Universität Stuttgart

Remote Sensing
Chapter 5: Classification

Prof. Dr.-Ing. Uwe Sörgel
soergel@ifp.uni-stuttgart.de

ifp

Inhalt

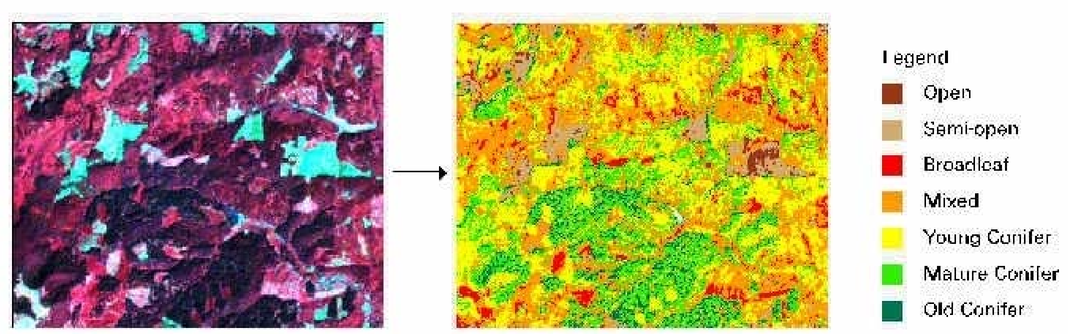
- Elektrisches und magnetisches Feld
- Schwingungen und Wellen
- Strahlungsbilanz
- Interaktion von Wellen und Materie
- Verschiedene Arten der Auflösung

Contents

- Introduction
- Unsupervised Classification
- Supervised Classification
- How to quantify classification performance?

Motivation: Derive Map from RS image

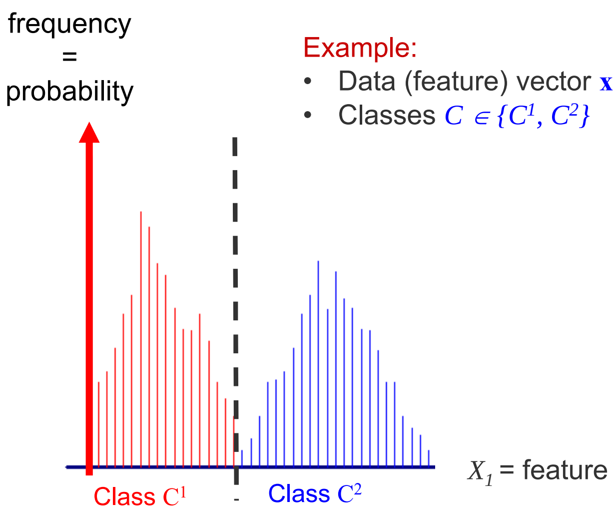
- Example Landsat TM image
- Aim: Distinguish types of forest



Land Cover vs. Land Use

- Classification gives LAND COVER
- Many users are more interested in how the terrain is being used: LAND USE
- Example
  - Grass is land cover
  - pasture and recreational parks are *land uses* of grass
- LAND USE extraction requires e.g. context

Separation of classes according to features



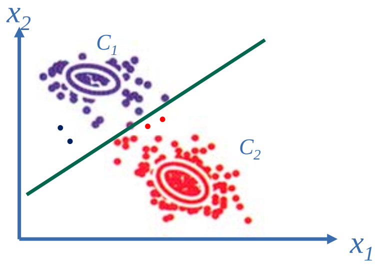
Example:
- Data (feature) vector $\mathbf{x} = [x_1]$
- Classes $C \in \{C^1, C^2\}$

Sometimes it is sufficient to consider the histogram of a single feature to separate classes → „1d analysis“

Enlarging the dimension of feature space

- Example:

$$\mathbf{x} = [x_1, x_2]^T$$
$$C \in \{C^1, C^2\}$$

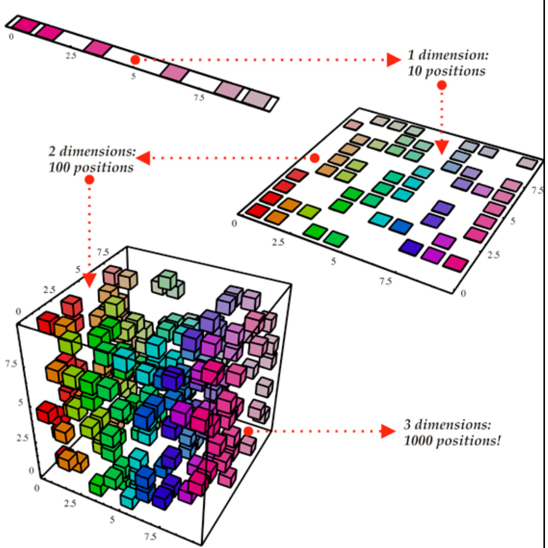


- Only from x_1 we are not able to separate C^1 and C^2.
- By enlarging the dimension of feature space with x_2 the classes C^1 and C^2 become separable.
- Note: Dimension of feature space must not get too large (Curse of Dimensionality)
- In remote sensing, for example, a pixel of a multi-spectral image is one dimension of the feature space.
- The choice of features is crucial for classification performance: usually **model knowledge** required!

