Christian-Albrechts-Universität zu Kiel

Intelligent Systems

Chapter 11: Evaluation

Winter Term 2019 / 2020

Prof. Dr.-Ing. habil. Sven Tomforde
Institute of Computer Science / Intelligent Systems group

About this Chapter

Christian-Albrechts-Universität zu Kiel

Content

- Evaluation
- Bias and variance
- Model testing
- Validation
- Conclusion
- Further readings

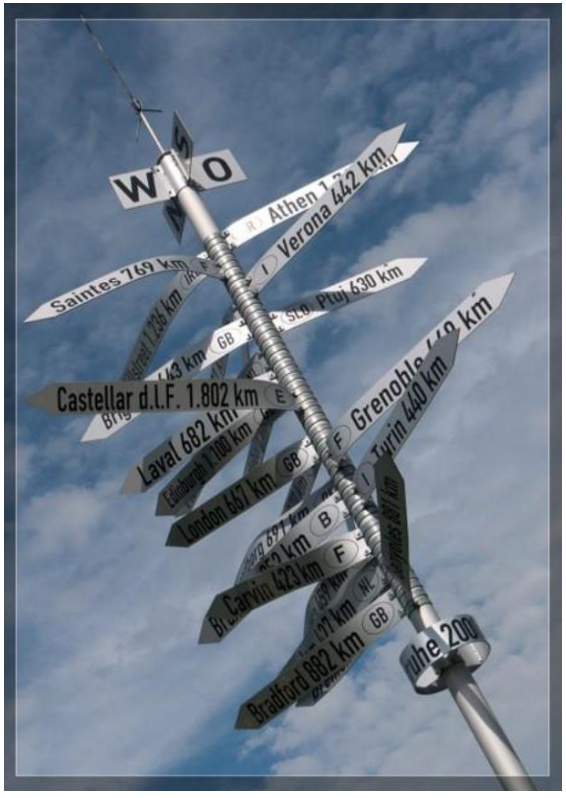
Goals

Students should be able to:

- explain how evaluation is done for classifiers.
- introduce the bias vs. variance dilemma.
- compare the advantages and disadvantages of model testing approaches.
- define and apply different validation techniques.

Agenda

- Evaluation
- Bias and variance
- Model testing
- Validation
- Conclusion
- Further readings



Evaluation

- So far: Training of a model based on example data
- Important for application: Assessment of the expected performance of a trained model
  - Can the model fulfil the task?
  - To what extent can the statements/decisions of the model be trusted?
- In general, there is only a finite amount of training data available
- However, the model is also intended to provide correct decisions for previously unseen samples
  - Good generalisation ability desired
  - How can this ability be numerically evaluated?

Evaluation (2)

Generalisation capability:

A model instance whose parameters have been set using a search algorithm and an evaluation function should provide minor errors not only for the data set used (when setting the parameters, i.e. learning), but also for (all) other data sets that are based on the same function (or task) to be modelled.

Goal:

A model instance should deliver good results in a later application to new (so-called "unknown") data.

Evaluation (3)

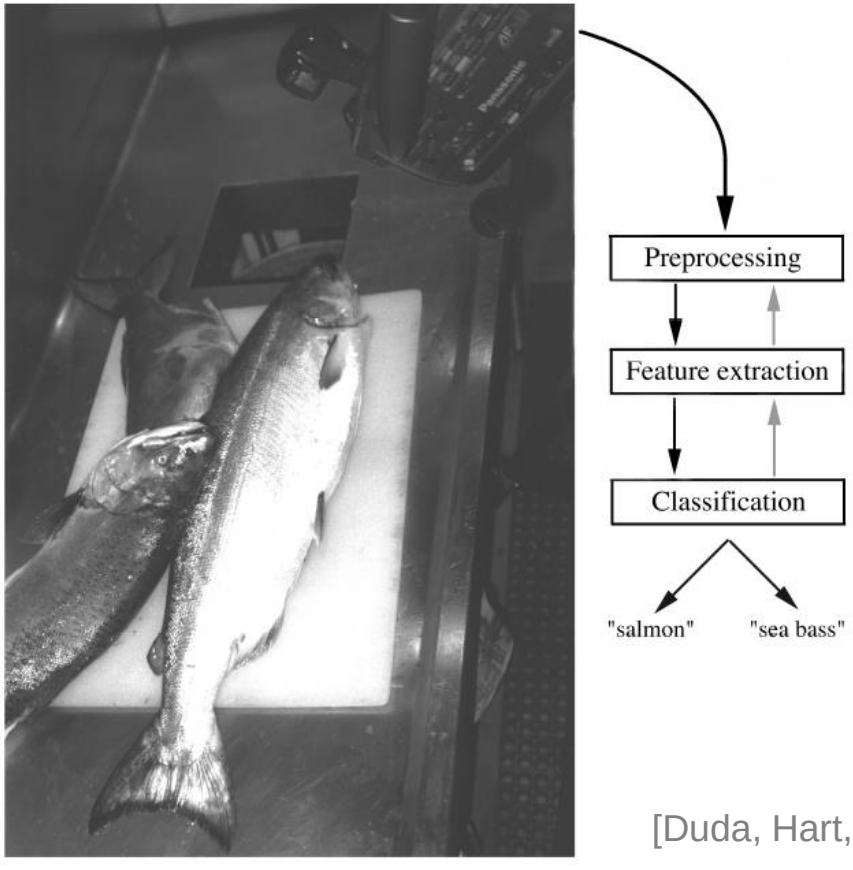
Evaluation of results

- Terms:
  - Training data: Data used to create the model instance.
  - Test data: Unknown data used to test the model instance prior to its application in order to evaluate its generalisation capability.
- Observations:
  1. In general, a model instance provides higher errors for test data than for training data.
  2. The more complex a model is, that is, the more parameters (degrees of freedom) it has, the smaller the training error will generally be and the more the test error becomes larger.

Evaluation (4)

Christian-Albrechts-Universität zu Kiel

Back to our fish example:



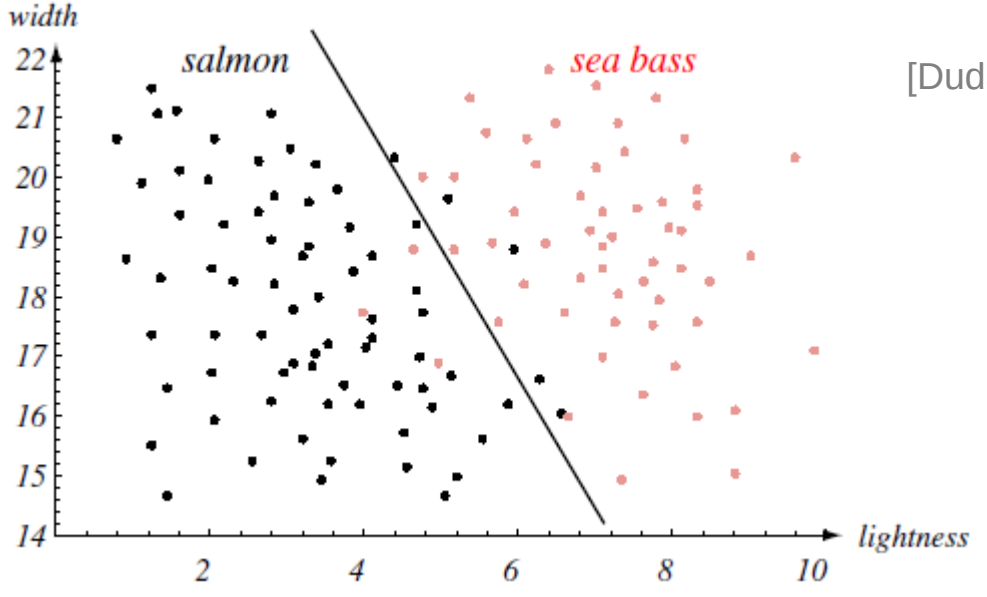
[Duda, Hart, Stork, 2001]

