

Decision Trees & Comparing classifiers

Course on Machine Learning

Mannes Poel

Gwenn Englebienne

Announcements

- Reading material for this week can be found on Canvas.

Questions?

- Lab session
- Homework

Use of K-CV

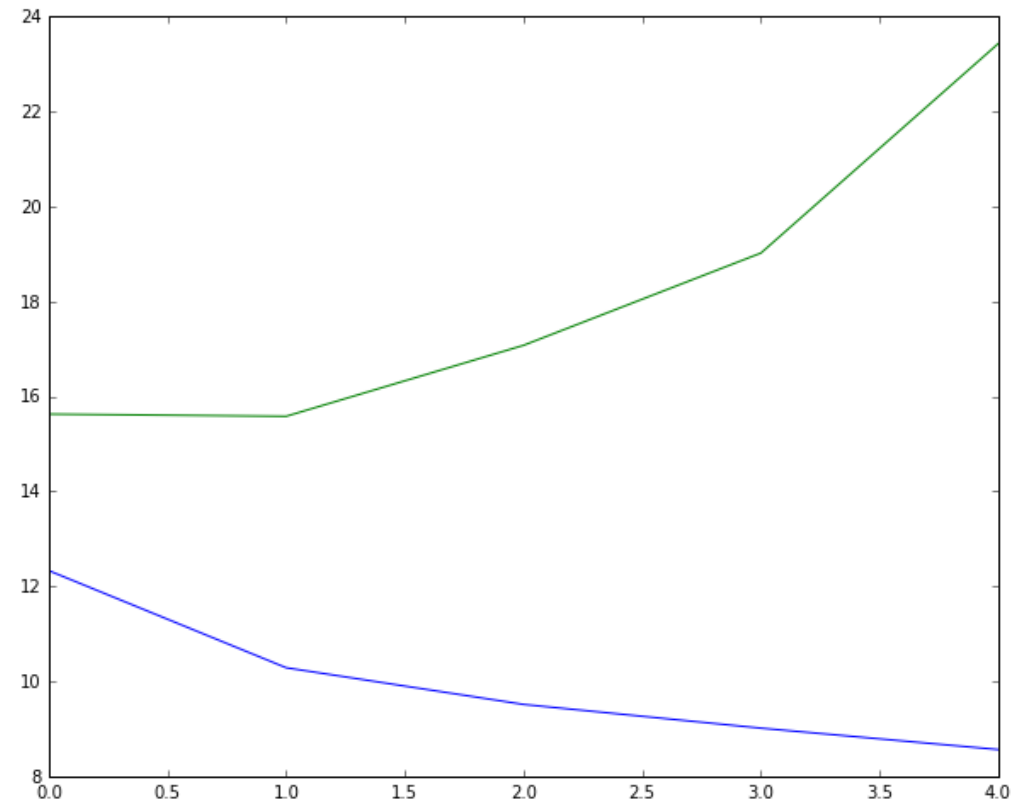
Lab: K-fold can be used to determine good parameter settings, such as the order.

But can also be used for determining other (hyper)parameter settings such as the regularization parameter.

After determining the good parameters one can train the model on the complete trainingset.

5	20													
1	11.5	17	2	10.5	15	3	9.48	21	4	8.86	24	5	8.76	35
1	14.2	7.1	2	12.1	6.4	3	11.1	6.5	4	10.8	7.1	5	10.6	8.8
1	10.4	28	2	7.27	34	3	7.16	34	4	7.09	35	5	5.97	37
1	13.5	9.4	2	11.2	9.6	3	10.4	10	4	9.71	11	5	9.6	11
1	12	16	2	10.4	13	3	9.41	14	4	8.57	17	5	7.89	25

[<matplotlib.lines.Line2D at 0x156e8128>]



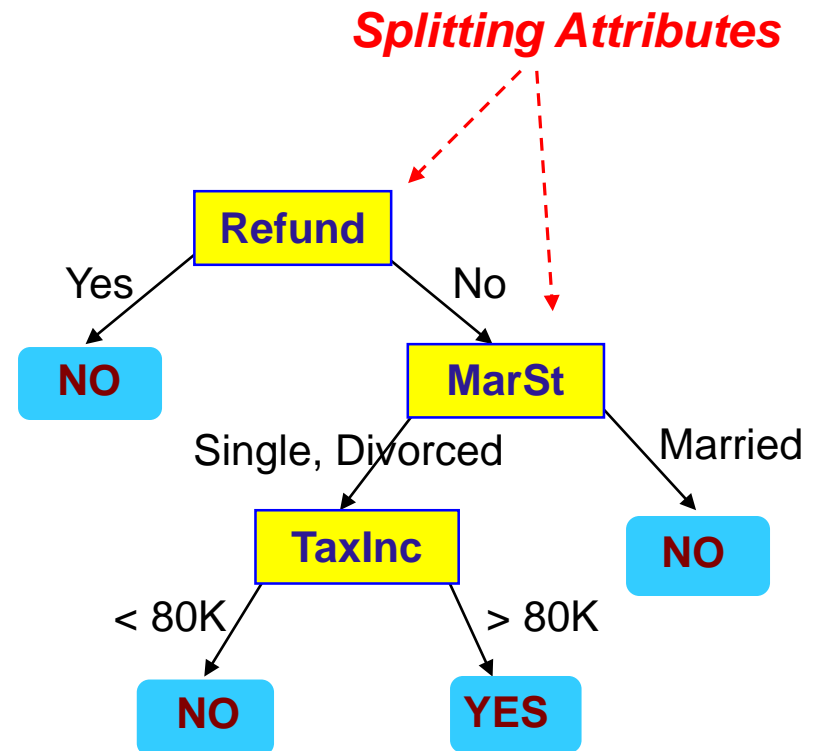
Decision Trees

How they work: Example

categorical
categorical
continuous
class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data

