Statistical Computing

Decision Trees

Uwe Menzel, 2011

uwe.menzel@matstat.org

www.matstat.org

What are decision trees?

Carl Kingsford & Steven L Salzberg

Decision trees have been applied to problems such as assigning protein function and predicting splice sites. How do these classifiers work, what types of problems can they solve and what are their advantages over alternatives?

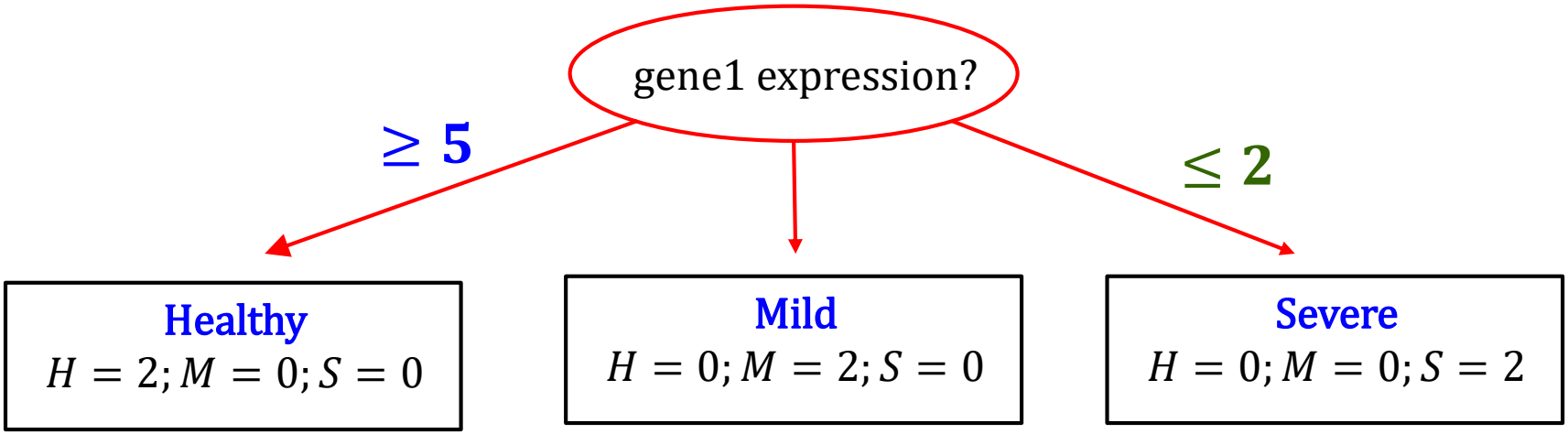
NATURE BIOTECHNOLOGY VOLUME 26 NUMBER 9 SEPTEMBER 2008

Constructing Decision Trees
- A trivial example -

- Finding the right criteria (questions) to subdivide a dataset into classes
- In the training dataset, the **classes** (attributes) must be known.
- **Example**:
  - Healthy, Mild infection, Severe infection
- The **expression values** of a single gene are sufficient to classify the patients:

#	gene1	class
1	6	H(ealthy)
2	5	H
3	4	M(ild)
4	3	M
5	1	S(evere)
6	2	S

#	gene1	class
1	6	H(ealthy)
2	5	H
3	4	M(ild)
4	3	M
5	1	S(evere)
6	2	S



- A new patient with unknown class ("**test set**") could now easily be classified.
- **But**: In most of the cases, constructing a tree is far more complicated.

Constructing Decision Trees

- Finding the right criteria (questions) to subdivide a dataset into classes
- In the training dataset, the **classes** (attributes) must be known:
- Healthy, Mild infection, Severe infection
- The **expression values** of gene1 and gene2 are sufficient to classify the patients, while the categorical variable "smoking" cannot be used for classification:

#	gene1	gene2	smoking	class
1	6	2	yes	H(ealthy)
2	5	2	no	H
3	1	5	yes	M(ild)
4	2	4	no	M
5	1	2	yes	S(evere)
6	1	3	no	S

Constructing Decision Trees

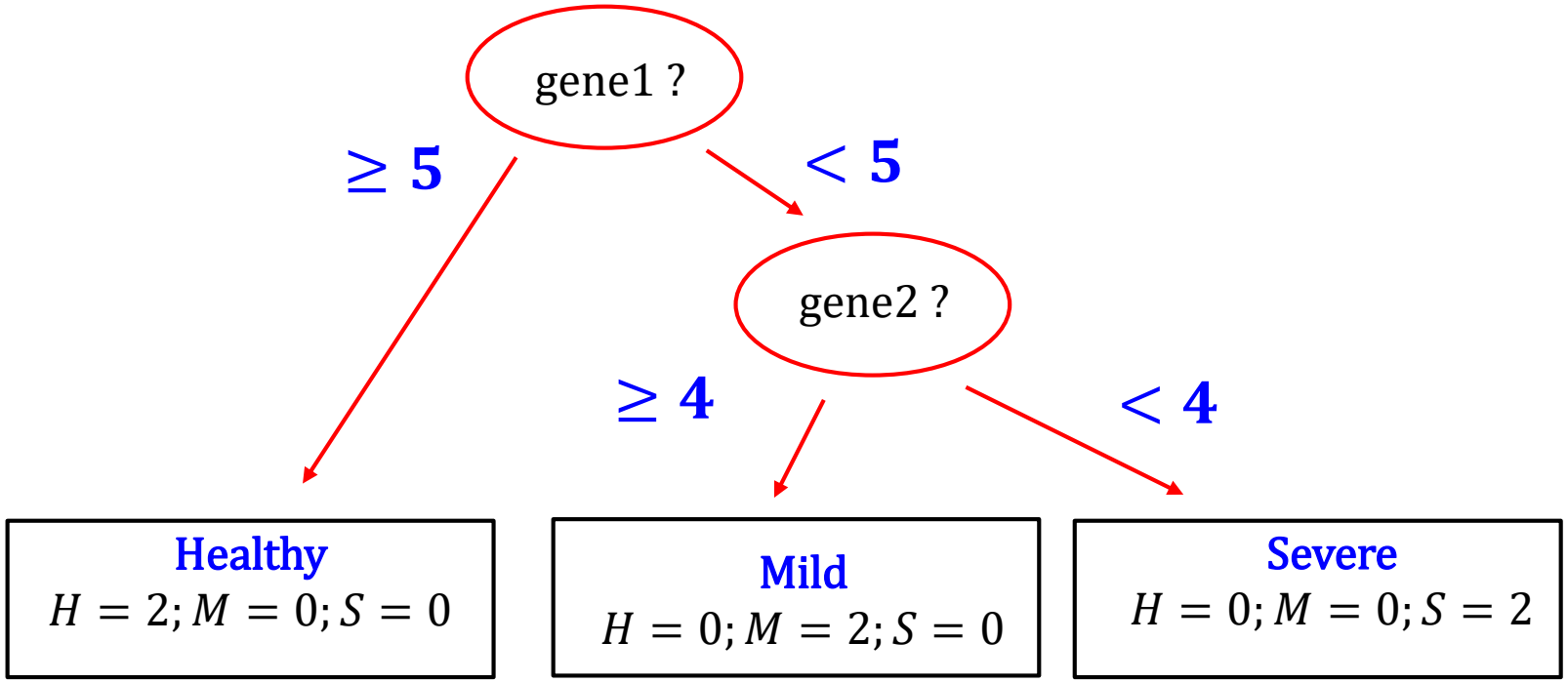
#	gene1	gene2	smoking	class
1	6	2	yes	H(ealthy)
2	5	2	no	H
3	1	5	yes	M(ild)
4	2	4	no	M
5	1	2	yes	S(evere)
6	1	3	no	S

- if gene1 $\geq 5 \rightarrow$ H
- gene1 is not sufficient to classify for mild and severe infection, we need more:
- if gene1 < 5 **and** gene2 $\geq 4 \rightarrow$ M
- if gene1 < 5 and gene2 $< 4 \rightarrow$ S ... we need two questions here!

uwe.menzel@matstat.org

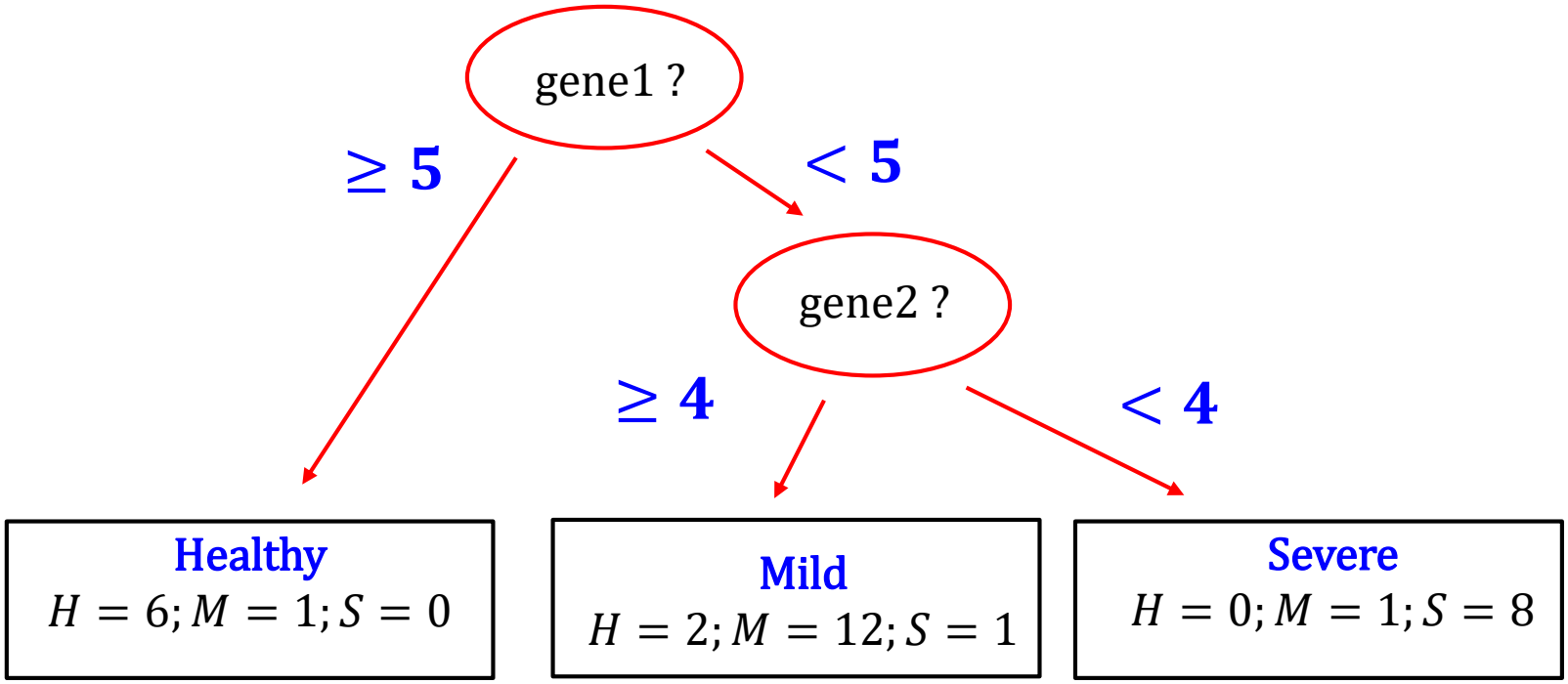
Constructing Decision Trees

- gene1 $\geq 5 \rightarrow$ **H**
- gene1 < 5 and gene2 $\geq 4 \rightarrow$ **M**
- gene1 < 5 and gene2 $< 4 \rightarrow$ **S**



Often, samples cannot be classified perfectly, i.e. the leafs remain **impure**.