Evaluation (5)

Christian-Albrechts-Universität zu Kiel

Evaluation and assessment of results

- Approach 1: Simple model, simple decision boundary (attempt of linear separation)



[Duda, Hart, Stork, 2001]

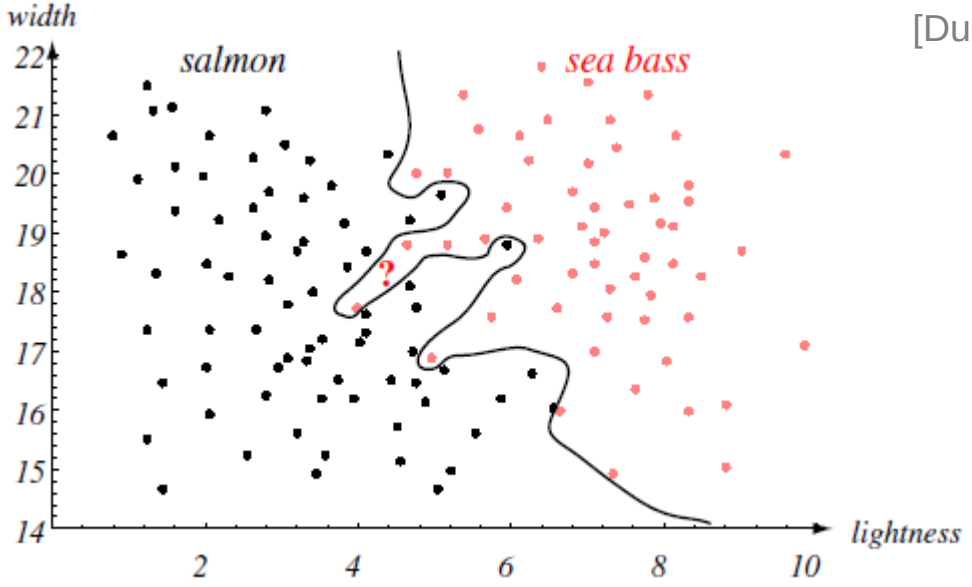
- The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier

Evaluation (6)

Christian-Albrechts-Universität zu Kiel

Evaluation and assessment of results

- Approach 2: Complex model, very complex decision boundary (perfect separation)



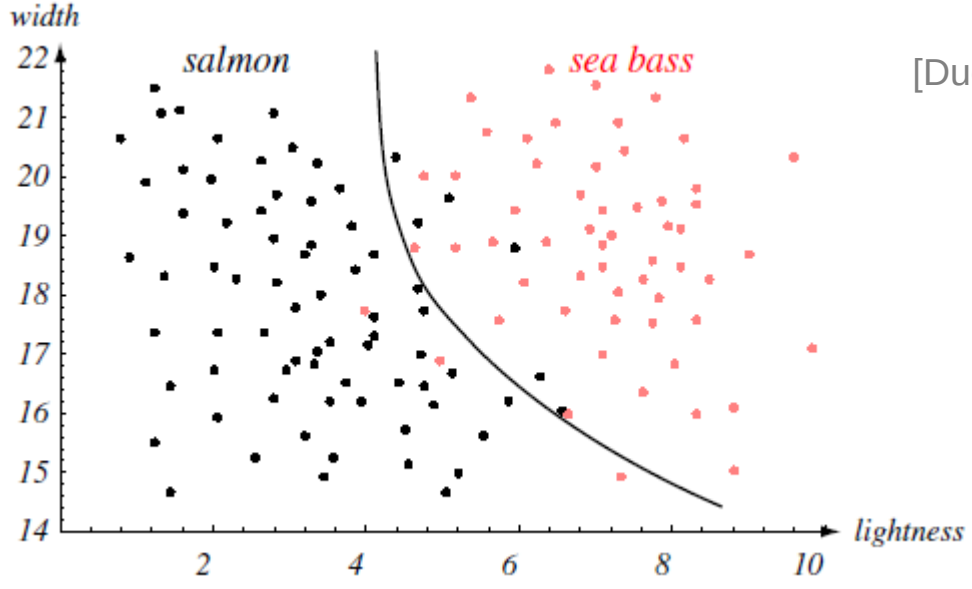
[Duda, Hart, Stork, 2001]

Evaluation (7)

Christian-Albrechts-Universität zu Kiel

Evaluation and assessment of results

- Approach 3: Compromise (may come with the best generalisation capability)



[Duda, Hart, Stork, 2001]

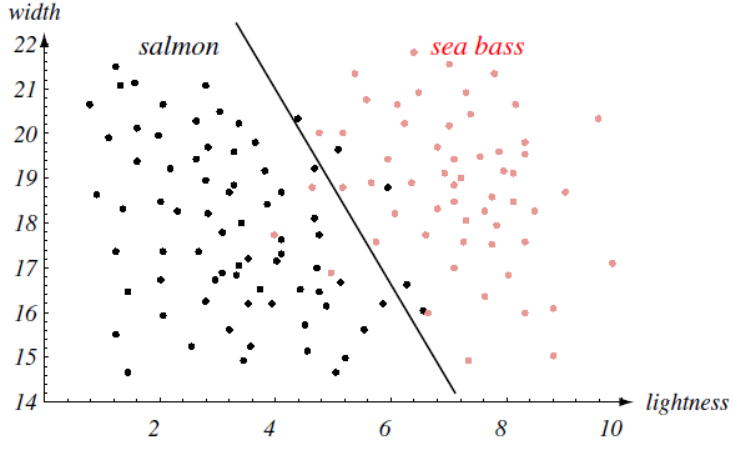
Evaluation (8)

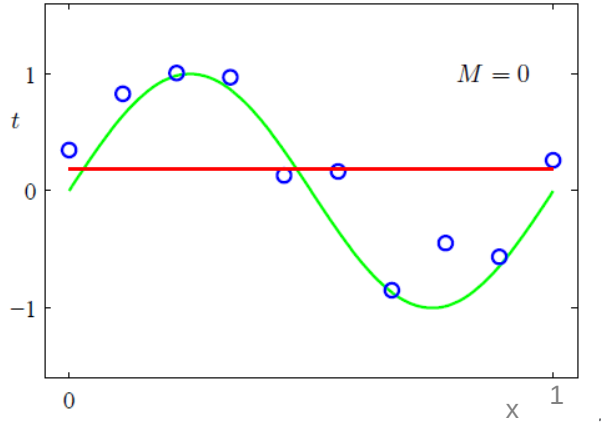
Christian-Albrechts-Universität zu Kiel

Underfitting

- If the model is too simple (too inflexible), it cannot describe the data (or their underlying structure) well.
  → Underfitting
- Consequence: No good prediction (or classification) of new data, as the relevant relationships are not mapped.
- Model is not sufficiently complex
  - Is "too easy" for the task

[Duda, Hart, Stork, 2001]