Problem: too many features

- Curse of Dimensionality
  - In case of high-dimension very many training data required.
  - Sometimes only a few of those dimensions important.
  - Irrelevant dimensions might be misleading

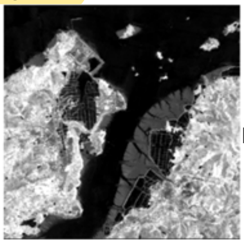
In order to ensure equal sample density, the data volume must grow exponentially with dimension d, here:
1-dim: 10^1
2-dim: 10^2
3-dim: 10^3



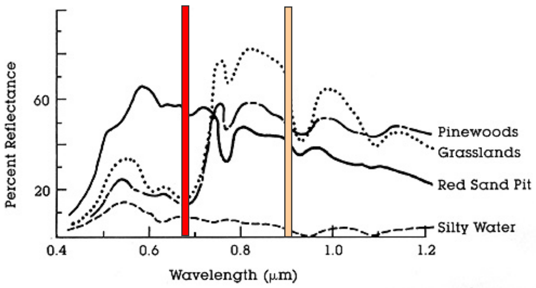
From images to features

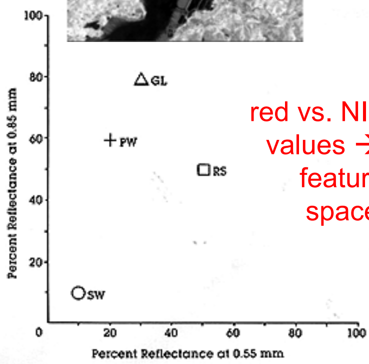
- Separate different materials by different spectral response
- Number n of bands gives dimension of feature space

Red band
(TM 3)



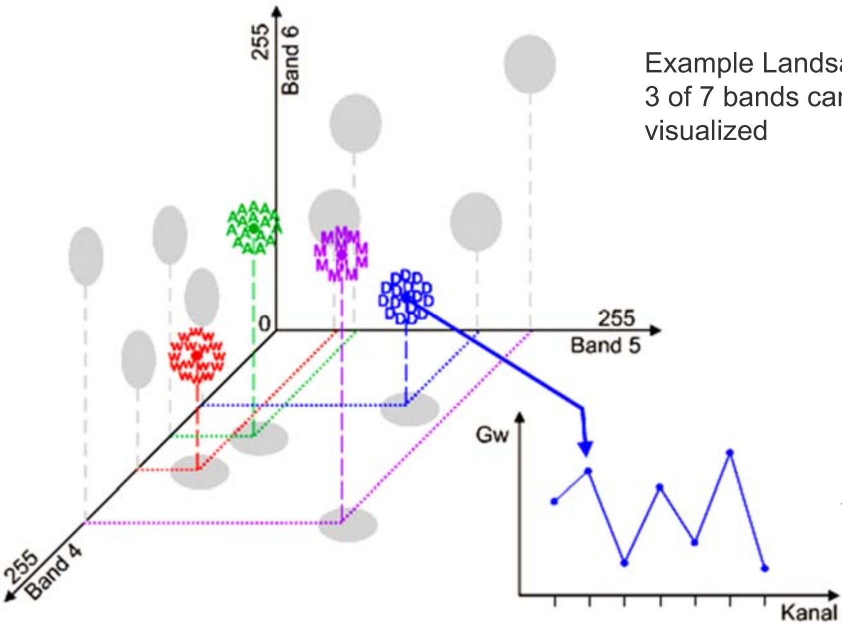
NIR band
(TM 4)





red vs. NIR DN values → 2d feature space

Feature Space: Multi-dimensional coordinate system



Example Landsat TM:
3 of 7 bands can be visualized

Example TM:
7 bands as vector

Land classification

- Aims to label each pixel in a scene to specific land cover types.
- Pixels can then be either correctly classified, incorrectly classified, or unclassified.
- Two main types of classification
  - Unsupervised
  - Supervised

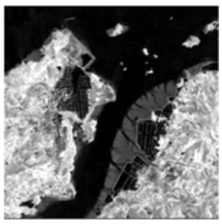
Unsupervised Classification

Unsupervised classification

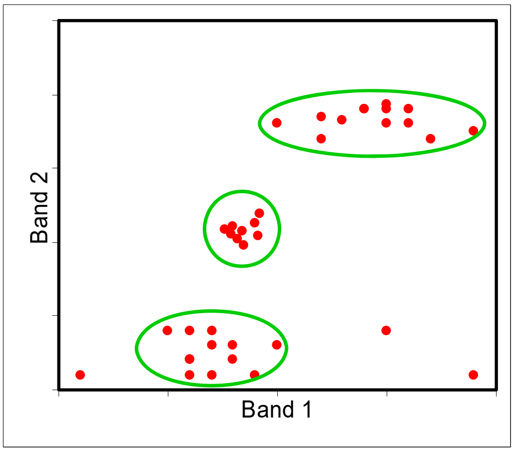
- No a priori knowledge assumed about data.
- Tries to spectrally separate the pixels.
- User has control over:
  - Number of classes
  - Number of iterations
  - Convergence thresholds
- Two main algorithms: Isodata and k-means

Example spectral plot

Red channel
(band 1)



NIR channel
(band 2)



- Two bands of data.
- Each pixel marks a location in this 2d spectral space
- Our eyes can split the data into clusters
- Some points do not fit to clusters.

K-means (unsupervised)

1. A set number K of cluster centres are positioned randomly through the spectral space.

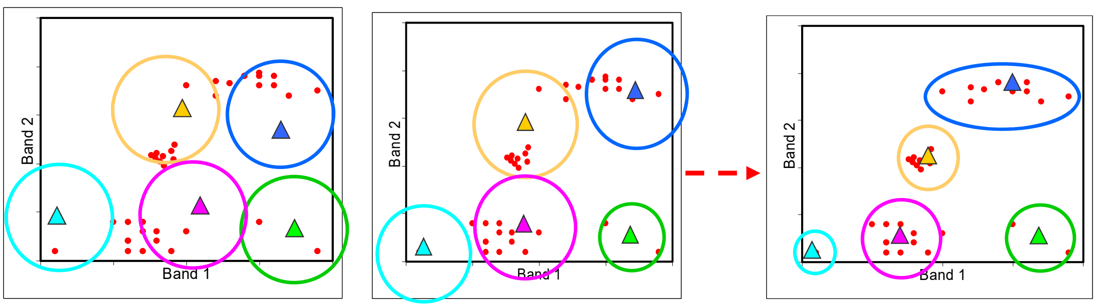
2. Pixels are assigned to their nearest cluster.