Impure leafs



- After tracking down the tree, we end up in a leaf
- In each leaf, if we pick some patient, there is a certain **probability** that the patient is healthy, mildly infected, or severely infected

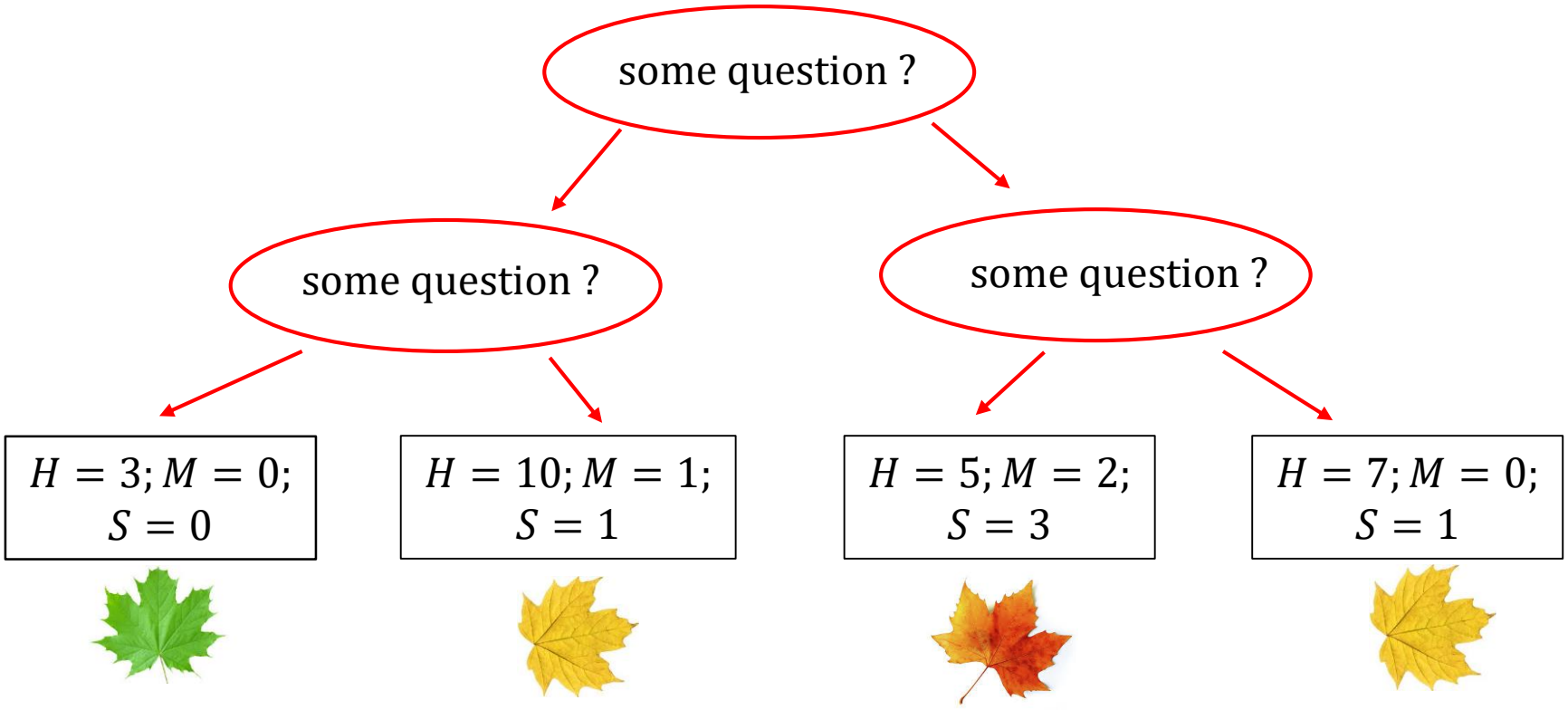
Impurity of the leaves in a Decision Tree

- Example: 33 **samples** subdivided into 3 **classes**:
- classes (**attributes**): healthy (H); mild infection (M); severe infection (S)

#	gene1	gene2	gene3	gene4	gene5	...	class
1	33.5	45.3	56.7	3455.8	34.4	...	H
2	543.5	567.3	6677.4	335.9	334.5	...	M
3	445.5	345.4	667.4	556.4	3445.4	...	M
4	233.1	985.2	33.2	43.4	45.9	...	S
...	...	...	...	...	...	...	...
33	345.4	6677.4	335.9	567.3	33.5	...	S

- The numbers in the table could be expression values or whatever.
- The table could also contain categorical variables like gender (f/m), smoking (y/n), ...

Assume we have constructed a decision tree:



- The **green leaf** contains 3 healthy, no mild, no severe case. That's what we wish: the leaf is **pure**. When ending up in this leaf after asking two questions, we can be sure that every person in this leaf is definitely healthy.
- The **red leaf** contains 5 healthy, 2 mild, and 3 severe. This is an unwanted situation because the samples were not very well classified. Persons in this leaf can be healthy, or have a mild or a severe infection, each with some **probability**.