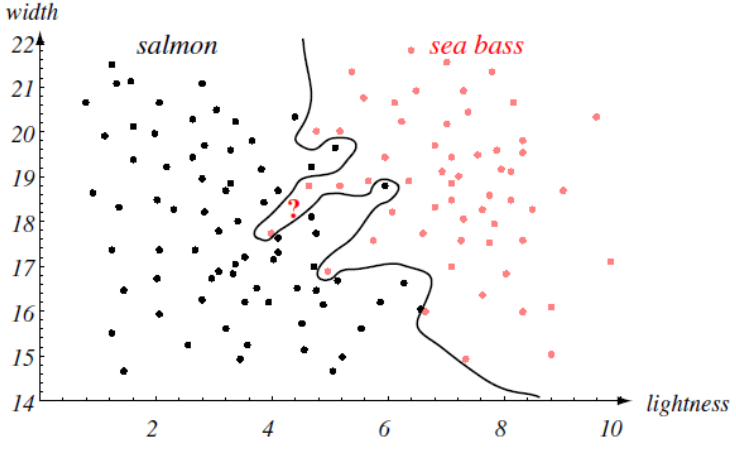
Evaluation (9)

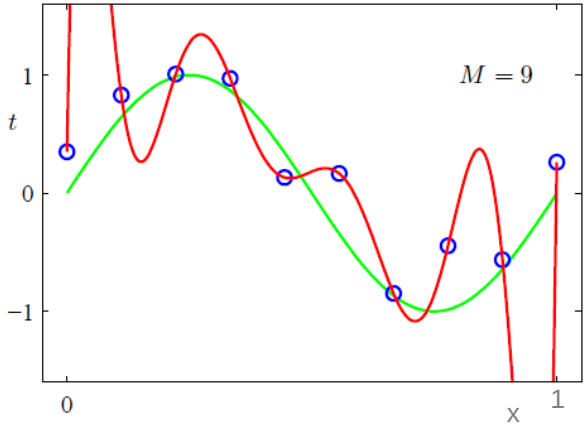
Christian-Albrechts-Universität zu Kiel

Overfitting

- An overfitting of models to training data (poor generalisation capability) is possible:
  - with high model complexity (many degrees of freedom),
  - with few training patterns,
  - in the course of a parameter adaptation process with many iterative search algorithms,
  - ...

[Duda, Hart, Stork, 2001]

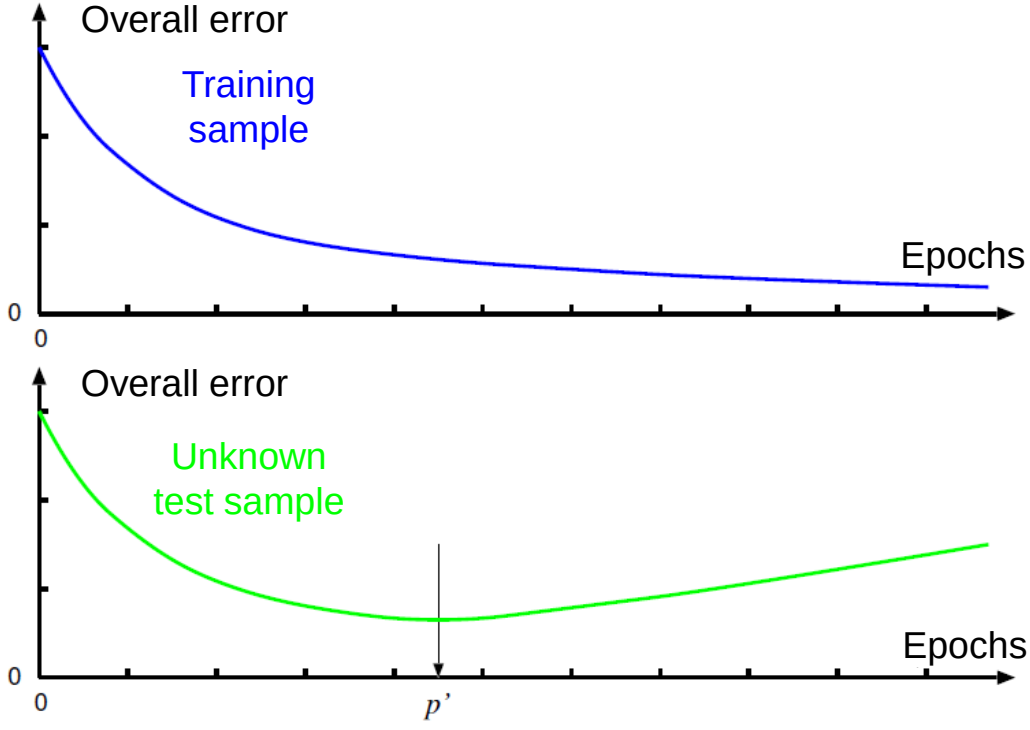




Evaluation (10)

Christian-Albrechts-Universität zu Kiel

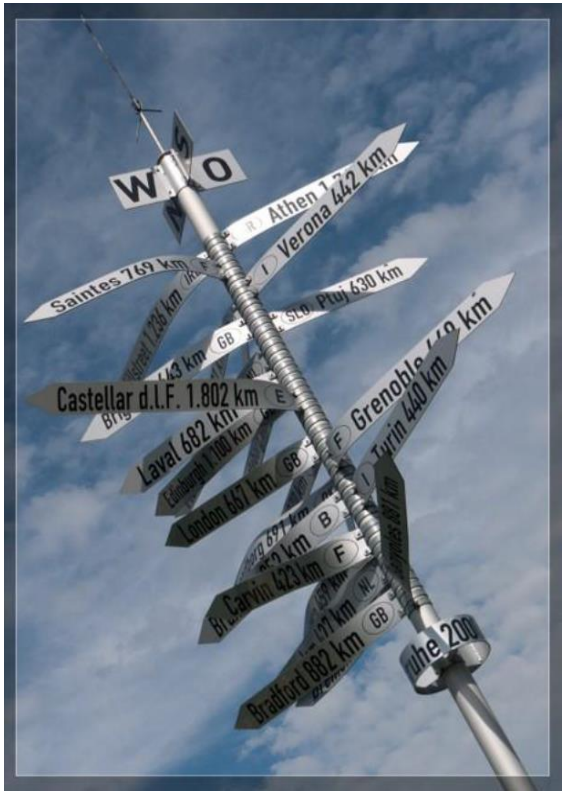
Overfitting



- Example: When training a neural network, but also with other paradigms

Agenda

- Evaluation
- Bias and variance
- Model testing
- Validation
- Conclusion
- Further readings



Bias and variance

Bias and Variance

- Previously: More informal view of over-/under-adjustment
- Goal: Mathematical/Probabilistic consideration of the problem of model complexity
  → Bias vs. Variance Dilemma

Terms:

- Bias: describes the accuracy of models (high bias - low accuracy)
- Variance: describes the "stability"/variability of models (high variance - low stability)