3. The mean location and variance (shape) are re-calculated for each cluster.

4. Repeat 2 and 3 until movement of cluster centres is below threshold.

5. Assign class types to spectral clusters.

Example k-means



1. First iteration. The cluster centres are set at random. Pixels will be assigned to the nearest centre.

2. Second iteration. The centres move to the mean-centre of all pixels in this cluster.

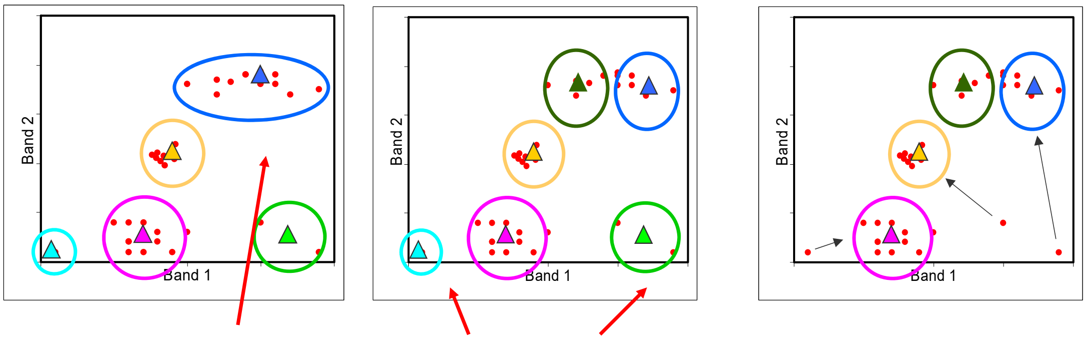
3. N-th iteration. The centres have stabilised.

Universität Stuttgart
ifp

ISODATA (unsupervised)

- Extends k-means. Also consider shape (standard deviation) of clusters.
- After stage 3 we can either:
  - Combine clusters if centres are close.
  - Split clusters with large standard deviation in any dimension.
  - Delete clusters that are too small.
- Then reclassify each pixel and repeat.
- Stop after max. iterations or at convergence limit.
- Assign class types to spectral clusters.

Example ISODATA



1. Data is clustered but blue cluster is very stretched in band 1.

2. Cyan and green clusters only have 2 or less pixels. So they will be removed.

3. Either assign outliers to nearest cluster, or mark as unclassified.

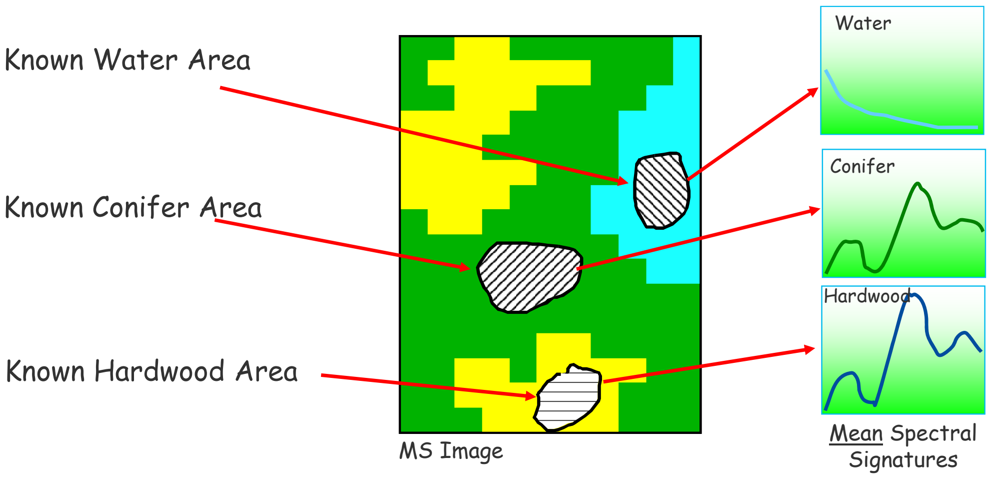
Application of unsupervised classification

- Unsupervised classification can often produce information that is not obvious to visual inspection.
- Very useful for areas where ‘ground truth’ data is difficult to obtain.
- However, results may not coincide with desired land cover classes.
- Often useful to trigger subsequent supervised classification.

Supervised Classification

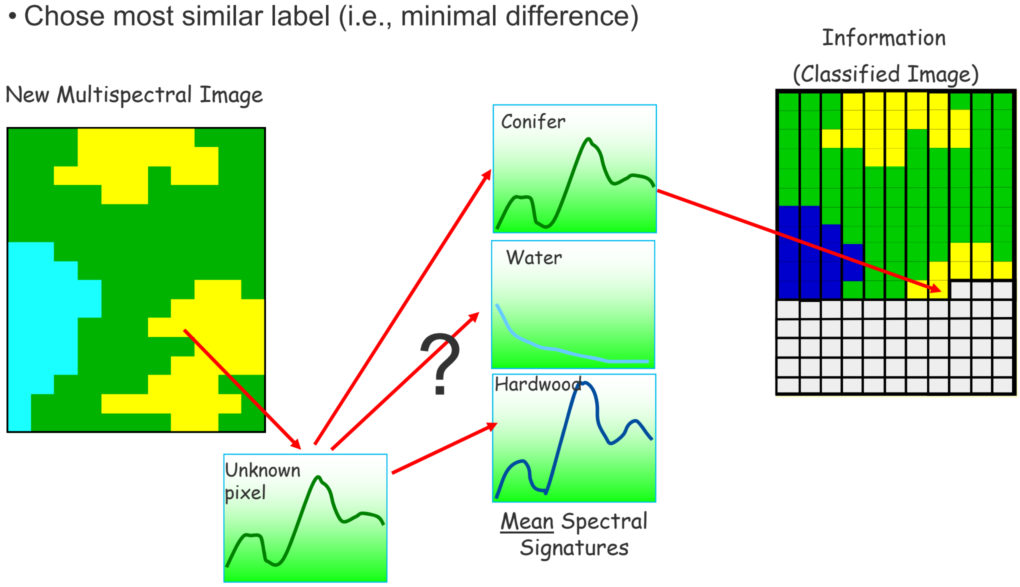
Supervised classification: Training

- Number m and type of land cover classes are known.
- Training regions are created for each class
- Classifier "learns" *mean signature* (n-dimensional vector) each class from set of samples



