Cube

Computational Systems Biology

Quantitative Methoden in der Molekularbiologie

11. Classification and performance; factorial designs

WS 2015/16

Exam question

Erklären Sie den Begriff „cycle threshold“ in der quantitativen PCR. Wie wird der cycle threshold bestimmt?

Explain the term „cycle threshold“ in quantitative PCR. How is the cycle threshold determined?

Exam question

a) Welche Methoden der Transkriptomanalyse können eingesetzt werden, um micro-RNAs zu identifizieren welche in Alzheimer-Patienten, verglichen mit Gesunden, differenziell exprimiert sind?

Which methods for transcriptome analysis can be used to identify micro-RNAs that are differentially expressed in Alzheimer patients, compared to healthy subjects?

b) Welche der Methoden würden Sie für eine neu zu planende Studie mit 1250 Probanden vorschlagen? Warum?

Which method would you recommend for a new study with 1250 individuals. Why?

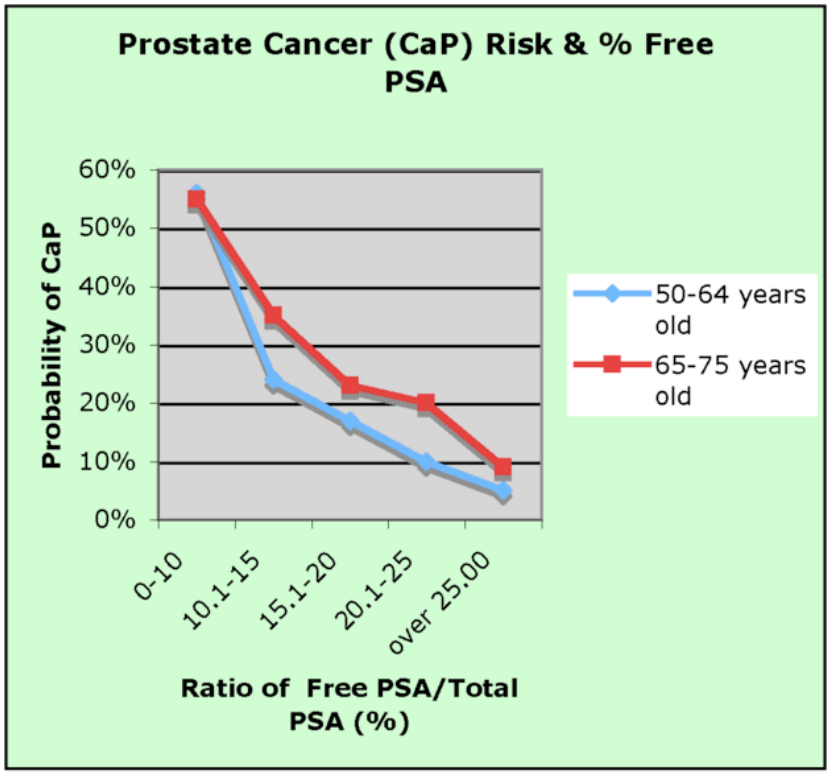
Outline

1. Classification and performance measures

2. Factorial designs

Classification

- Identifying sub-populations based on observations
- Needs correlation between observation and subpopulation features (no causation)
- Example: Biomarkers



Prostata cancer tests using Prostate-Specific Antigen (PSA)

Classification test

Based on:

- Training data (used to determine classifier function): sub-populations known

- Test data (independent from training data, but comprehensive): sub-populations to be predicted and evaluated

Confusion matrix

		Condition (as determined by "Gold standard")	
		Positive	***Negative***
Test outcome	***Positive***	**True Positive**	**False Positive** (Type I error)
	Negative	**False Negative** (Type II error)	**True Negative**

Classification performance

		Condition (as determined by "Gold standard")		
		Positive	***Negative***	
Test outcome	***Positive***	**True Positive**	**False Positive** (Type I error)	→ Positive predictive value $= \frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test outcome Positive}}$
	Negative	**False Negative** (Type II error)	**True Negative**	→ Negative predictive value $= \frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test outcome Negative}}$
		↓ Sensitivity $= \frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	↓ Specificity $= \frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

Classification performance example

		Patients with bowel cancer (as confirmed on endoscopy)		
		Positive	***Negative***	
Fecal occult blood screen test outcome	***Positive***	**True Positive (TP) = 20**	**False Positive (FP) = 180**	→ Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = 20 / 200 **= 10%**
	Negative	**False Negative (FN) = 10**	**True Negative (TN) = 1820**	→ Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) = 1820 / 1830 **≈ 99.5%**
		↓ Sensitivity = TP / (TP + FN) = 20 / (20 + 10) = 20 / 30 **≈ 66.67%**	↓ Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = 1820 / 2000 **= 91%**	

Instead of conclusion: how not to lie with classification statistics

- Test data:
  - Remove redundancies
  - Remove elements not independent from training data
  - Sample all degrees of freedom comprehensively
- Assessing performance:
  - Sensitivity and selectivity may be misleading for unbalanced data -> Provide also PPV/NPV
  - Calculate costs of false classifications -> classification still useful??