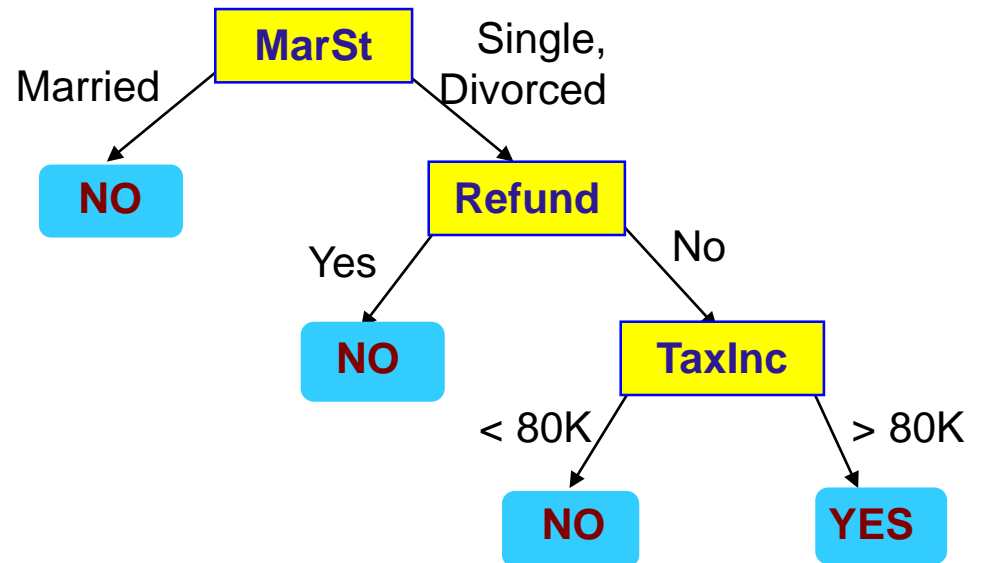


Model: Decision Tree

How they work: Example 2

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

class



There could be more than one tree that fits the same data!