Example: Bias and Variance in Regression Tasks

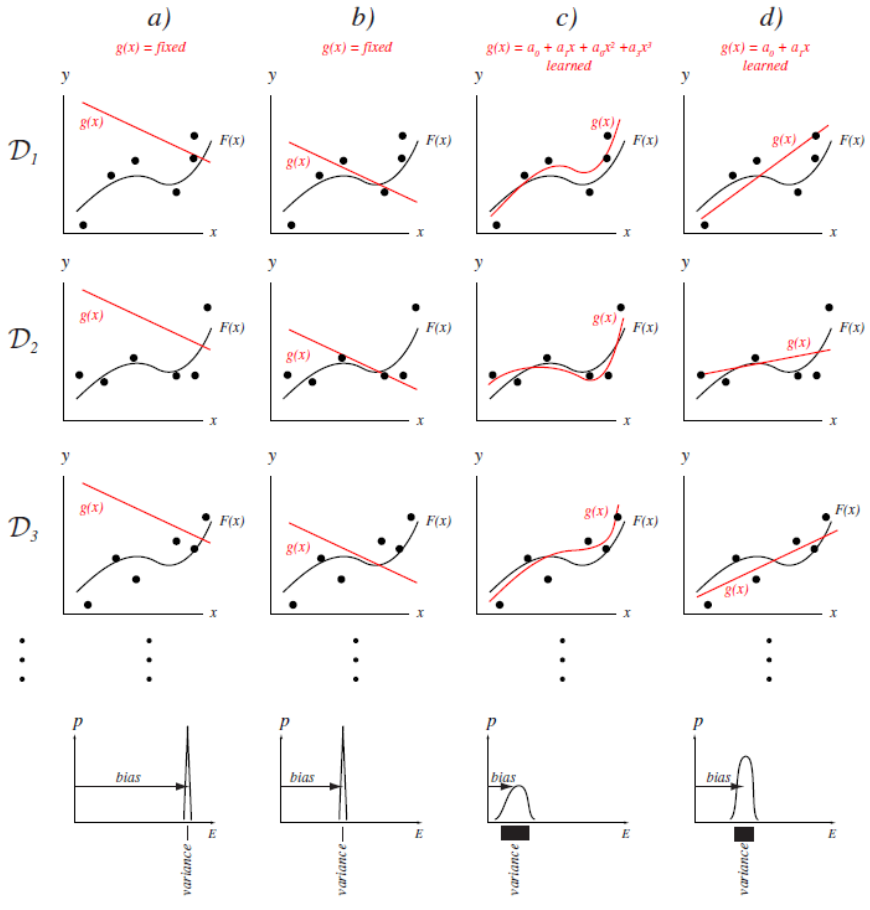
Bias and variance (2)

Definitions

- $F(x)$: function to be modelled
- D_i: Sets of samples of this function
- $g(x)$: Model
- E: Error of a model
- p: Frequency

Illustration

- Columns: different model types
- Lines: different sample sets



[Duda, Hart, Stork, 2001]

Bias and variance (3)

Bias and Variance

- Let $g(x, D)$ be the approximation of $F(x)$ based on the data set D.
- Depending on the choice of the data set, the approximation is
  - sometimes be good
  - sometimes be bad
- Formal: Quality defined as square distance between model and approximated function.

$$(g(x, D) - F(x))^2$$

- Based on the notation $\mathcal{E}[.]$ for the expected value, holds for all possible datasets D:

$$\mathcal{E}_D[(g(x, D) - F(x))^2] =$$

$$\underbrace{\left(\mathcal{E}_D[g(x, D)] - F(x)\right)^2}_{\text{Bias}^2} + \underbrace{\mathcal{E}_D[(g(x, D) - \mathcal{E}_D[g(x, D)])^2]}_{\text{Variance}}$$

Bias and variance (4)

Derivation

$$\begin{aligned}
&\big(g(x,D) - F(x)\big)^2 \\
&= \big(g(x,D) - \mathcal{E}_D[g(x,D)] + \mathcal{E}_D[g(x,D)] + F(x)\big)^2 \\
&= (g(x,D) - \mathcal{E}_D[g(x,D)])^2 + \big(\mathcal{E}_D[g(x,D)] + F(x)\big)^2 \\
&+2 \cdot (g(x,D) - \mathcal{E}_D[g(x,D)]) \cdot \big(\mathcal{E}_D[g(x,D)] + F(x)\big)
\end{aligned}$$

- This only applies to a specific record.
- Therefore: Average of all data records D.
  - Expected value $\mathcal{E}_D$: The last summand disappears.

$$\begin{aligned}
&\mathcal{E}_D\left[\big(g(x,D) - F(x)\big)^2\right] \\
&= \big(\mathcal{E}_D[g(x,D)] - F(x)\big)^2 + \mathcal{E}_D[(g(x,D) - \mathcal{E}_D[g(x,D)])^2]
\end{aligned}$$

Bias and variance (5)

Bias and Variance - Interpretation

$$\underbrace{\left(\mathcal{E}_D[g(x,D)] - F(x)\right)^2}_{\text{Bias}^2} + \underbrace{\mathcal{E}_D[(g(x,D) - \mathcal{E}_D[g(x,D)])^2]}_{\text{Variance}}$$

The expected quadratic difference between $g(x, D)$ and $F(x)$ can therefore be expressed by the sum of two terms:

- The first summand (squared bias) describes how strongly the average prediction of all data sets deviates from the optimal prediction.
  - Low bias: $F(x)$ is accurately approximated on average
- The second summand (variance) describes the extent to which the solutions for individual data sets vary around their average i.e. the extent to which the function $g(x, D)$ varies from the specific selection of the data set.
  - Low variance: the approximation of $F(x)$ does not change much with another data record

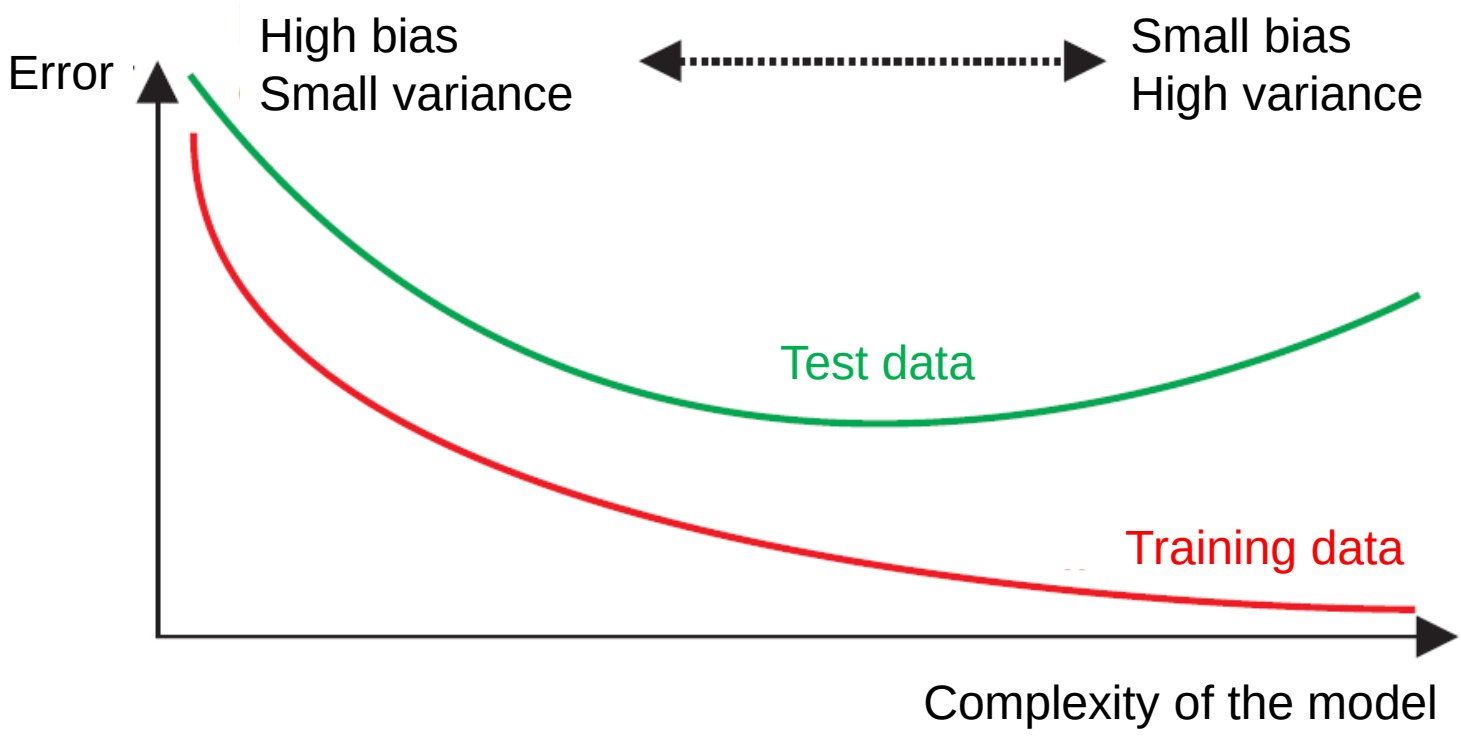
Bias and variance (6)