Probabilities of the classes in leaf I

$H = 3; M = 0;$
$S = 0$

the leftmost leaf in the tree above

- Assume we ended up in this leaf after judging two criteria ("asking two questions")
- If we then pick a patient from this leaf, we have the following probabilities that the patient is healthly, mildly infected, or severely infected:

- $n = 3$: total number of samples in this leaf
- $p_1 = {}^{3}/_{3} = 1$: probability that a patient in this leaf is healthy.
- $p_2 = {}^{0}/_{3} = 0$: probability that a patient in this leaf has a mild infection.
- $p_3 = {}^{0}/_{3} = 0$: probability that a patient in this leaf has a severe infection.

uwe.menzel@matstat.org

Probabilities of the classes in leaf III

$H = 5; M = 2;$
$S = 3$

the 3rd leaf in the tree above

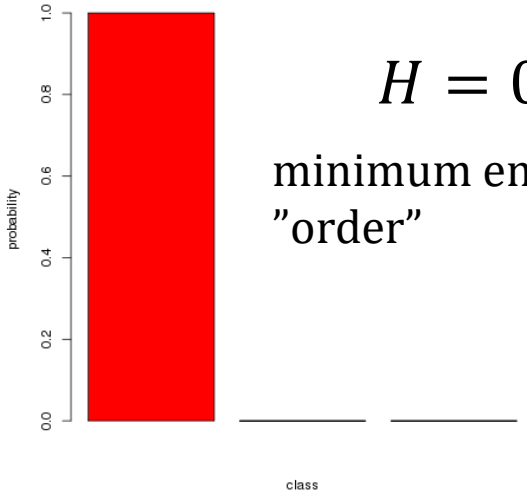
- Assume we ended up in this leaf after judging two criteria ("asking two questions")
- If we then pick a patient from this leaf, we have the following probabilities that the patient is healthly, mildly infected, or severely infected:

- $n = 10$: total number of samples in this leaf
- $p_1 = {}^{5}/_{10} = 0.5$: probability that a patient in this leaf is healthy.
- $p_2 = {}^{2}/_{10} = 0.2$: probability that a patient in this leaf has a mild infection.
- $p_3 = {}^{3}/_{10} = 0.3$: probability that a patient in this leaf has a severe infection.

Entropy definition based on discrete probabilities

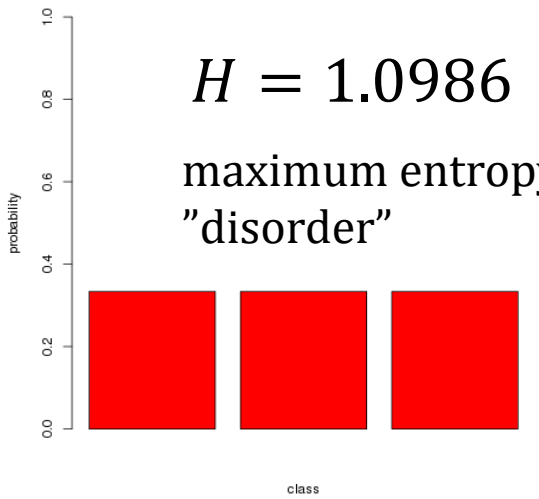
$$H = -\sum_{i=1}^{n} p_i \cdot \ln(p_i)$$ always positive because $p_i \leq 1$

- The entropy has a minimum (of zero) if all probability is concentrated on a single category, for instance if $p_1 = 1$ while $p_2 = 0$ and $p_3 = 0$ (left plot).
- The entropy has a maximum if the probabilities are evenly distributed between the categories, for instance if $p_1 = {}^1/_3$; $p_2 = {}^1/_3$ and $p_3 = {}^1/_3$



$H = 0$

minimum entropy
"order"



$H = 1.0986$

maximum entropy
"disorder"

2nd law of thermodynamics: isolated systems spontaneously evolve towards maximum entropy.

uwe.menzel@matstat.org

Entropy in leaf I

$H = 3; M = 0;$
$S = 0$

- $n = 3$: total number of samples in this leaf
- $p_1 = {}^3/_3 = 1$: probability that a patient in this leaf is healthy.
- $p_2 = {}^0/_3 = 0$: probability that a patient in this leaf has a mild infection.
- $p_3 = {}^0/_3 = 0$: probability that a patient in this leaf has a severe infection.

$$
\begin{aligned}
H &= -\left[p_1 \cdot \ln(p_1) + p_2 \cdot \ln(p_2) + p_3 \cdot \ln(p_3)\right] \\
&= -\left[\frac{3}{3} \cdot \ln\left(\frac{3}{3}\right) + \frac{0}{3} \cdot \ln\left(\frac{0}{3}\right) + \frac{0}{3} \cdot \ln\left(\frac{0}{3}\right)\right] \\
&= -\ln(1) = 0
\end{aligned}
$$

What is $0 \cdot \ln(0)$? see Appendix

Entropy in leaf II

$H = 10; M = 1;$
$S = 1$

- $n = 12$: total number of samples in this leaf
- $p_1 = {}^{10}/_{12}$: probability that a patient in this leaf is healthy.
- $p_2 = {}^{1}/_{12}$: probability that a patient in this leaf has a mild infection.
- $p_3 = {}^{1}/_{12}$: probability that a patient in this leaf has a severe infection.

$$H = -[p_1 \cdot \ln(p_1) + p_2 \cdot \ln(p_2) + p_3 \cdot \ln(p_3)]$$

$$= -\left[\frac{10}{12} \cdot \ln\left(\frac{10}{12}\right) + \frac{1}{12} \cdot \ln\left(\frac{1}{12}\right) + \frac{1}{12} \cdot \ln\left(\frac{1}{12}\right)\right] = 0.5660857$$