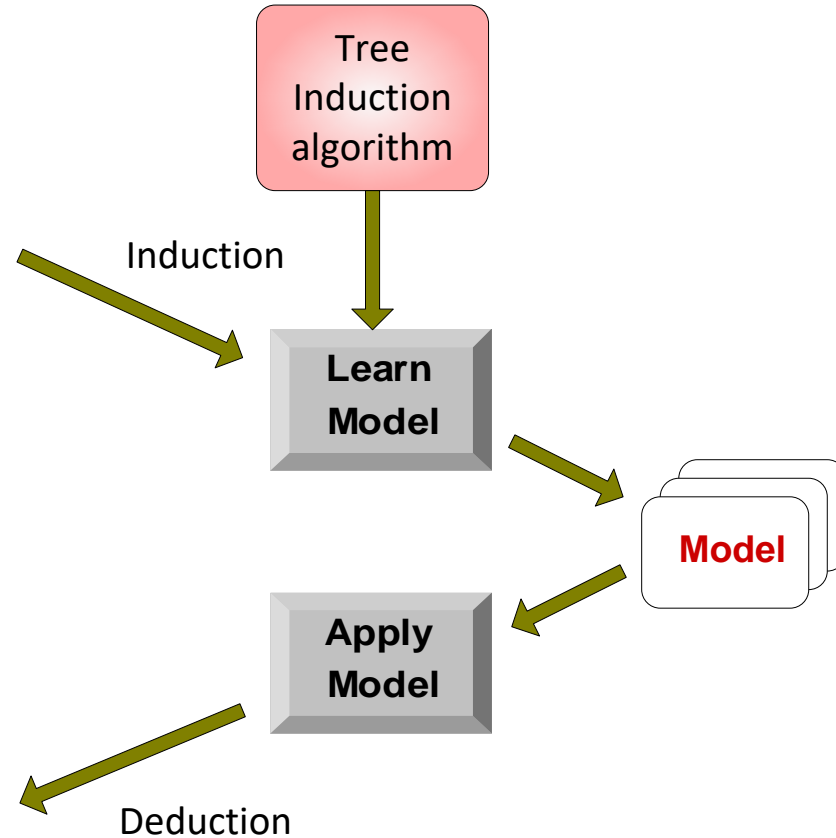
DT: How they are generated

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

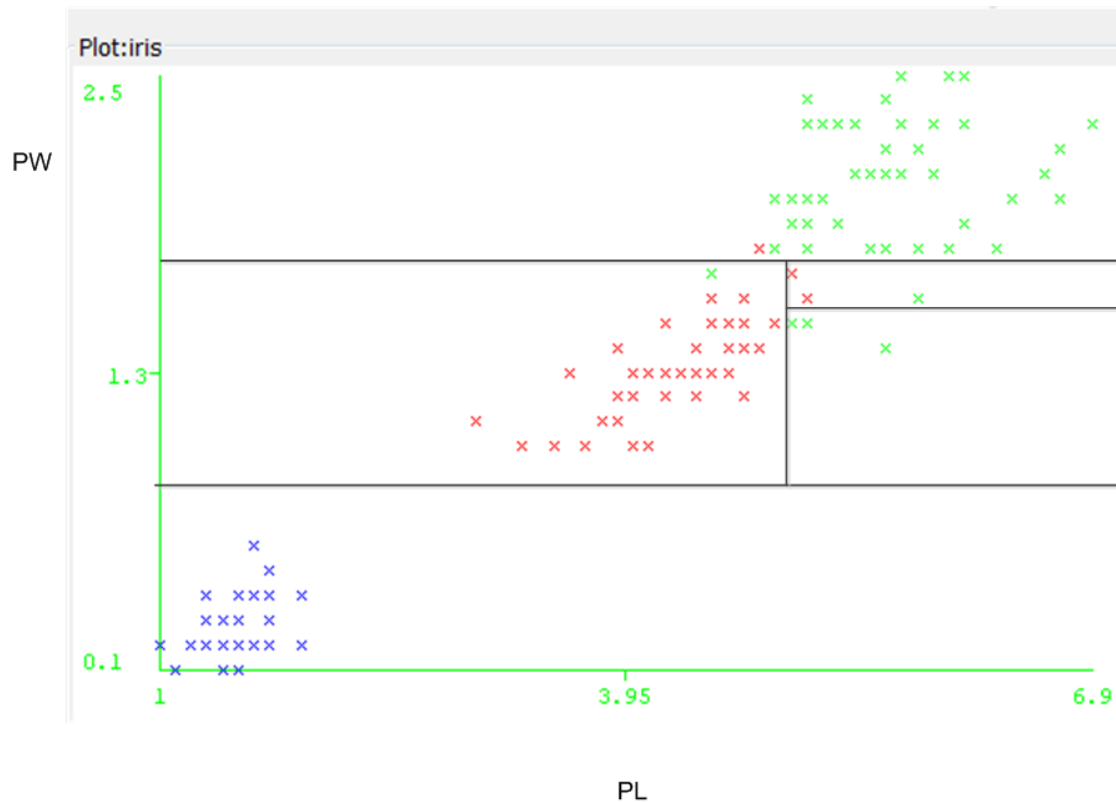
<i>Tid</i>	<i>Attrib1</i>	<i>Attrib2</i>	<i>Attrib3</i>	<i>Class</i>
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



DECISION TREES

IRIS DATA SET



Decision trees: separation between classes using horizontal and vertical lines.

DECISION TREES

TOY EXAMPLE

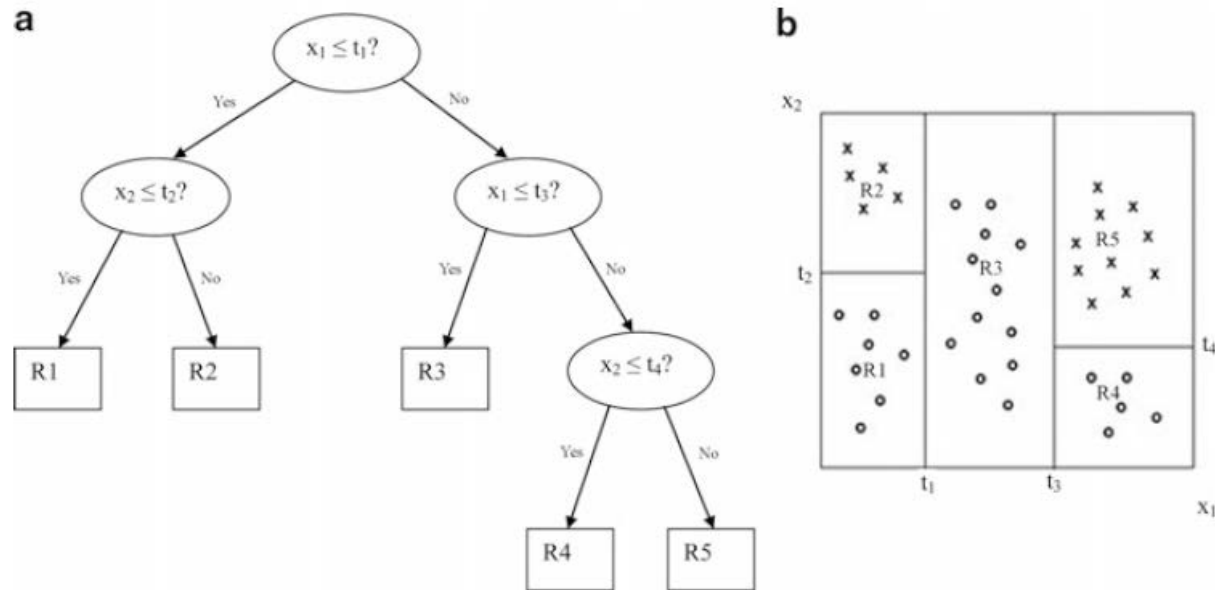
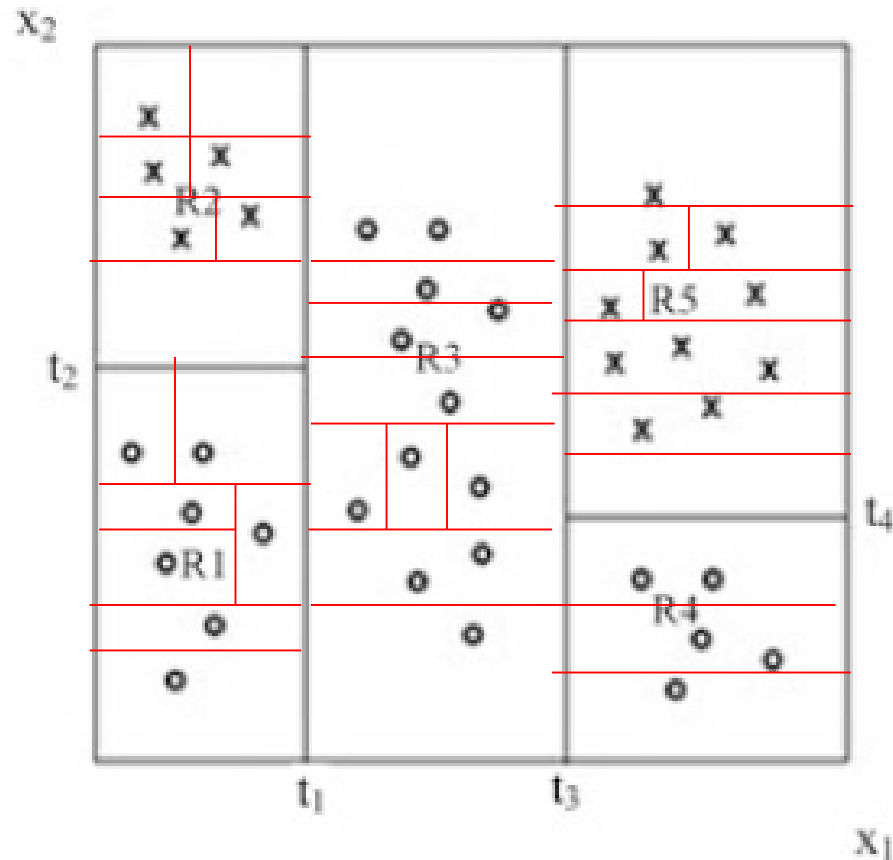