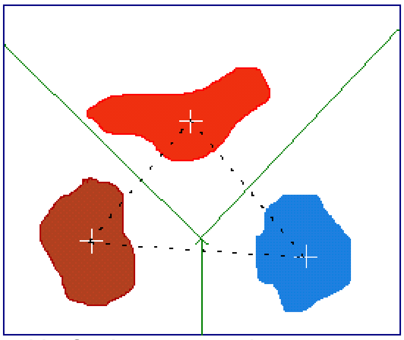
Classification phase

- Compare features of each pixel with set of signatures
- Chose most similar label (i.e., minimal difference)

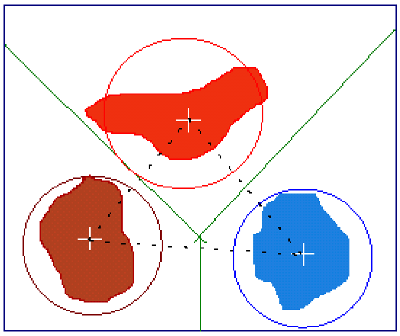


Minimum Distance Classifier

- Calculates mean of the spectral values for the training set for each class.
- Measures the distance from a pixel of unknown class to the mean of each class.
- Assigns the pixel to the class with the shortest distance.
- Assigns a pixel as "unknown" if the pixel is beyond the distances defined by the analyst (optional).

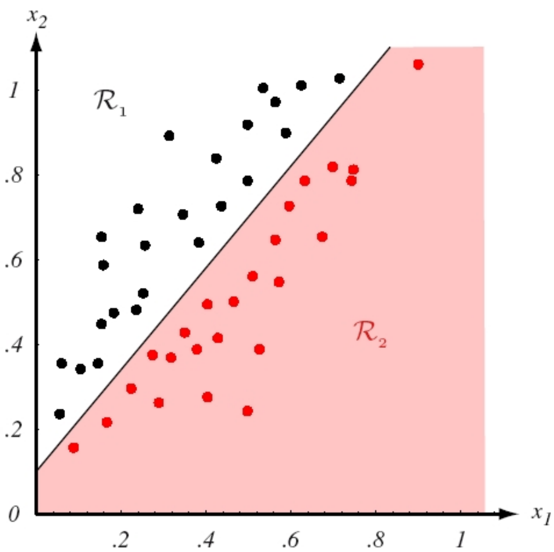


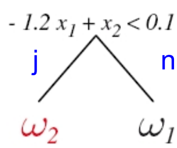
No further constraint



Reject if beyond radius

Hierarchical decision tree

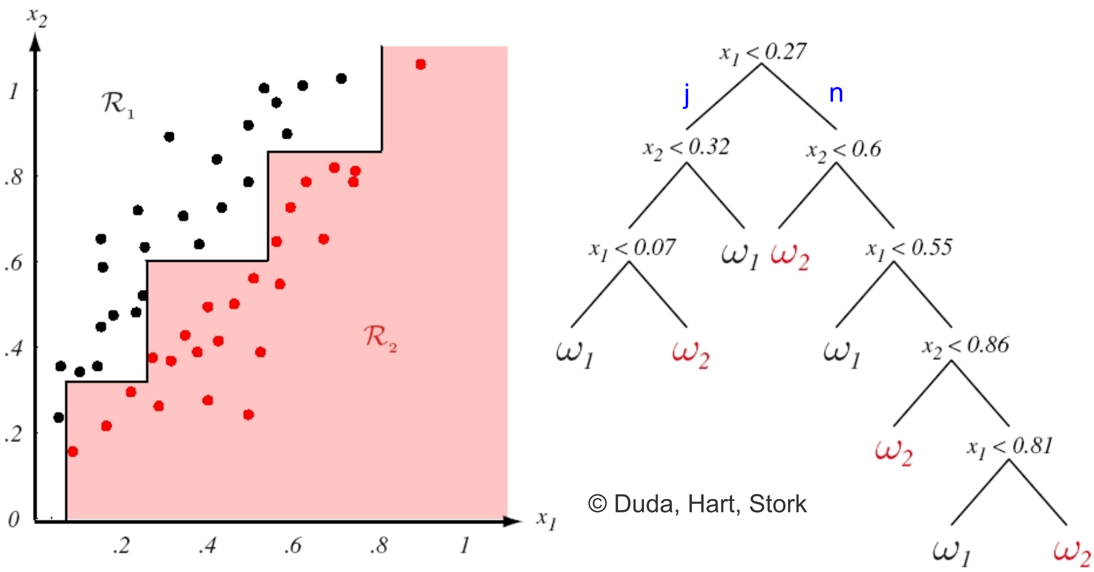




© Duda, Hart, Stork

- According to thresholds the data are divided step by step
- The thresholds are either set manually or derived by training

Hierarchical decision tree

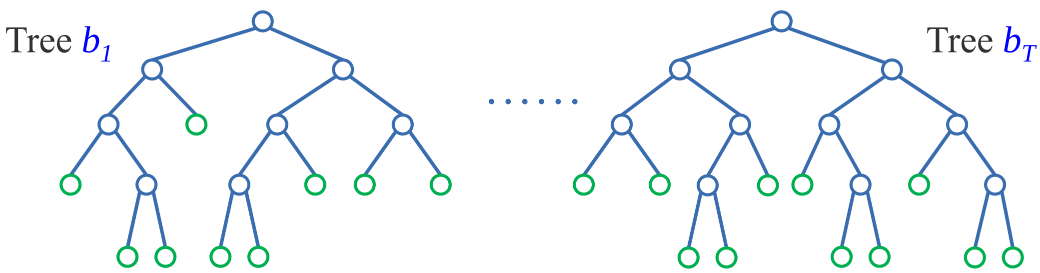


© Duda, Hart, Stork

- However, the same data can be separated in many ways.
- Subjective, tendency to “overfitting”