Entropy in leaf III

$H = 5; M = 2;$
$S = 3$

- $n = 10$: total number of samples in this leaf
- $p_1 = {}^{5}/_{10}$: probability that a patient in this leaf is healthy.
- $p_2 = {}^{2}/_{10}$: probability that a patient in this leaf has a mild infection.
- $p_3 = {}^{3}/_{10}$: probability that a patient in this leaf has a severe infection.

$$H = -[p_1 \cdot \ln(p_1) + p_2 \cdot \ln(p_2) + p_3 \cdot \ln(p_3)]$$

$$= -\left[\frac{5}{10} \cdot \ln\left(\frac{5}{10}\right) + \frac{2}{10} \cdot \ln\left(\frac{2}{10}\right) + \frac{3}{10} \cdot \ln\left(\frac{3}{10}\right)\right] = 1.029653$$

Entropy in leaf IV

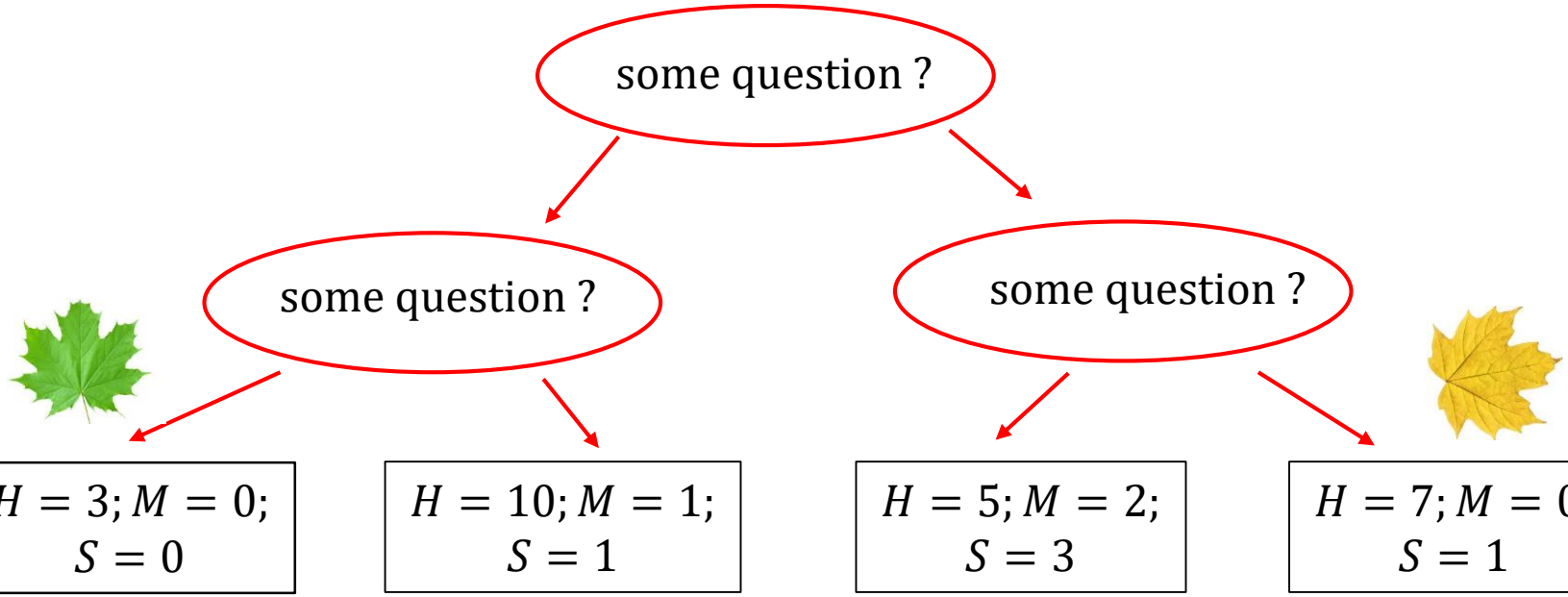
$H = 7; M = 0;$
$S = 1$

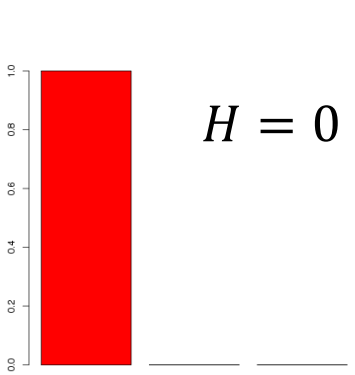
- $n = 8$: total number of samples in this leaf
- $p_1 = {}^{7}/_{8}$: probability that a patient in this leaf is healthy.
- $p_2 = {}^{0}/_{8}$: probability that a patient in this leaf has a mild infection.
- $p_3 = {}^{1}/_{8}$: probability that a patient in this leaf has a severe infection.