Bias variance dilemma:

- There is a trade-off between bias and variance: models with more free parameters usually have lower bias, but higher variance (and vice versa).
- Goal: low bias and low variance

Bias and variance (7)

Bias variance dilemma:



[Hastie, Tibshitani, Friedman, 2001]

Bias and variance (8)

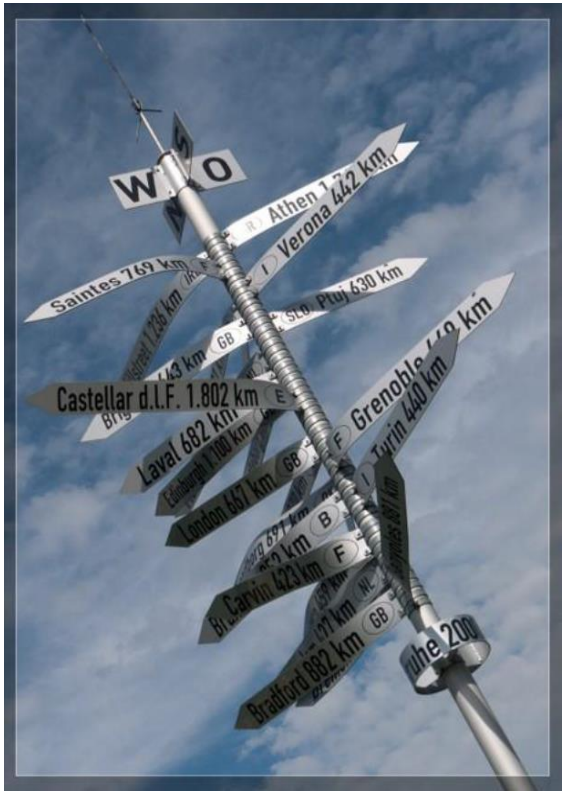
Bias variance dilemma:

- For classification tasks, the decomposition into bias and variance is more difficult.
- Usually a 0/1 error measure is used (sample is either correctly or incorrectly classified).
- It can be shown that the influence of variance in classification tasks is significantly higher than that of bias under the assumption of this error measure.

Agenda

Christian-Albrechts-Universität zu Kiel

- Evaluation
- Bias and variance
- **Model testing**
- Validation
- Conclusion
- Further readings



Model testing

Test of models

- Means in general test / evaluation of generalisation capability
- Methods:
  - Holdout Method
  - Cross validation
  - Jackknife (Leave-One-Out)
  - Bootstrap
  - and many more …
- Estimates for bias and variance: see e.g. [Duda, Hart, Stork, 2001].

Model testing (2)

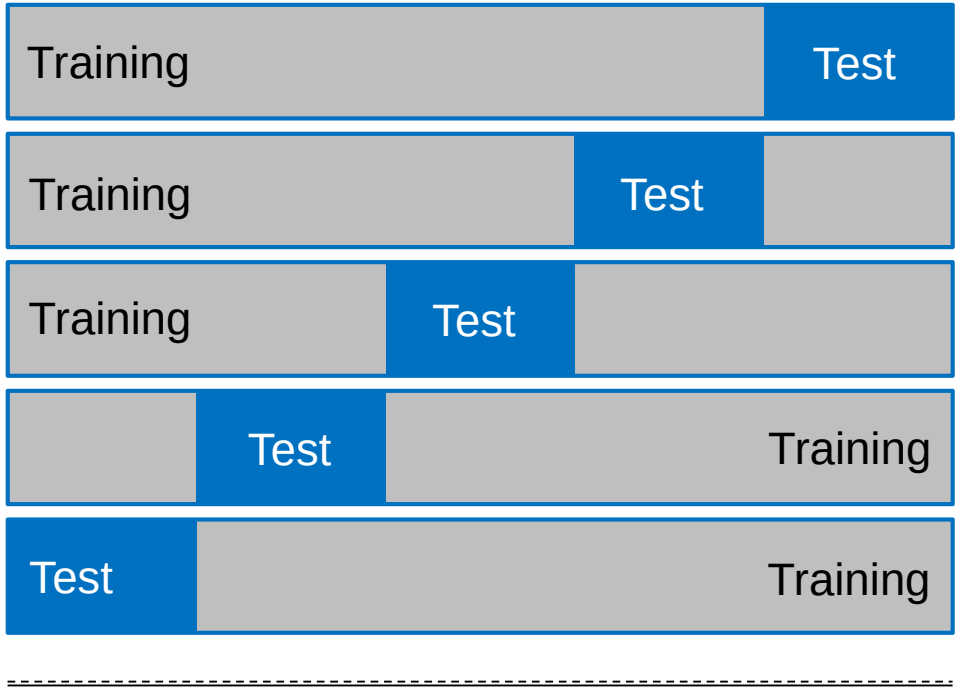
Holdout method:

- Part of available data for modelling (training), other part for testing (unknown data)
- Often: one third for testing, two thirds for training
- Problem: poor estimation of generalisation performance (strong dependence on data distribution)

Model testing (3)

Cross validation

- Repetition of the holdout method with different partial data sets:
- (here: 5-fold cross validation)
- Variant stratified cross validation: Each subset approximately equal distribution



Model testing (4)

Jackknife (Leave-One-Out):

- Corresponds to X-fold cross validation for X samples in the data set.
- In each pass of the cross validation a sample is used for testing.

Model testing (5)

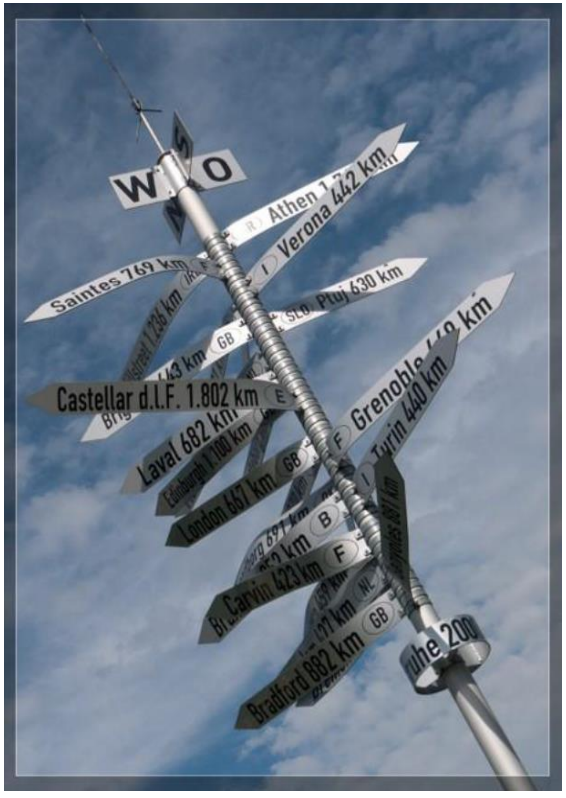
Bootstrap:

- From a data set of X patterns, X patterns are selected by drawing with replacement (i.e., putting the sample back in the set after drawing it).
- The remaining patterns are used for testing.
- The probability that a pattern will not be selected for training is about 0.368.