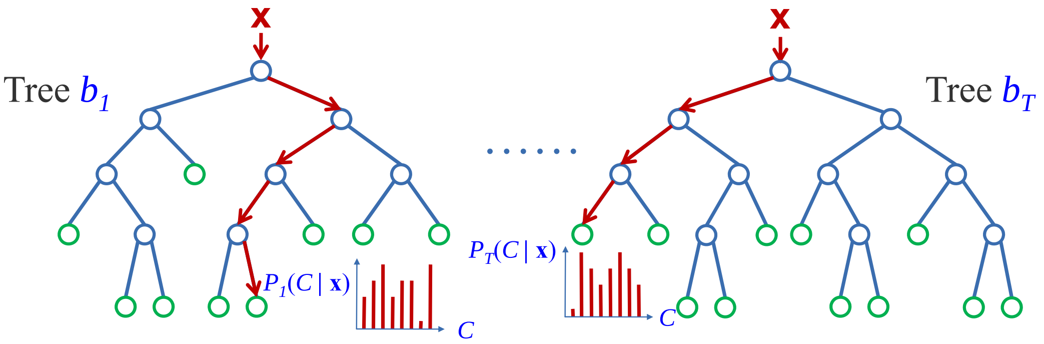
Random Forests

- A Random Forest [Breiman, 2001] consists of T decision trees.
- Training: fetures are randomly selected.
- Each tree is a weak classifier, the **ensemble classifier** however is strong!
- Classification: A feature vector $\mathbf{x}$ is classified by each tree.



Random Forests

- A Random Forest [Breiman, 2001] consists of T decision trees.
- Training: fetures are randomly selected.
- Each tree is a weak classifier, the **ensemble classifier** however is strong!
- Classification: A feature vector $\mathbf{x}$ is classified by each tree.



In every tree t: posterior distribution $P_t(C \mid \mathbf{x})$ according to tree t

后验

- Posterior: average probability: $$P(C \mid \mathbf{x}) = \frac{1}{T} \cdot \sum_{t=1}^{T} P_t(C \mid \mathbf{x})$$

Training of Random Forest

- Draw with replacement T so-called bootstrap test data sets (e.g. $T = 50$), subsets of initial set.
- For each of those sets t we train one tree b_t.
- Important: independent drawing of bootstrap subsets.
- By combining results from different trees:
  - Better generalization
  - Higher stability

并发

- Easy to implement as parallel process (concurrency)

Bayesian Classification

- Generative approach:
  - The posterior probability $p(C|\mathbf{x})$ is maximized.
  - Posterior $p(C|\mathbf{x})$ is modelled indirectly according to the Theorem of Bayes.
  - This requires a model of the joint distribution $p(C,\mathbf{x})$ of the data $\mathbf{x}$ and the class labels C.
  - It is possible to *generate* synthetic data sets by sampling from the joint distribution.

- Strong theoretical foundation:
  - If the required distributions are known, Bayesian classification will deliver the result with the lowest proportion of classification errors!

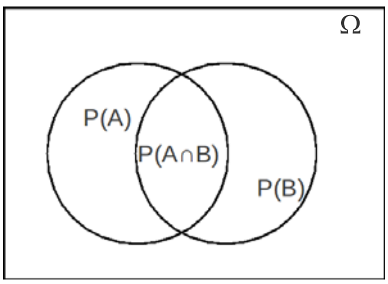
Motivation: Recap probabilities I

- A subset A of a population Ω suffers from cancer. By normalization we yield a probability that a person we sample carries this disease:

$$\frac{|A|}{|\Omega|} = P(A)$$

- A drug company invents some screening test, which is either "positive" (indicating cancer) for some people (set B) and "negative" for the rest:

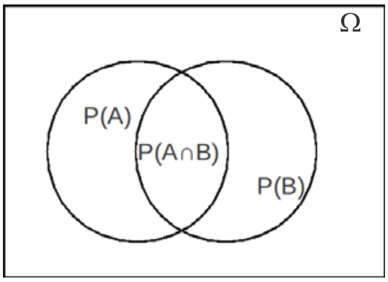
$$\frac{|B|}{|\Omega|} = P(B)$$



Motivation: Recap probabilities II

- The joint probability A,B (shorthand $A \cap B$) is: $\frac{|A,B|}{|\Omega|} = P(A,B)$

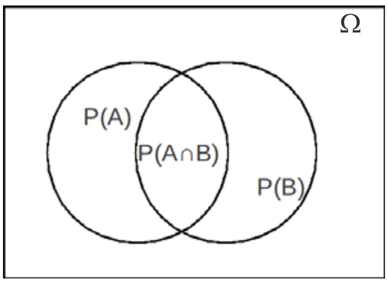
- We ask: “Given that the test is positive for a randomly selected individual, what is the probability that said individual has cancer?”
  - This is a conditional probability $P(A|B) = \frac{|A,B|}{|B|} = \frac{\frac{|A,B|}{\Omega}}{\frac{|B|}{\Omega}} = \frac{P(A,B)}{P(B)}$



Motivation: Recap probabilities III

- Now let us ask “Given that a randomly selected individual has cancer (event A), what is the probability that the test is positive for that individual (event A,B)?”
  - This is of course again a conditional probability: $P(B|A) = \frac{P(A,B)}{P(A)}$
  - We have now: $P(A|B) = \frac{P(A,B)}{P(B)}$ and $P(B|A) = \frac{P(A,B)}{P(A)}$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$



Theorem of Bayes: Derivation for our purpose

- For the joint distribution $p(\mathbf{x},C)$ of data $\mathbf{x}$ and classes C the product rule applies:

$$p(\mathbf{x},C) = p(C|\mathbf{x}) \cdot p(\mathbf{x})$$

- Likewise: $$p(C,\mathbf{x}) = p(\mathbf{x}|C) \cdot p(C)$$

- Due to $p(\mathbf{x},C) = p(C,\mathbf{x})$:

$$p(C|\mathbf{x}) \cdot p(\mathbf{x}) = p(\mathbf{x}|C) \cdot p(C)$$

- Therefore: $$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C) \cdot p(C)}{p(\mathbf{x})}$$

Theorem of Bayes

Theorem of Bayes: Interpretation

- Causal relation between object type and observed features: the observed features are a function of the object type.