Fig. 3.8 (a) A decision tree and (b) the resulting decision boundaries in feature space

- Decision trees: separation between classes using horizontal and vertical lines.
- Can a Decision Tree overfit?

OVERFITTING



By increasing the complexity one can construct/learn a look-up table

One of the first successful applications of ML

- The [Sloan Digital Sky Survey](#) (SDSS), a large-scale digital survey of the celestial sphere (universe).
- How many objects: 1231051050 $\approx 10^9$
- All need to be classified: stars of different types, galaxies, quasars.
- First successful applications of Decision Trees

How to construct a Decision Tree

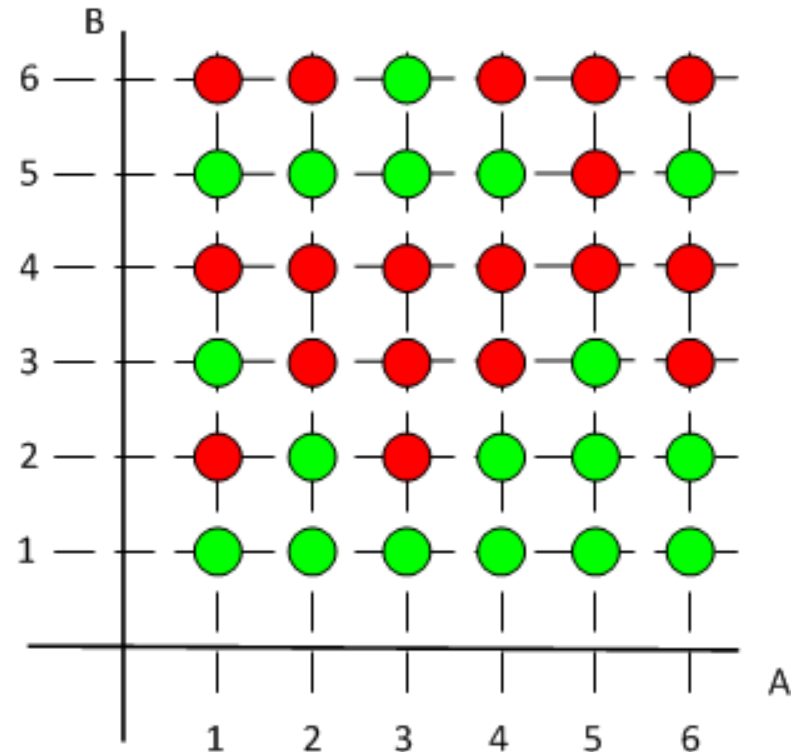
- Needed:
 - Data
 - Algorithmic Method to select best attribute for root of the tree. This method can then be applied, using *recursion*, to the rest of the tree.

Algorithmic Method

- Several methods, but we will only look at the one based on *Information Gain*.
- Information Gain is based on *entropy* of a dataset. It is measure of chaos, the less chaos the better. The more chaos the more information you need to get an 100% correct answer.

Intuitive example

- Given a urn of red and green balls. You have to guess for an unseen ball what the color is.
- Each ball has two attributes A and B with color division as given.
- The value of which feature gives you the most information A or B ?



Entropy example

- Set of data elements with the following class distribution $\{y:6, n:4\}$. That is 6 *yes* elements and 4 *no* elements.
- Entropy gives the average amount of information needed (in bits) in order to resolve the class. Entropy is a fundamental quantity of information theory (Shannon and Weaver)
- Entropy $H(p_y, p_n)$:
$$-p_y \log_2(p_y) - p_n \log_2(p_n)$$