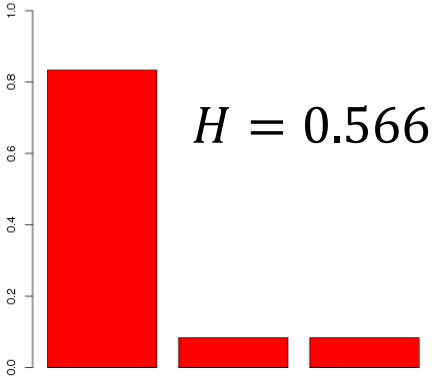
$$H = -[p_1 \cdot \ln(p_1) + p_2 \cdot \ln(p_2) + p_3 \cdot \ln(p_3)]$$

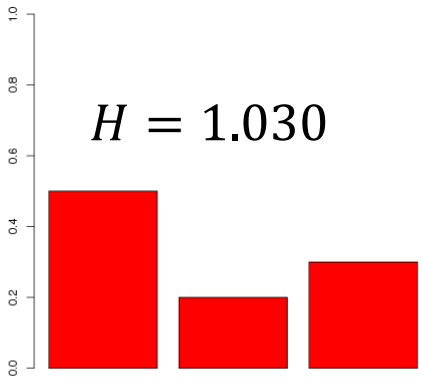
$$= -\left[\frac{7}{8} \cdot \ln\left(\frac{7}{8}\right) + \frac{0}{8} \cdot \ln\left(\frac{0}{8}\right) + \frac{1}{8} \cdot \ln\left(\frac{1}{8}\right)\right] = 0.3767702$$

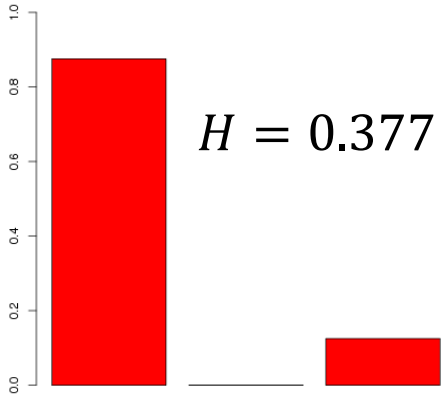
Entropy in the leaves



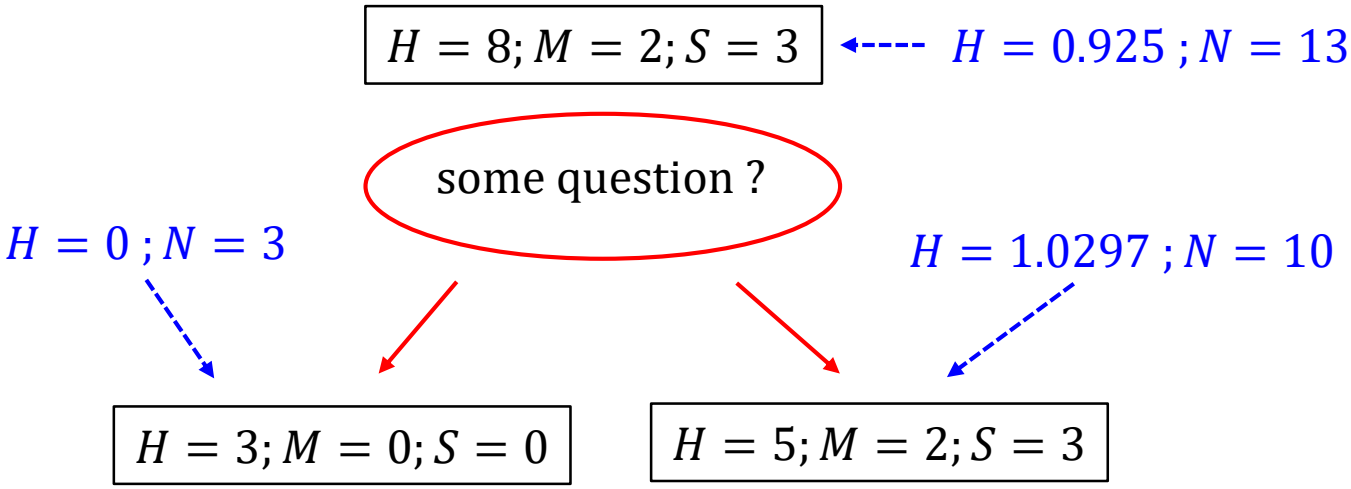








Is a question (defining a split) improving classification?



The child leaves should be "purer" than the parent leaf. How to calculate purity?

$$
\begin{aligned}
I &= \sum_{i=1}^{childs} \frac{N_i}{N} \cdot H_i \\
&= \frac{N_1}{N_1 + N_2} \cdot H_1 + \frac{N_2}{N_1 + N_2} \cdot H_2 \\
&= \frac{3}{13} \cdot 0 + \frac{10}{13} \cdot 1.0297 = 0.792
\end{aligned}
$$

Weighted average of the impurity of the resulting child nodes.

- N_i = nr. samples in a leaf
- H_i= entropy in a leaf
- $N = \sum N_i$

Oops: Check if the number of H, M, and S is the same on each level of the tree!: e.g for H: 8 = 3 + 5

Information gain of a split

- Information gain = difference between:
  - the entropy in the parent node and
  - the weighted average of the children's entropy
- always has a non-negative value

Gain = 0.925 - 0.792 = 0.133 in the example above.

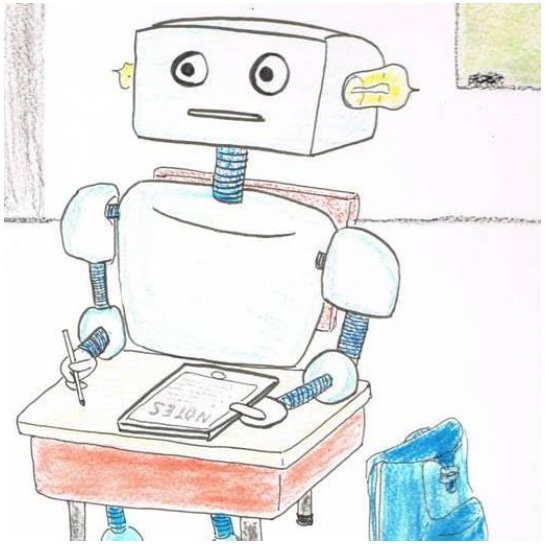
The information gain is the change in information entropy H from a prior state to a state that takes some information as given:

$$IG(T, a) = H(T) - H(T \mid a)$$

Kullback-Leibler divergence

Machine Learning

- If we once have created a tree, we can classify new samples by "running them down the tree" (Breiman, Cutler [1])
- we have learned a classification scheme → machine learning



http://gureckislab.org/blog

[1] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Limiting the complexity of the learned tree