- Usually it is easier to deduce the effect from the cause, i.e., it would seem to be easier to deduce the features from the object type.

- The theorem of Bayes allows inverse reasoning : derive information about the cause (the object type) from the effect (the observed features).

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C) \cdot p(C)}{p(\mathbf{x})}$$

Theorem of Bayes: Meaning of the terms I

• $p(C)$: Prior probability

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C)\cdot p(C)}{p(\mathbf{x})}$$

- Corresponds to knowledge (bias) for the occurrence of C.
- If no information is available: Uniform Distribution
  → MAP becomes Maximum-Likelihood (ML)
- $p(C)$ can be determined iteratively:
  1. Classification under the assumption of a uniform distribution of the occurrence of the individual classes.
  2. Determination of $p(C)$ from the relative frequencies of occurrence of the individual classes C^k.
  3. Classification according to the theorem of Bayes.

Theorem of Bayes: Meaning of the terms II

• $p(\mathbf{x}|C)$: Likelihood

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C)\cdot p(C)}{p(\mathbf{x})}$$

- Probability to observe $\mathbf{x}$ if it is known to belong to class C.
- Note: the Likelihood is **no probability density function of the Classes C**!
- For each class C^k there is a model for $p(x|C= C^k)$, which describes the distribution of the features for this class.
- Determination from data in training areas
- *Non-parametric Models*: Direct determination of $p(\mathbf{x}|C)$ from the **training data**.
- *Parametric Models*: Based on the assumption of an **analytical model** for $p(\mathbf{x}|C)$, whose **parameters** are estimated from the training data.

Theorem of Bayes: Meaning of the terms III

- $p(\mathbf{x})$: Probability of the data
  (also called evidence)

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C)\cdot p(C)}{p(\mathbf{x})}$$

  - Equal for all values of C becauses it does not depend on C.
    $\Rightarrow$ MAP can also be applied without knowing $p(\mathbf{x})$:

$$p(C|\mathbf{x}) \propto p(\mathbf{x}|C)\cdot p(C)$$

$$\Rightarrow \max\left(p(C|\mathbf{x})\right) = \max\left(p(\mathbf{x}|C)\cdot p(C)\right)$$

  - $p(\mathbf{x})$ ensures that $p(C|\mathbf{x})$ can be interpreted as a probability and can be used as such in further probabilistic processes.
  - $p(\mathbf{x})$ can be determined as the 边界 **marginal distribution** of $p(\mathbf{x},C)$:

$$p(\mathbf{x}) = \sum_k p(\mathbf{x}|C^k)\cdot p(C^k)$$

Theorem of Bayes: Example

- It is known that from 100000 people 20 suffer from a certain severe illness:

$$p(K = ill) = 0.0002,\ p(\bar{K} = healthy) = 0.9998$$

- It exists a screening method for this disease:
  - Sensitivity of the tests: 95% of all ill persons are detected ($T=1$):

$$p(T|K) = 0.95,\ p(\bar{T}|K) = 0.05$$

  - Unfortunately, the test delivers false positive result for 1% of healthy persons:

$$p(T|\bar{K}) = 0.01,\ p(\bar{T}|\bar{K}) = 0.99$$

- We may be interested in the portion of ill persons in the set of all persons with positive test results:

$$p(K|T) = \frac{p(T|K)\,p(K)}{p(T|K)\,p(K) + p(T|\bar{K})\,p(\bar{K})} = \frac{0.95*0.0002}{0.95*0.0002+0.01*0.9998} = 0.0186$$

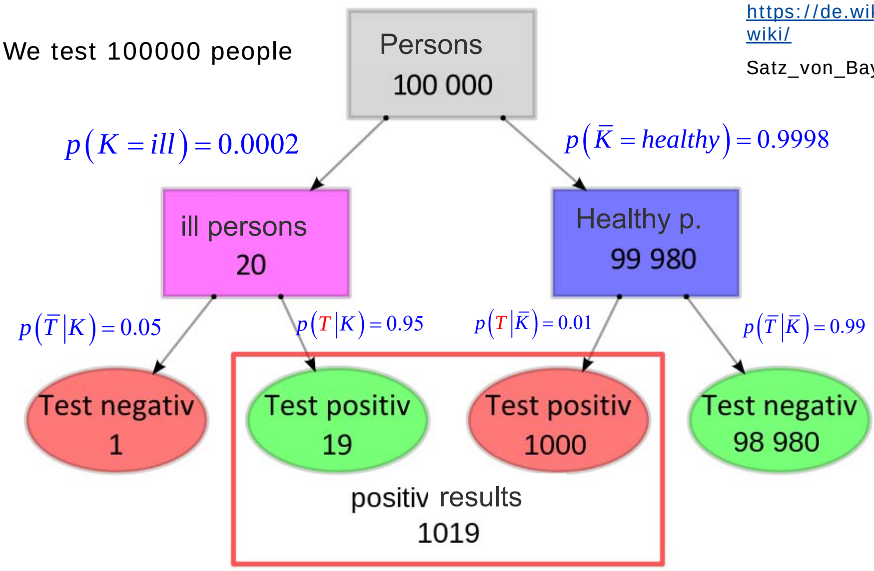
$p(T)$: Sum over all classes (here: 2)