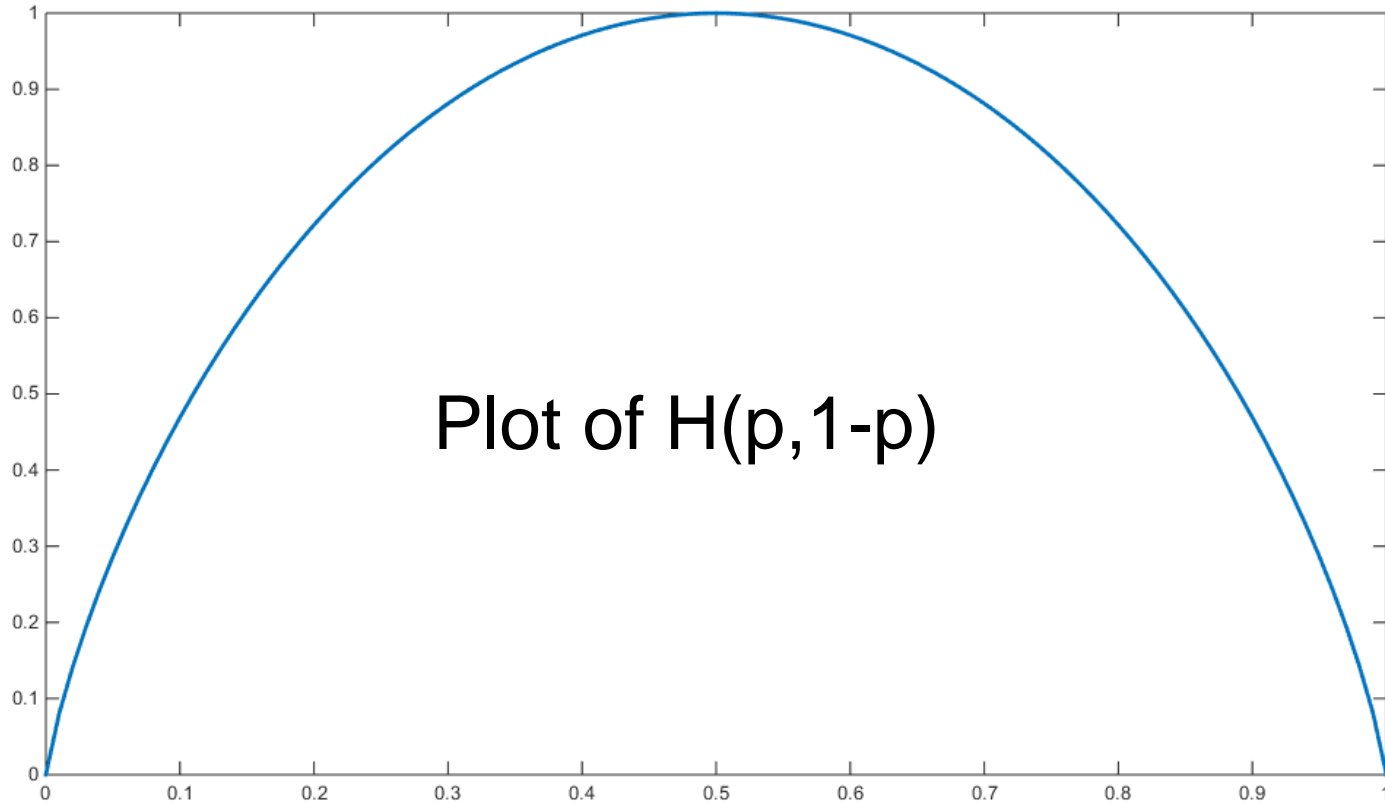
with in this case $p_y=0.6$ and $p_n=0.4$
- $H(0.6, 0.4)=0.97$

Entropy

- Suppose we have a dataset with p positive elements and n negative elements. Define $p_y = p/(p+n)$ and $p_n = n/(p+n)$. The entropy of this dataset is given by:

$$H(p_y, p_n) = -p_y \log_2(p_y) - p_n \log_2(p_n)$$


Plot of Entropy



Information gain example

Features/Attributes

Class label



Ex.	U	H	I	T	S	A
1	Y	M	N	P	M	N
2	N	S	N	P	L	N
3	Y	M	N	A	M	Y
4	N	M	N	P	S	N
5	N	M	Y	P	M	N
6	Y	N	N	A	S	N
7	N	N	N	G	S	Y
8	N	S	N	A	M	N
9	N	L	Y	P	L	Y
10	N	M	N	P	S	N
11	N	L	Y	A	M	Y
12	N	N	N	G	L	Y
13	Y	S	N	P	L	N
14	N	L	Y	P	L	Y
15	N	M	N	A	M	N
16	Y	N	N	G	S	Y
17	N	L	N	A	M	N
18	N	L	N	P	S	N
19	N	N	N	G	L	Y
20	N	N	N	G	S	Y

- Initial Entropy:
 $p_y = 9/20$, entropy is
 $H(9/20, 11/20) = 0.99$
- Entropy for feature T:

Feature T	Y	N	p_y	p_n	Entropy
P	2	7	$2/9$	$7/9$	0.76
A	2	4	$2/6$	$4/6$	0.92
G	5	0	1	0	0

- Average entropy:
 $9/20 * 0.76 + 6/20 * 0.92 + 5/20 * 0 = 0.62$
- Information gain: $0.99 - 0.62 = 0.37$

Exercise

- Compute information gain for attribute (feature) U and H .

Ex.	U	H	I	T	S	A
1	Y	M	N	P	M	N
2	N	S	N	P	L	N
3	Y	M	N	A	M	Y
4	N	M	N	P	S	N
5	N	M	Y	P	M	N
6	Y	N	N	A	S	N
7	N	N	N	G	S	Y
8	N	S	N	A	M	N
9	N	L	Y	P	L	Y
10	N	M	N	P	S	N
11	N	L	Y	A	M	Y
12	N	N	N	G	L	Y
13	Y	S	N	P	L	N
14	N	L	Y	P	L	Y
15	N	M	N	A	M	N
16	Y	N	N	G	S	Y
17	N	L	N	A	M	N
18	N	L	N	P	S	N
19	N	N	N	G	L	Y
20	N	N	N	G	S	Y

Feature U

Feature U	Y	N	p_y	p_n	Entropy
Y	2	3	0.4	0.6	0.97
N	7	8	7/15	8/15	1.00

Average entropy: 0.99

Information gain: $0.99 - 0.99 = 0$

Feature H

Feature H	Y	N	p_y	p_n	Entropy
L	3	2	0.6	0.4	0.97
M	1	5	1/6	5/6	0.65
N	5	1	5/6	1/6	0.65
S	0	3	0	1	0