- **Q**: Why not create a tree being so big that all leafs are pure?
  - in an extreme case they would just contain a single sample.
- **A**: Avoid overfitting!
  - overfitting means that the model (tree) is too closely adapted to the actual dataset under consideration

- **Measures to avoid overfitting**:
- Stop splitting when gain is not higher than some threshold, or
- build tree until no leaf can be further subdivided, prune the tree afterwards by deleting nodes
  - pruning: collapse internal nodes into leafs if this reduces the classification error on a held-out test set.
  - minimum description length
- Check the models by performing cross-validation

Thanks.
I hope you understood that trees are indeed of great importance!

Appendix

Limits of expressions "0/0" or "±∞/±∞"

- L'Hopital's Rule -

- http://tutorial.math.lamar.edu/Classes/CalcI/LHospitalsRule.aspx

Suppose that we have one of the following cases,

$$\lim_{x\to a}\frac{f(x)}{g(x)}=\frac{0}{0} \qquad \text{OR} \qquad \lim_{x\to a}\frac{f(x)}{g(x)}=\frac{\pm\infty}{\pm\infty}$$

where a can be any real number, infinity or negative infinity. In these cases we have,

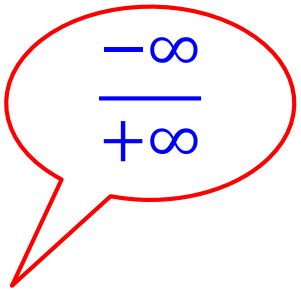
$$\lim_{x\to a}\frac{f(x)}{g(x)}=\lim_{x\to a}\frac{f'(x)}{g'(x)}$$

L'Hopital's rule tells us that if we have an indeterminate form 0/0 or ∞/∞ all we need to do is to differentiate the numerator and denominator and then take the limit.

$$\lim_{x\to 0}\frac{sinx}{x}=\lim_{x\to 0}\frac{cosx}{1}=\frac{1}{1}=1$$

Limits of expressions "0/0" or "±∞/±∞"

- L'Hopital's Rule -



$$\lim_{x\to 0} x\cdot\ln(x) = \lim_{x\to 0}\frac{\ln(x)}{\frac{1}{x}} = \lim_{x\to 0}\frac{\frac{1}{x}}{\frac{-1}{x^2}} = \lim_{x\to 0}(-x) = 0$$

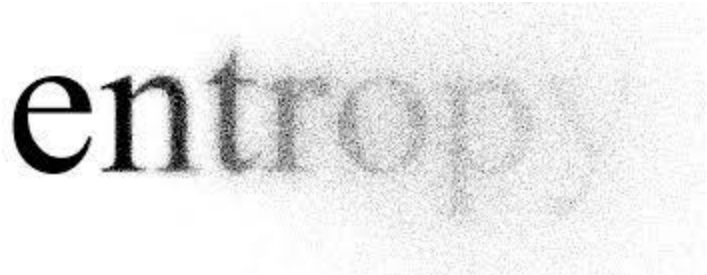


Image: http://creepypasta.wikia.com/wiki/Entropy

The entropy is defined as the negated expectation value of the logarithm of the probability distribution:

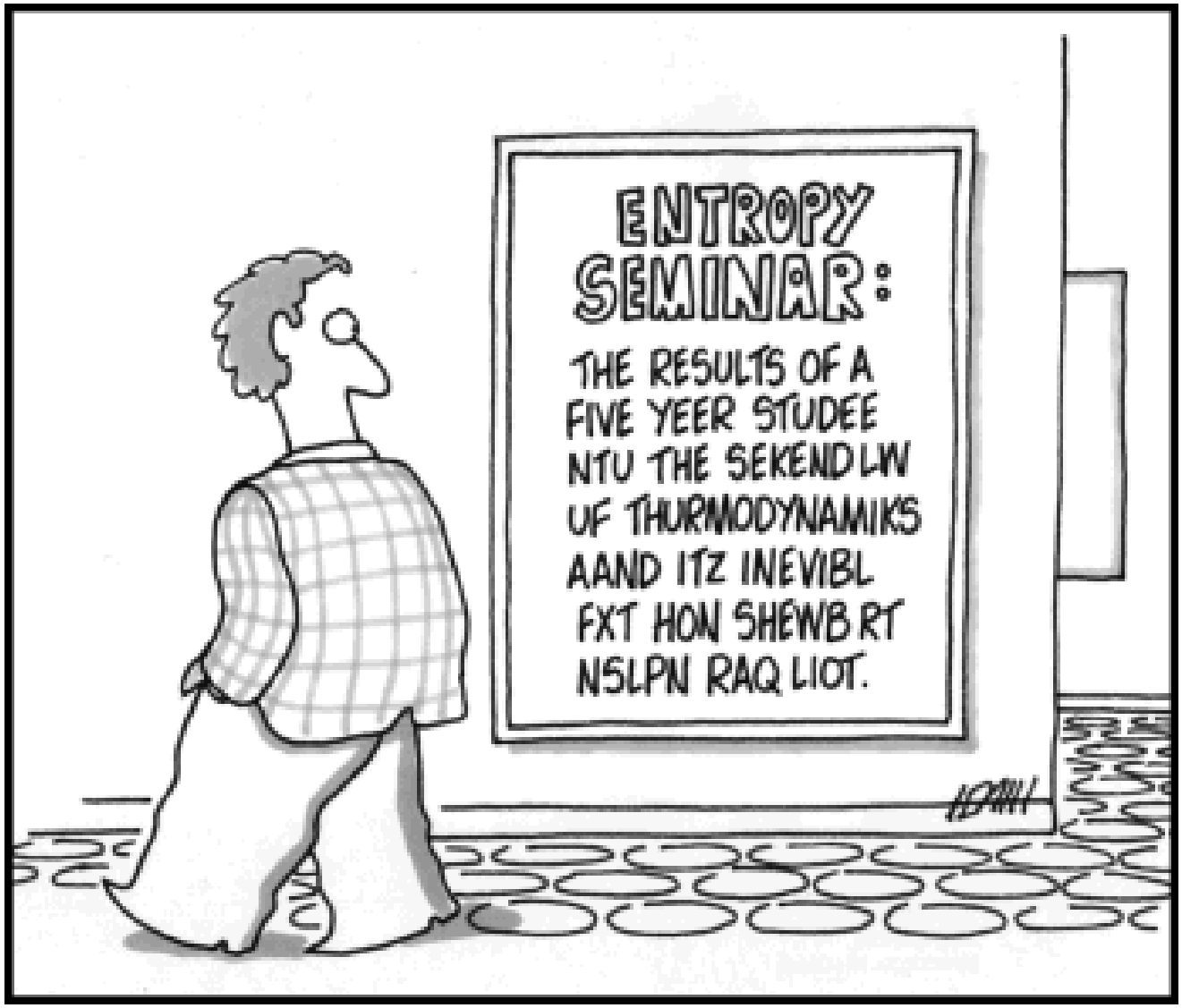
$$H = -E\left[\ln P\right]$$

For discrete probability distributions, this translates to:

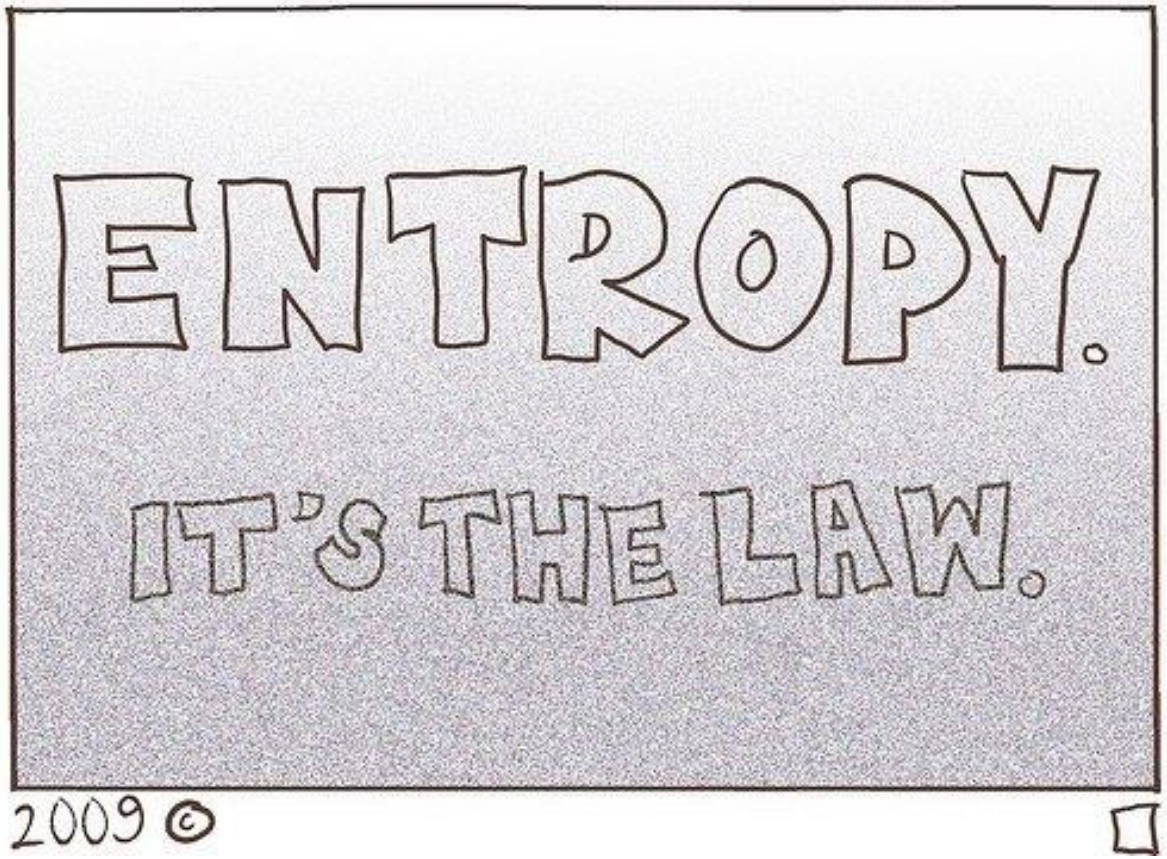
$$H = -\sum_{i=1}^{n} p_i \cdot \ln(p_i)$$

always positive because $p_i \leq 1$

- Entropy captures the amount of randomness or uncertainty in a variable.
- The entropy has a minimum (of zero) if all probability is concentrated on a single category, for instance if $p_1 = 1$ while $p_2 = 0$ and $p_3 = 0$.
- The entropy has a maximum if the probabilities are evenly distributed between the categories, for instance if $p_1 = {}^1/_3$; $p_2 = {}^1/_3$ and $p_3 = {}^1/_3$.



http://uncyclopedia.wikia.com/wiki/Entropy



http://brownsharpie.courtneygibbons.org