https://de.wikipedia.org/wiki/Satz_von_Bayes

Visualization using event tree I

We test 100000 people

https://de.wikipedia.org/wiki/
Satz_von_Bayes

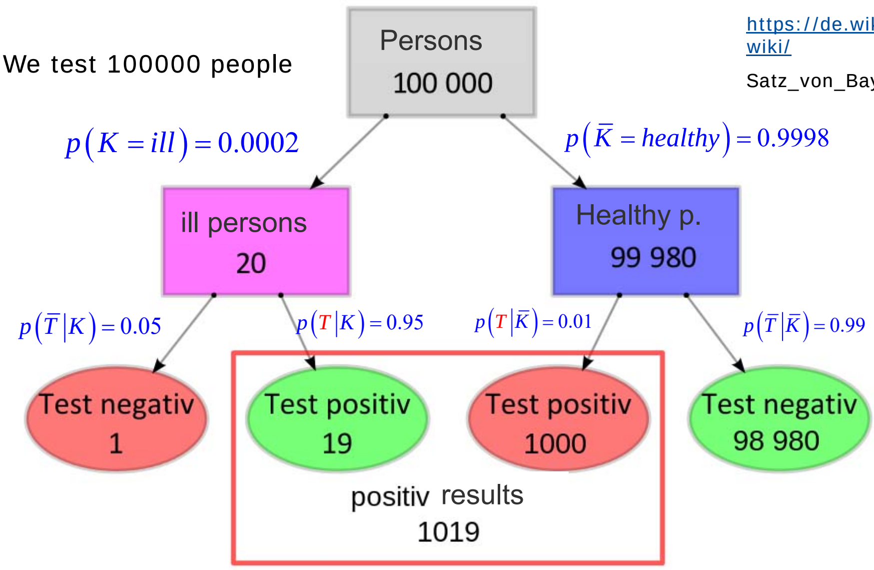


$$p(K|T) = \frac{p(T|K)\,p(K)}{p(T|K)\,p(K) + p(T|\bar{K})\,p(\bar{K})} = \frac{0.95 \cdot 0.0002}{0.95 \cdot 0.0002 + 0.01 \cdot 0.9998} \approx 0.0186$$

Visualization using event tree II

We test 100000 people

https://de.wikipedia.org/wiki/
Satz_von_Bayes



Other conclusion possible, e.g.:

$$p(\bar{K}|T) = 1 - p(K|T) = \frac{p(T|\bar{K})\,p(\bar{K})}{p(T|K)\,p(K) + p(T|\bar{K})\,p(\bar{K})} = \frac{0.01 \cdot 0.9998}{0.95 \cdot 0.0002 + 0.01 \cdot 0.9998} \approx 0.9814$$

Workflow of Bayesian classification

- Given:
  - Models for the likelihoods $p(\mathbf{x}|C^k)$ of all classes C^k
  - Priori probabilities $p(C^k)$ of all classes C^k
  - A feature vector $\mathbf{x}$ to be classified

- Wanted: Class C_{map} of $\mathbf{x}$ according to the MAP criterion.

- Procedure:
  1. For all C^k: calculate $p(\mathbf{x}, C^k) = p(\mathbf{x}|C^k) \cdot p(C^k)$
  2. Calculate $p(\mathbf{x}) = \sum_k p(\mathbf{x}|C^k) \cdot p(C^k)$
  3. For all C^k: calculate $p(C^k|\mathbf{x}) = p(\mathbf{x}, C^k) / p(\mathbf{x})$
  4. C_{map} results as the label C^k for which $p(C^k|\mathbf{x})$ is a maximum.

Training

- **Training: provision of examples**
  - User marks image regions which correspond to a class C^k.
  - Assumption: all pixels in the selected region belong to C^k.
  - Training areas must be provided for all classes
  - The training data must be representative for all classes
- **Modelling of the likelihood for the classes:**
  - Based on training data
  - Different for parametric and non-parametric methods.

Maximum Likelihood Method

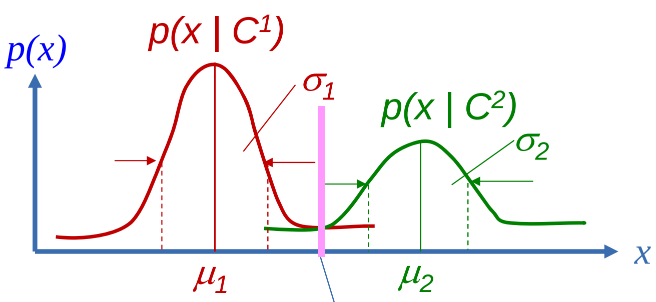
- Special case: prior probability unknown → only Likelihood
- The classes C are often modelled to as multivariate normal distribution over feature space $\mathbf{x}$.
- Estimation of expectation value vector μ_k and Covariance matrix Σ_k of features $\mathbf{x}$ from training areas of each class C_k.
- For each pixel we infer the probability $p(\mathbf{x} \mid C_k)$ of each class C_k from the features $\mathbf{x}$:

$$p(\mathbf{x} \mid C_k) = \frac{1}{(2\pi)^{N/2} \cdot |\boldsymbol{\Sigma}_k|^{1/2}} \cdot e^{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_k)^T \cdot \boldsymbol{\Sigma}^{-1} \cdot (\mathbf{x}-\boldsymbol{\mu}_k)\right]}$$

- The pixel to be classified is labeled to belong to the class of the highest probability.