Average entropy:

$$5/20 * 0.97 + 6/20 * 0.65 + 6/20 * 0.65 + 3/20 * 0 = 0.64$$

$$\text{Information gain: } 0.99 - 0.64 = 0.35$$

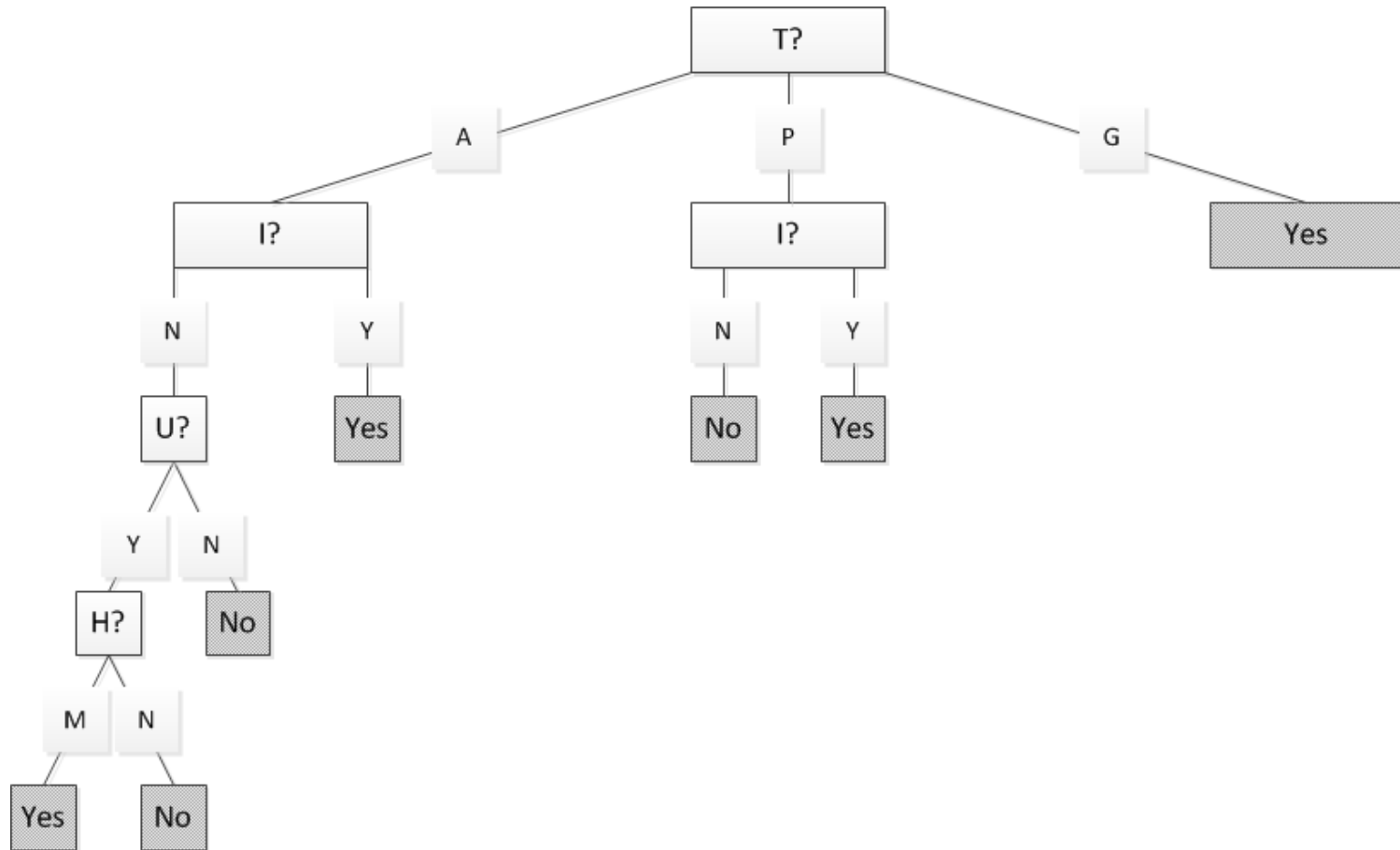
Final result

- Feature T has highest information gain. So will be the test attribute at the root of the tree.
- How to proceed?

Feature T	Examples
P	$\{1,2,4,5,9,10,13,14,18\}$
A	$\{3,6,8,11,15,17\}$
G	$\{7,12,16,19,20\}$

- For branch $T=P$ apply same method to examples for which $T=P$: $\{1,2,5,9,10,13,14,18\}$