Agenda

Christian-Albrechts-Universität zu Kiel

- Evaluation
- Bias and variance
- Model testing
- Validation
- Conclusion
- Further readings



Validation

- Evaluation of a trained classifier with test samples (completely unknown beforehand and not considered in any step of the model creation)
- Many different valuation sizes possible
- Some well-known metrics are based on the so-called confusion matrix

Validation (2)

Confusion matrix

- Confusion matrix using the example of a classification problem with two classes, e.g. signature of one person (positive) and signature of another person or fake (negative)
- For each unknown test sample, the classifier predicts whether it belongs to the first class (positive examples) or to the second class (negative examples).
- For each prediction, a decision is made as to whether:
  A. the sample is indeed positive and has been correctly identified as such
  B. the sample is indeed negative and erroneously considered positive
  C. the sample is actually positive and incorrectly considered negative
  D. the sample is indeed negative and has been correctly recognised as such

Validation (3)

Christian-Albrechts-Universität zu Kiel

Confusion matrix

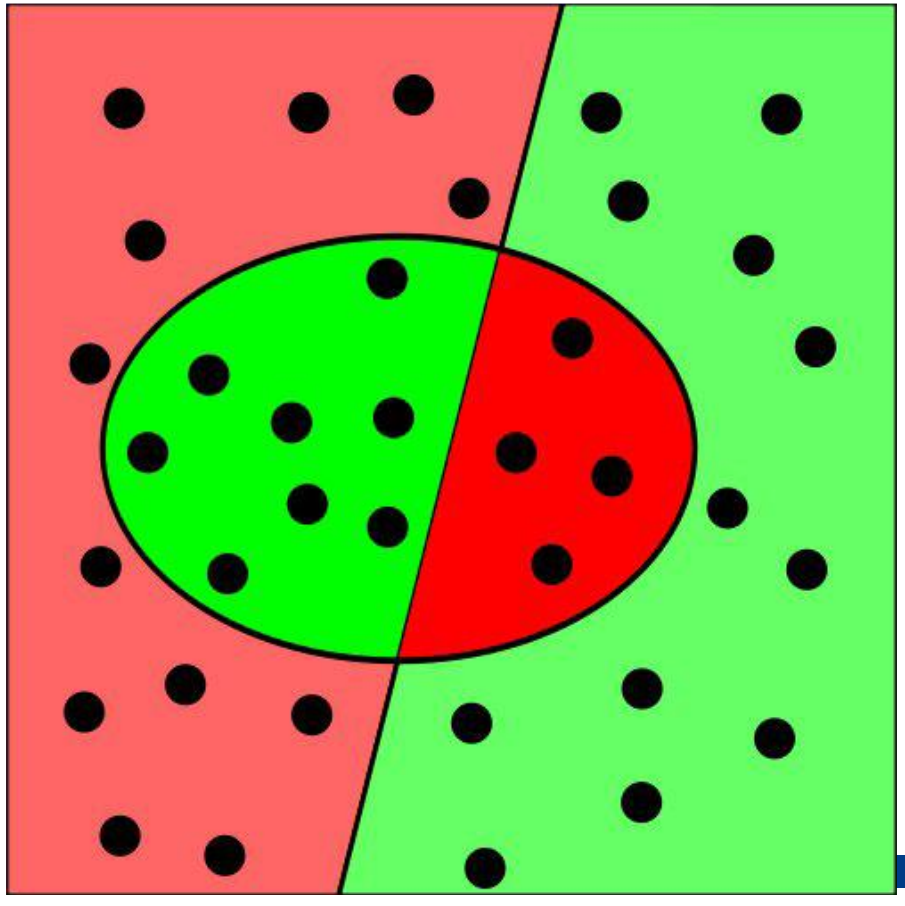
	pred. pos.	pred. neg.	
true pos.	A	C	A+C
true neg.	B	D	B+D
	A+B	C+D	A+B+C+D=N

Validation (4)

Christian-Albrechts-Universität zu Kiel

Possible classification errors:

- Classification of patients
  - Left side: Sick persons
  - Right side: Healthy persons
- Classifier: Inside the ellipse ill
- Right decision: Green
- Wrong decision: Red



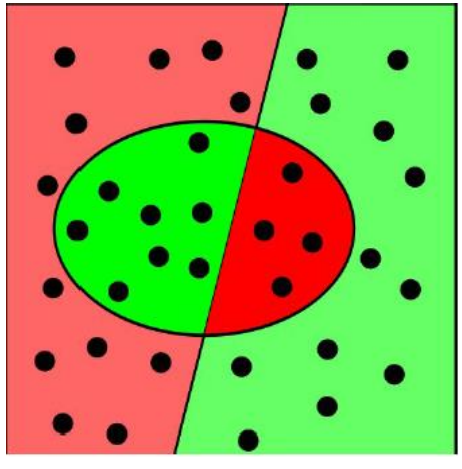
[wikipedia]

Validation (5)

Christian-Albrechts-Universität zu Kiel

Further evaluation measures for classification

	pred. pos.	pred. neg.	
true pos.	A	C	A+C
true neg.	B	D	B+D
	A+B	C+D	A+B+C+D=N



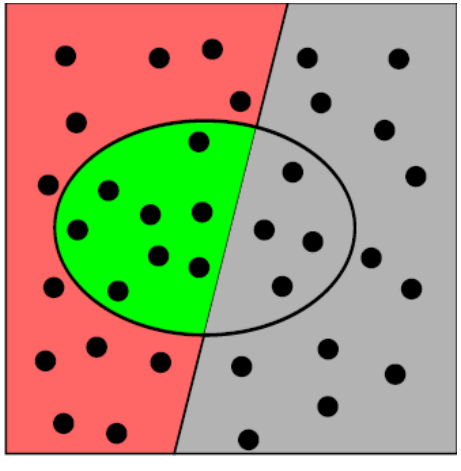
- Accuracy: $\frac{A+D}{N}$

Validation (6)

Further evaluation measures for classification

	pred. pos.	pred. neg.	
true pos.	A	C	A+C
true neg.	B	D	B+D
	A+B	C+D	A+B+C+D=N

- Sensitivity, true positive rate, recall: $\frac{A}{A+C}$
- False negative rate, miss rate: $\frac{C}{A+C}$

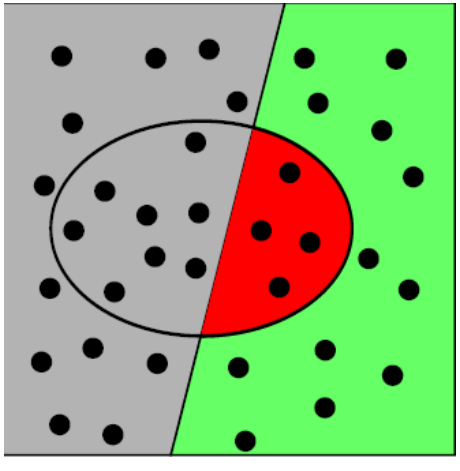


Validation (7)