Maximum Likelihood: Example

- **Example** (single Band, two classes C^1, C^2)
  - The feature space is of dimension 1

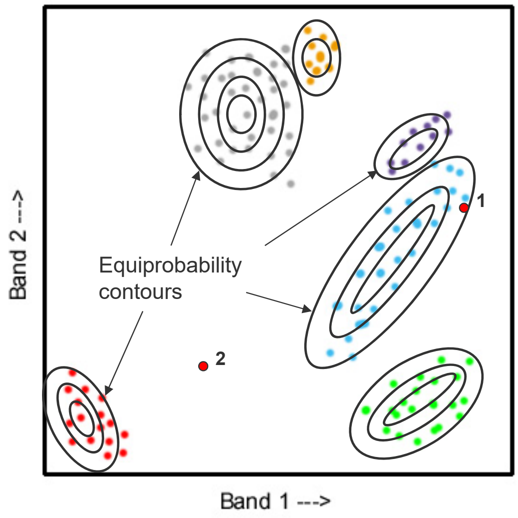


Threshold: value x, for which holds:

$$p(x \mid C^1) = p(x \mid C^2)$$

Maximum likelihood: 2D example

- Normal distributions are fitted to each training class.
- The lines in the diagram show regions of equal probability.
- Point 1 would be assigned to bright blue class as this is most probable.
- Point 2 would generally be unclassified as the probabilities of fitting into one for the classes would be below threshold.



How to quantify classification performance?

Accuracy Assessment: Error Matrix

- The error matrix reveals the classification accuracy
- Quantifying accuracy
  - *Total Accuracy*: Number of correct plots / total number of plots

	Class types determined from reference source			
Class types determined from classified map				
# Plots	Conifer	Hardwood	Water	Totals
Conifer	50	5	2	57
Hardwood	14	13	0	27
Water	3	5	8	16
Totals	67	23	10	100

Diagonals represent sites classified correctly according to reference data

Off-diagonals were mis-classified

$$Accuracy_{Total} = \frac{50+13+8}{\text{total number of plots}} \cdot 100 = 71\%$$

Accuracy Assessment: Total Accuracy

- Total Accuracy:
  - Number of correct plots / total number of plots
- Problem with total accuracy:
  - Summary value is an average
    - Does not reveal if error was evenly distributed between classes or if some classes were really bad and some really good.
- Therefore, include other forms:
  - User's accuracy
  - Producer's accuracy

Accuracy Assessment: User's Accuracy

- From the perspective of the user of the classified map, how accurate is the map?
  - For a given class, how many of the pixels on the map are actually what they say they are?
  - Calculated as:

Number correctly identified in a given map class /
Number claimed to be in that map class

Accuracy Assessment: User's Accuracy

- User's accuracy corresponds to error of commission (inclusion):
  - E.g., 5 hardwood and 2 water pixels labeled erroneously as conifer

	Class types determined from reference source				
Class types determined from classified map	# Plots	Conifer	Hardwood	Water	Totals
	Conifer	50	5	2	57
	Hardwood	14	13	0	27
	Water	3	5	8	16
	Totals	67	23	10	100

Example: Conifer

$$Accuracy_{User's,Conifer} = \frac{50}{57} \cdot 100 = 88\%$$

Accuracy Assessment: Producer's Accuracy

- From the perspective of the maker of the classified map, how accurate is the map?
  - For a given class in reference plots, how many of the pixels on the map are labeled correctly?
  - Calculated as:

Number correctly identified in ref. plots of a given class /
Number actually in that reference class

Accuracy Assessment: Producer's Accuracy

- Producer's accuracy corresponds to error of omission (exclusion):
  - Here 14 hardwood and 3 water pixels are excluded from correct label

	Class types determined from reference source				
	# Plots	Conifer	Hardwood	Water	Totals
Class types determined from classified map	Conifer	50	5	2	57
	Hardwood	14	13	0	27
	Water	3	5	8	16
	Totals	67	23	10	100

Example: Conifer

$$Accuracy_{producers,Conifer} = \frac{50}{67} \cdot 100 = 75\%$$

Error Matrix with User's and Producer's Accuracy

	Class types determined from reference source					
	# Plots	Conifer	Hardwood	Water	Totals	User's Accuracy
Class types determined from classified map	Conifer	50	5	2	57	88%
	Hardwood	14	13	0	27	48%
	Water	3	5	8	16	50%
	Totals	67	23	10	100	
Producer's Accuracy		75%	57%	80%		Total: 71%

Example: Corine Land Cover (CLC) 1990

EU program, member states obliged to contribute

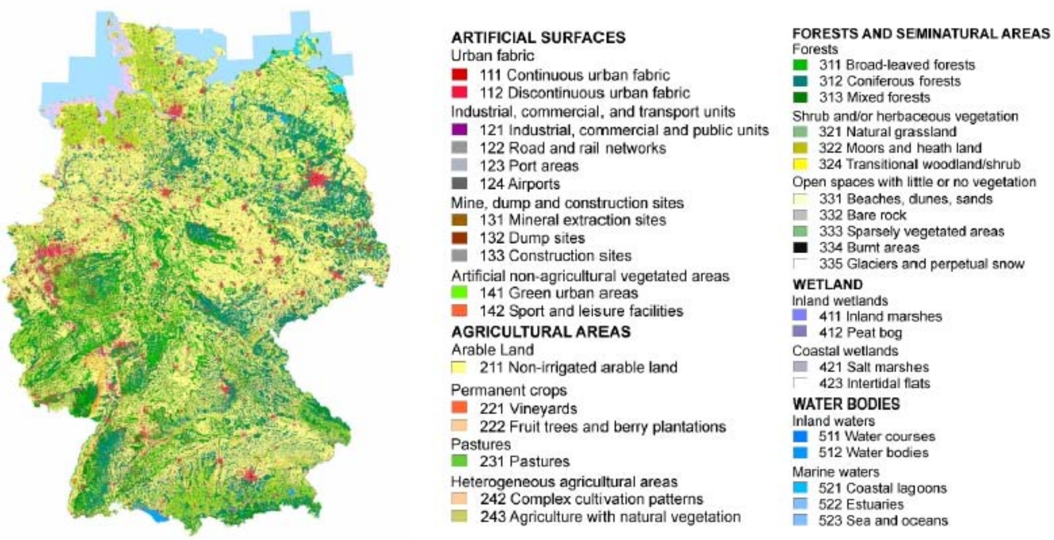


Figure 3. Land cover map Germany CLC1990 and legend showing 36 land cover classes for Germany