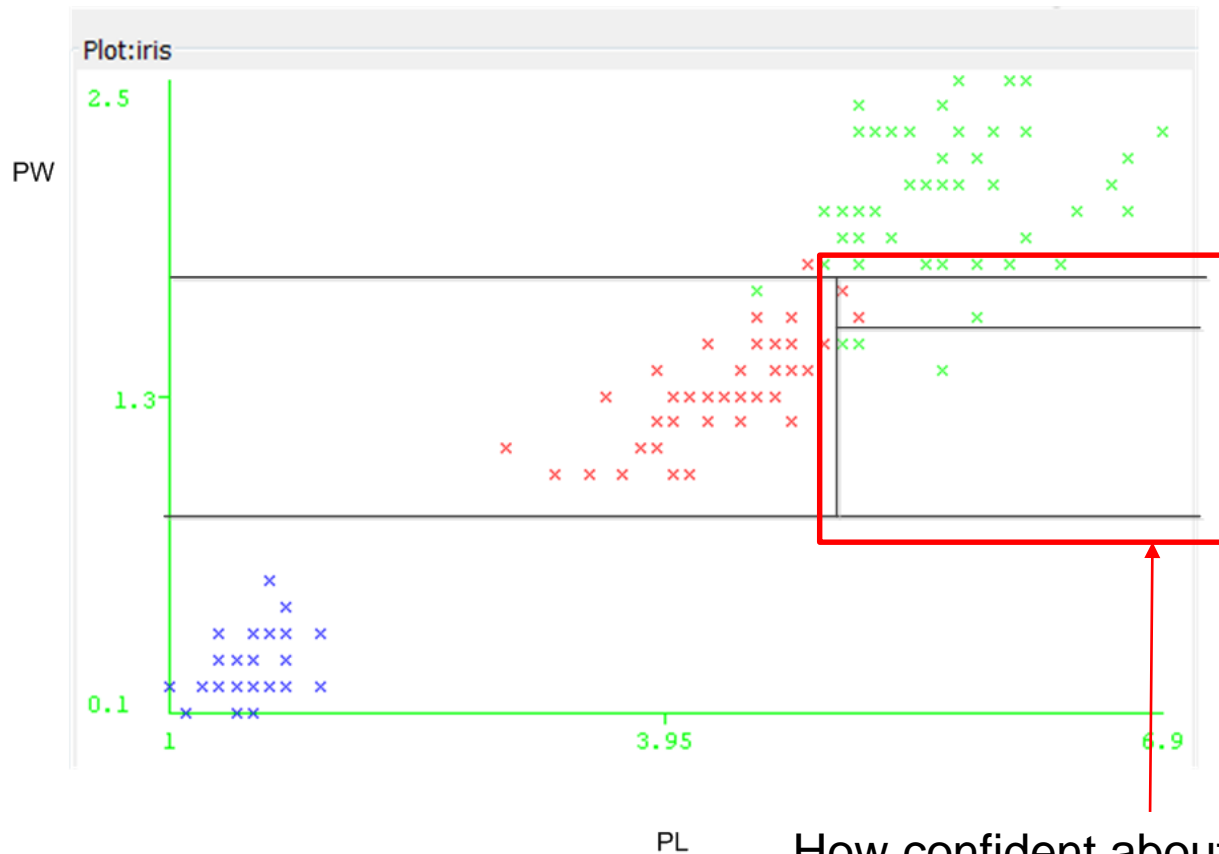
Final result



Preventing Overfitting

Reducing Complexity

Iris Example



Two approaches

- First one: lower bound on number of elements in a node. Number of elements should be larger than a certain threshold. So do not expand a node if the number of elements in the node is below or equal to a certain threshold.
- Q: How to determine the value of this threshold?

Second method: Pruning

- Prune branches for which the information gain is below a certain threshold.

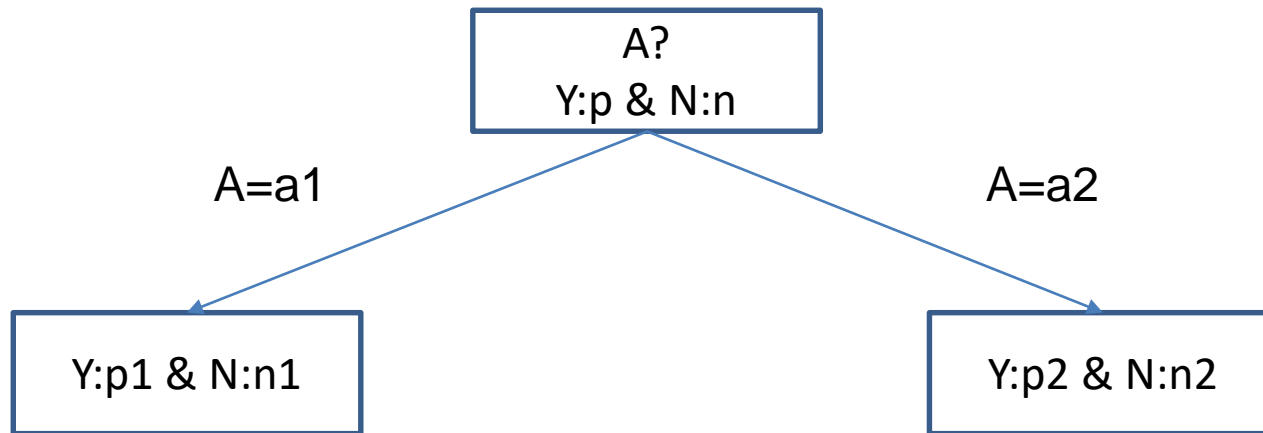
How to prune?

- Pruning during building the Decision Tree does not work.
- Example (classical XOR problem):

A	B	Class
a1	b1	Y:50 N:0
a1	b2	Y:0 N:50
a2	b1	Y:0 N:50
a2	b2	Y:50 N:0

- A and B have no information gain, but in combination they can separate the classes

Bottom up pruning after construction



- No improvement then:
$$p/(p+n)=p1/(p1+n1)=p2/(p2+n2)$$
$$n/(p+n)=n1/(p1+n1)=n2/(p2+n2)$$

Bottom up pruning

- $\frac{p}{p+n} = \frac{p1}{p1+n1} \rightarrow p1 = p * \frac{p1+n1}{p+n}$
- Define $\widehat{p1} = p * \frac{p1+n1}{p+n}$, similar for $n1$, $p2$ and $n2$.
- Calculate:

