are successful.

Then we have an observed success rate of $p' = \frac{S}{N}$

What can we say about the true success rate p?

Confidence on a prediction

In statistics, a succession of independent events that either succeed or fail is called a Bernoulli process

Assume that out of N events, S are successful.

Then we have an observed success rate of $p' = \frac{S}{N}$

What can we say about the true success rate p?

Confidence Interval

The answer is expressed as a confidence interval:
p lies within an interval with a specified confidence

Confidence on a prediction

For a specific Bernoulli trial with success rate p we have

mean p

variance $p(1-p)$

For large N, the distribution of this random variable approaches the normal distribution

Confidence on a prediction

The probability that a random variable χ, with zero mean, lies within a certain confidence range of width $2z$ is

$$\mathcal{P}[-z \leq \chi \leq z] = c$$

Confidence on a prediction

The probability that a random variable χ, with zero mean, lies within a certain confidence range of width $2z$ is

$$\mathcal{P}[-z \leq \chi \leq z] = c$$

Confidence limits for the normal distribution are e.g.

$\mathcal{P}[\chi \geq z]$	0.001	0.005	0.01	0.05	0.1	0.2	0.4
z	3.09	2.58	2.33	1.65	1.28	0.84	0.25

Confidence on a prediction

The probability that a random variable χ, with zero mean, lies within a certain confidence range of width $2z$ is

$$\mathcal{P}[-z \leq \chi \leq z] = c$$

Confidence limits for the normal distribution are e.g.

$\mathcal{P}[\chi \geq z]$	0.001	0.005	0.01	0.05	0.1	0.2	0.4
z	3.09	2.58	2.33	1.65	1.28	0.84	0.25

Standard assumption in such tables on the random variable:

mean 0
variance 1

Confidence on a prediction

$\mathcal{P}[\chi \geq z]$	0.001	0.005	0.01	0.05	0.1	0.2	0.4
z	3.09	2.58	2.33	1.65	1.28	0.84	0.25

The z figures are measured in standard deviations from the mean:

Confidence on a prediction

$\mathcal{P}[\chi \geq z]$	0.001	0.005	0.01	0.05	0.1	0.2	0.4
z	3.09	2.58	2.33	1.65	1.28	0.84	0.25

The z figures are measured in standard deviations from the mean:

Example

E.g. the figure for $\mathcal{P}[\chi \geq z] = 0.05$ implies that there is a 5% chance that χ lies more than 1.65 standard deviations above the mean.

Confidence on a prediction

$\mathcal{P}[\chi \geq z]$	0.001	0.005	0.01	0.05	0.1	0.2	0.4
z	3.09	2.58	2.33	1.65	1.28	0.84	0.25

The z figures are measured in standard deviations from the mean:

Example

E.g. the figure for $\mathcal{P}[\chi \geq z] = 0.05$ implies that there is a 5% chance that χ lies more than 1.65 standard deviations above the mean.

Since the distribution is symmetric, the chance that X lies more than 1.65 standard deviations from the mean is 10%:

$$\mathcal{P}[-1.65 \leq \chi \leq 1.65] = 0.9$$

Confidence on a prediction

In order to apply this to the random variable p', we have to reduce it to have zero mean and unit variance.

Confidence on a prediction

In order to apply this to the random variable p', we have to reduce it to have zero mean and unit variance.

We do this by subtracting the mean p and by dividing by the standard deviation $\sqrt{\frac{p(1-p)}{N}}$

Confidence on a prediction

In order to apply this to the random variable p', we have to reduce it to have zero mean and unit variance.

We do this by subtracting the mean p and by dividing by the standard deviation $\sqrt{\frac{p(1-p)}{N}}$

This leads to

$$\mathcal{P}\left[-z < \frac{p' - p}{\sqrt{\frac{p(1-p)}{N}}} < z\right] = c$$

Confidence on a prediction

To find confidence limits, given a particular confidence figure c:

- consult a table with confidence limits for the normal distribution for the corresponding value z
- Note: since Success probabilities are displayed, we have to subtract our value c from 1 and divide by two:

$$\frac{1-c}{2}$$

- Then, write the inequality above as an equality and invert it to find an expression for p
- Finally, solving a quadratic equation will produce the respective value for p

Confidence on a prediction

- Finally, solving a quadratic equation will produce the respective value for p

$$p = \frac{\left(p' + \frac{z^2}{2N} \pm z\sqrt{\frac{p'}{N} - \frac{p'^2}{N} + \frac{z^2}{4N^2}}\right)}{1 + \frac{z^2}{N}}$$

The resulting two values are the upper and lower confidence boundaries

Confidence on a prediction

Example

$$p' = 0.75;\ N = 1000,\ c = 0.8\ (z = 1.28) \quad \rightarrow \quad [0.732, 0.767]$$
$$p' = 0.75;\ N = 100,\ c = 0.8\ (z = 1.28) \quad \rightarrow \quad [0.691, 0.801]$$

Note that the assumptions taken are only valid for large N

Outline

Decision Tree

C4.5

Confidence on a prediction

Questions?

Stephan Sigg
stephan.sigg@cs.uni-goettingen.de

Literature

- C.M. Bishop: Pattern recognition and machine learning, Springer, 2007.
- R.O. Duda, P.E. Hart, D.G. Stork: Pattern Classification, Wiley, 2001.