Further evaluation measures for classification

	pred. pos.	pred. neg.	
true pos.	A	C	A+C
true neg.	B	D	B+D
	A+B	C+D	A+B+C+D=N

- Specificity, true negative rate: $\frac{D}{B+D}$
- False positive rate, fallout: $\frac{B}{B+D}$



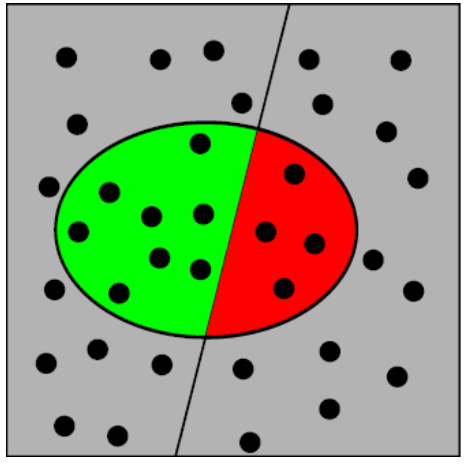
Validation (8)

Christian-Albrechts-Universität zu Kiel

Further evaluation measures for classification

	pred. pos.	pred. neg.	
true pos.	A	C	A+C
true neg.	B	D	B+D
	A+B	C+D	A+B+C+D=N

- Precision: $\frac{A}{A+B}$

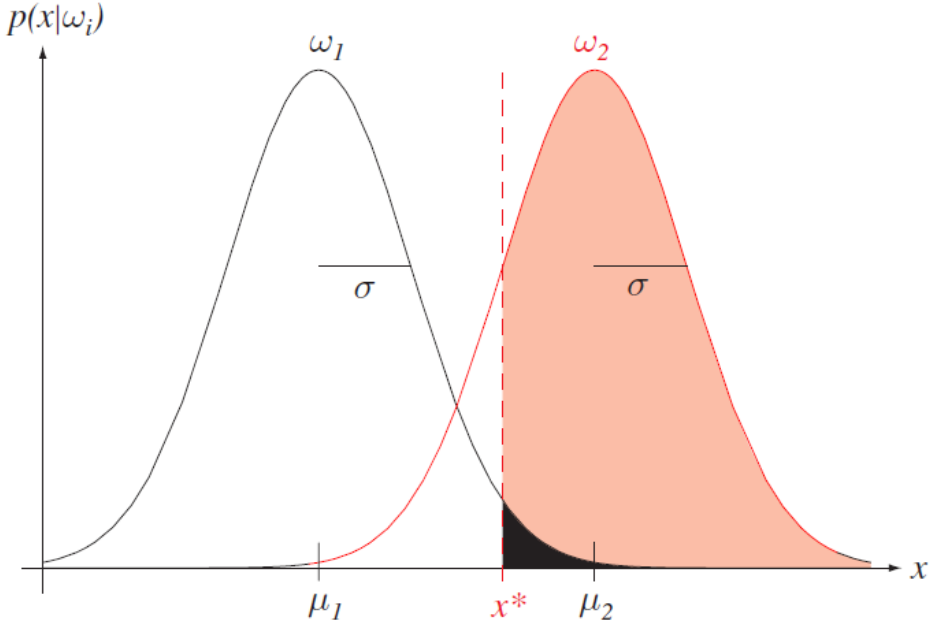


Validation (9)

Christian-Albrechts-Universität zu Kiel

ROC Curve

- Sensitivity and specificity are generally dependent on a threshold value (also discriminant threshold or operating point).
- Here an example with x^*:

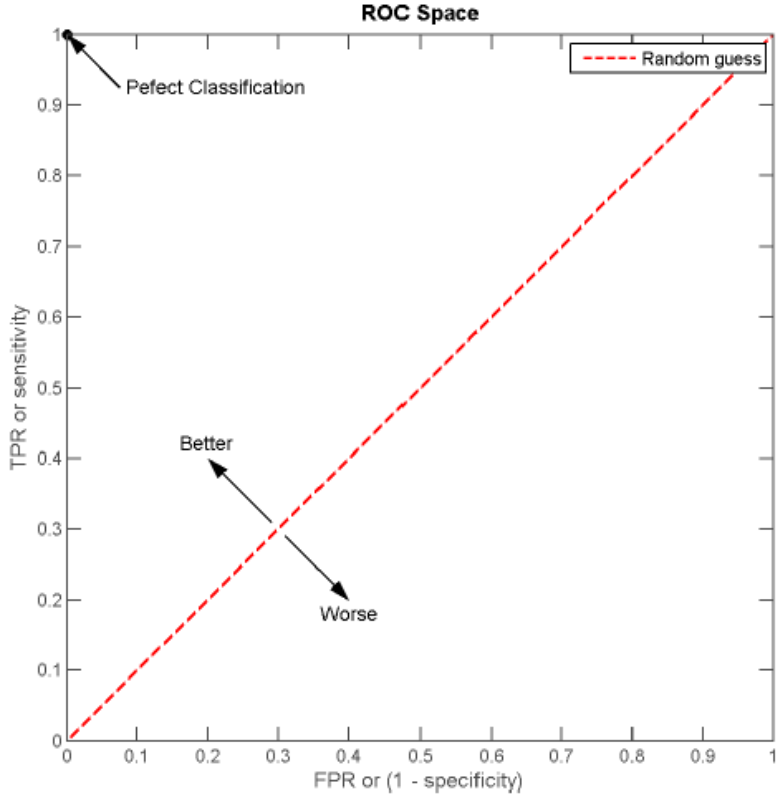


[wikipedia]

Validation (10)

Receiver operating characteristic (ROC) Curve

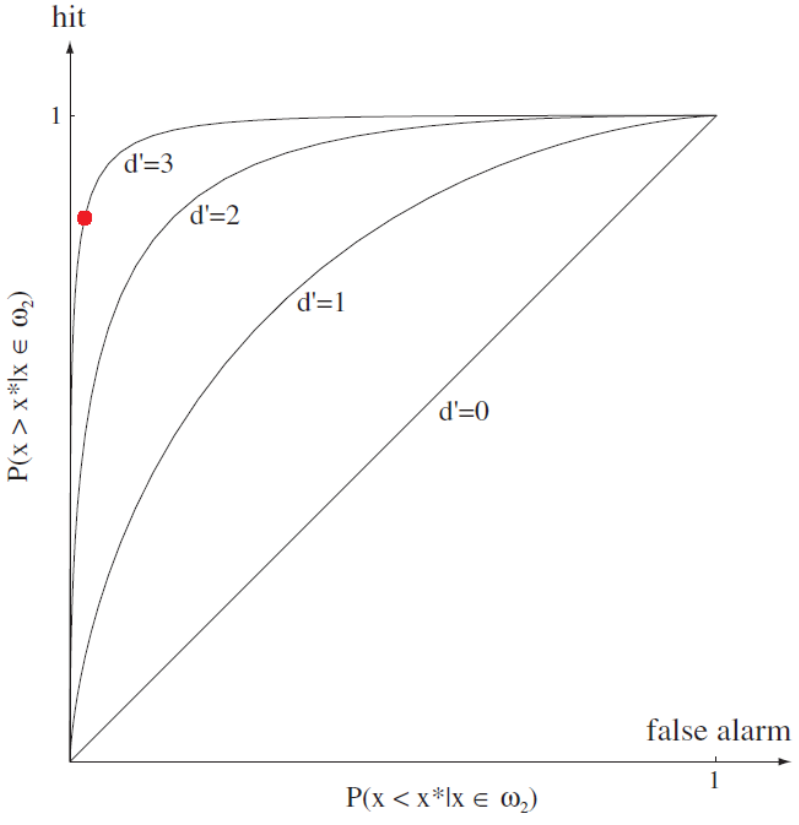
- Representation of sensitivity and specificity at different threshold values
- Sensitivity (true positive) plotted as ordinate and (1 – specificity), i.e. false positive, plotted as abscissa