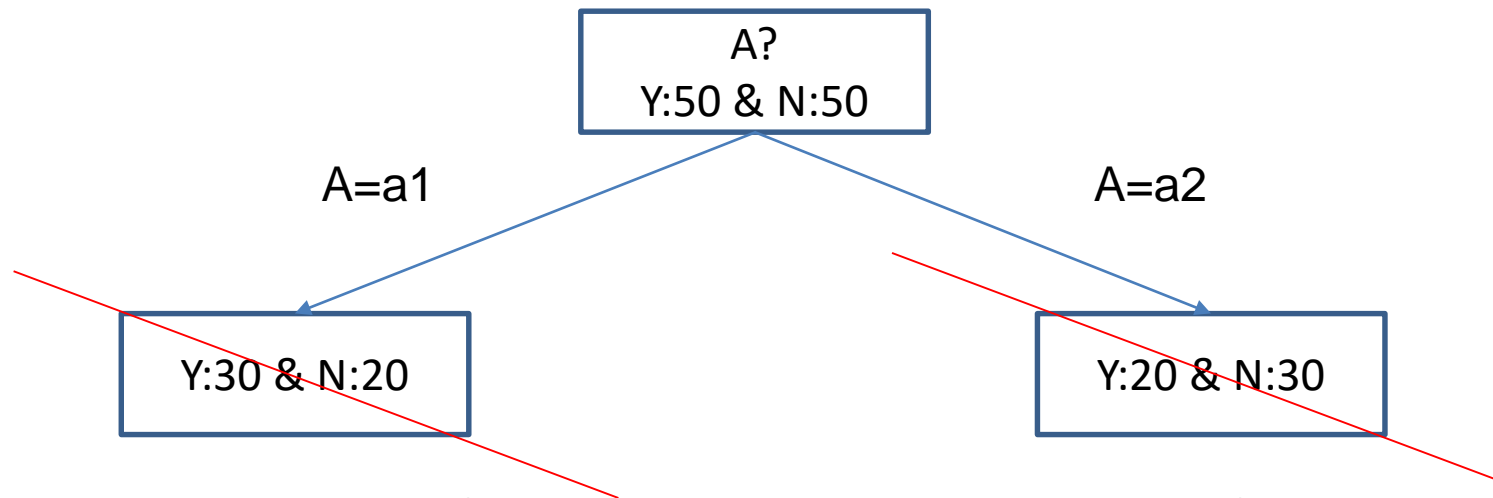
$$\Delta = \frac{(p1 - \widehat{p1})^2}{\widehat{p1}} + \frac{(n1 - \widehat{n1})^2}{\widehat{n1}} + \frac{(p2 - \widehat{p2})^2}{\widehat{p2}} + \frac{(n2 - \widehat{n2})^2}{\widehat{n2}}$$

Bottom up pruning: χ^2 pruning

- Under the null hypothesis Δ is χ^2 distributed with $d-1$ degrees of freedom, with d the number of branches.
- For 2 branches, so $d=2$ and $d-1=1$:
 $\Delta = 3.84$ reject null hypothesis at 5% level.
 $\Delta = 6.64$ reject null hypothesis at 1% level.

Exercise

- Calculate Δ for the following case:



- $\Delta = 4$ hence H_0 (no pattern, no difference) is rejected at the 5% level but not rejected for the 1% level.

Comparing classifiers

Confidence intervals for classifiers

- Test set consisting of N elements, performance of model is $acc=X/N$. What is the 95% confidence interval for this model.
- Predicting the true class label can be seen as Bernoulli experiment with N trials. For each trail the true probability is p .
- So $acc=X/N$ is a Bernoulli distribution with mean p and variance $p(1-p)$.

Confidence intervals for classifiers

- One can use the Bernoulli distribution for estimating the confidence interval. But for N large this can be approximated by a normal distribution. Thus

$$P\left(-Z_{\alpha/2} \leq \frac{acc - p}{\sqrt{p(1-p)/N}} \leq Z_{1-\alpha/2}\right) = 1 - \alpha$$

Normal distribution: $Z_{1-\alpha/2} = Z_{\alpha/2}$

Confidence intervals for classifiers

And hence the confidence interval for p is:

$$\frac{2 \times N \times acc + Z_{\alpha/2}^2 \pm Z_{\alpha/2} \sqrt{Z_{\alpha/2}^2 + 4Nacc - 4Nacc^2}}{2(N + Z_{\alpha/2}^2)}.$$

With

$1 - \alpha$	0.99	0.98	0.95	0.9	0.8	0.7	0.5
$Z_{\alpha/2}$	2.58	2.33	1.96	1.65	1.28	1.04	0.67

Example confidence interval

- Accuracy=0.8 and $N=80$. What is 95% confidence interval?
- Answer: [0.71, 0.87]

Comparing classifiers

- Two cases:
 - Different test sets (but from same data set)
 - Same datasets

Several approaches can be found in handout about DT and article of Dietterich.

Same datasets:

5 x 2 CV paired t test

- 5 replications of 2-CV.
- In replication i , in fold j : $p_i^{(j)}$ is the difference of the error between the two classifiers.
- Average on replication i : $\bar{p}_i = (p_i^{(1)} + p_i^{(2)})/2$
- Estimated variance on replication i :

$$s_i^2 = (p_i^{(1)} - \bar{p}_i)^2 + (p_i^{(2)} - \bar{p}_i)^2$$

Same datasets:

5 x 2 CV paired t test

- Under the null hypothesis, no difference, $p_i^{(j)}$ is approximately normally distributed.
- Then one can deduce that:

$$t = \frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^5 s_i^2 / 5}} \sim t_5$$

- A t statistics with 5 degrees of freedom.
- Null hypothesis rejected at significance level α if the value is outside the interval $(-t_{\alpha/2,5}, t_{\alpha/2,5})$.
- $t_{0.025,5} = 2.57$ the 95% confidence interval