Validation (11)

Christian-Albrechts-Universität zu Kiel

Receiver operating characteristic (ROC) Curve



Validation (12)

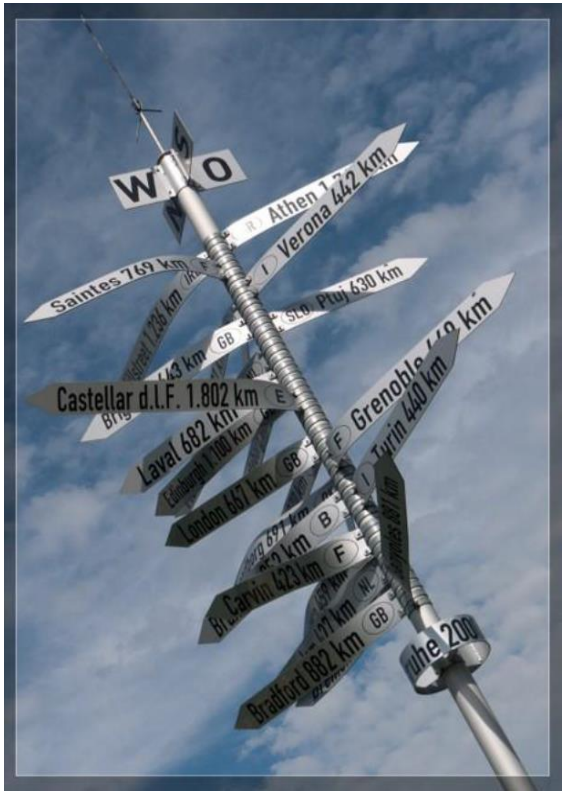
Receiver operating characteristic (ROC) Curve

- The evaluation of the ROC curve can be performed using the ROC AUC (area under curve)
- In principle, a diagonal shape is the worst possible result, i.e. an area of 0.5
  - An area > 0.5 (ideally close to 1) indicates a good Close classification result
  - An area < 0.5 can also be considered a correspondingly good result if the statements "positive" and "negative“ are interpreted in reverse
- The larger the area (i.e. a course strongly apart from the diagonals) shows a good quality of the classifier
  - Classifier comparison

Agenda

Christian-Albrechts-Universität zu Kiel

- Evaluation
- Bias and variance
- Model testing
- Validation
- Conclusion
- Further readings



Conclusion

- Random and pseudo-random influences must be taken into account when evaluating the generalisation performance:
- Errors in the determination or measurement of samples
- Description of the function to be modelled by a limited, possibly very small set of samples.
- Influences on search algorithms, e.g. stochastic optimisation methods, such as
  - Random initial values for iterative processes
  - Sequence in which samples are considered during parameter adaptation
  - ...

[wikipedia]

Conclusion

Christian-Albrechts-Universität zu Kiel

This chapter discussed:

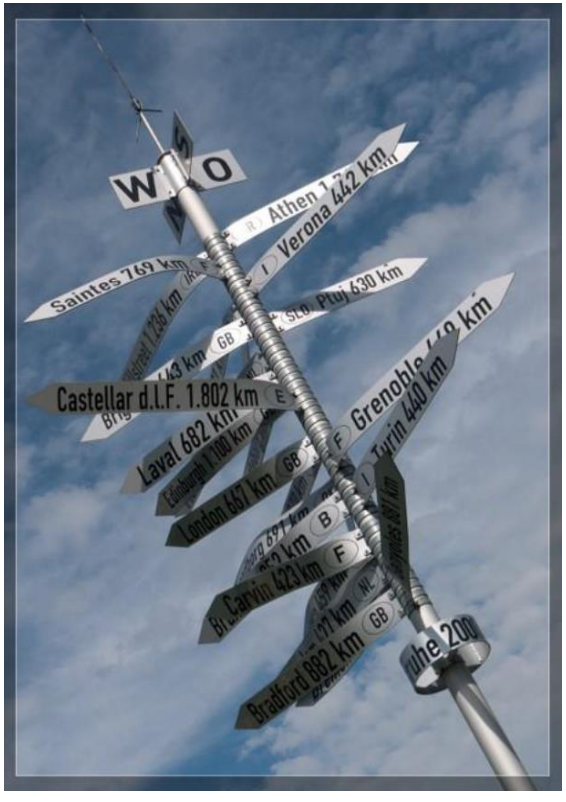
- Evaluation
- Bias and variance
- Model testing
- Validation
- Conclusion
- Further readings

Students should be able to:

- explain how evaluation is done for classifiers.
- introduce the bias vs. variance dilemma.
- compare the advantages and disadvantages of model testing approaches.
- define and apply different validation techniques.

Agenda

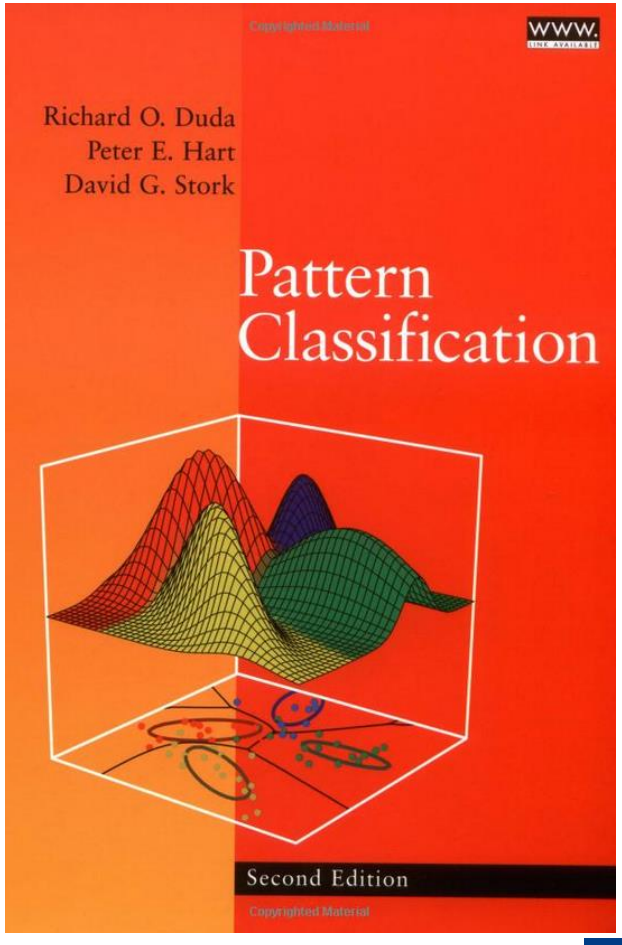
- Evaluation
- Bias and variance
- Model testing
- Validation
- Conclusion
- Further readings



Further readings

Basic readings:

- Duda, Richard O., Peter E. Hart, and David G. Stork.
- *Pattern classification*.
- John Wiley & Sons, 2012.
- ISBN: 978-0471056690



End

Christian-Albrechts-Universität zu Kiel

- Questions….?