Comparison of classifiers

- Balance between sensitivity and selectivity can be easily changed by changing thresholds
- Direct comparison of sensitivity and selectivity values is therefore impossible
- Cumulative performance of classifiers along a wide range of thresholds is compared insted

Receiver Operator Characteristics (ROC)

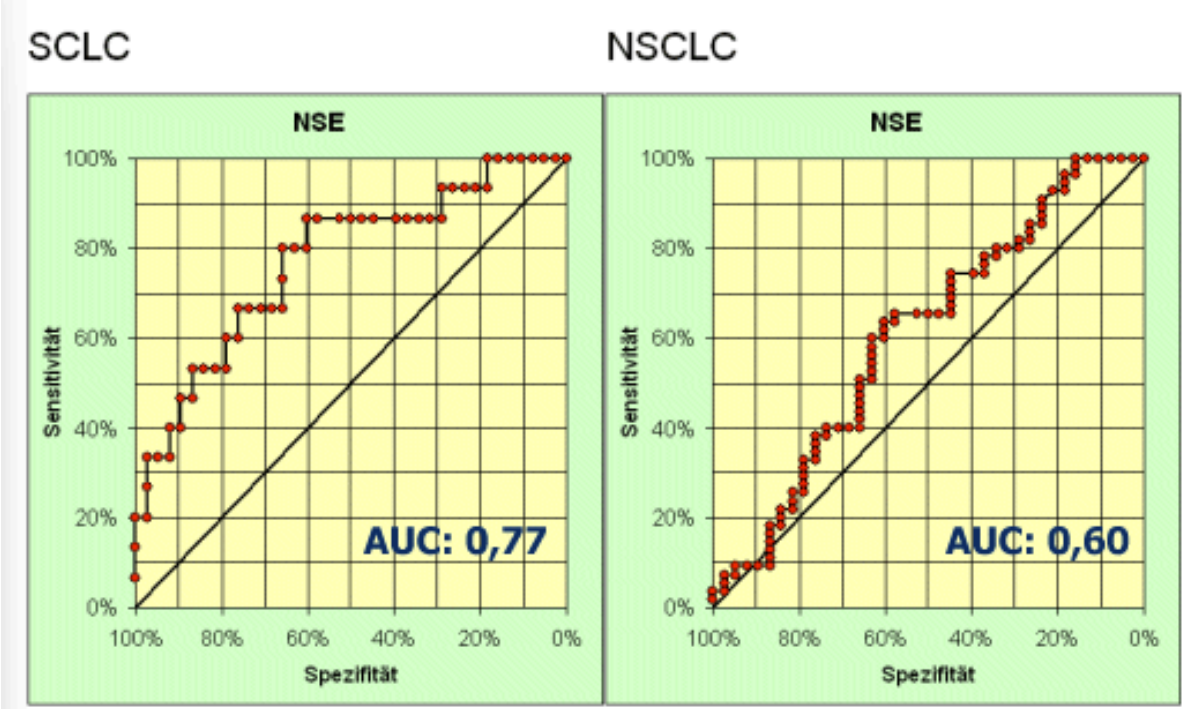


Abb. 1: ROC-Kurven für die NSE-Werte (NSE: Tumormarker Neuronen-spezifische Enolase) von Patienten mit Bronchialkarzinom bzw. benignen Lungenerkrankungen. Links: Patienten mit kleinzelligem BCa (SCLC), rechts: Patienten mit nichtkleinzelligem BCa (NSCLC). Angegeben sind die ROC-Kurve (rote Kreise), sowie die Diagonale (schwarze Linie).

AUC: Area under the curve

Online calculation

Input Data: (paste or enter)

```
1  0.378
1  1.250
1  0.225
1  1.373
1 -0.869
1  0.817
1  1.541
1  1.123
1  0.907
1  0.210
1  1.472
1 -0.099
1  2.951
1  1.254
1  0.789
1  0.882
1  0.554
1  0.560
1  1.273
1 -0.207
```

Program Output: (may be copied and pasted into other programs)

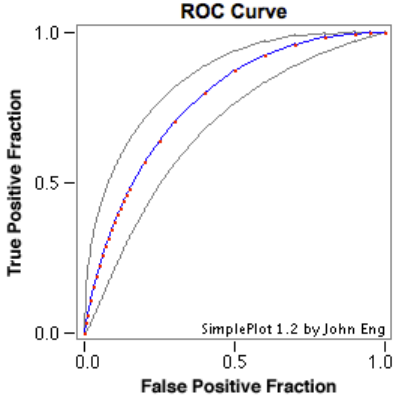
```
JLABROC4:
Maximum likelihood estimation of a binormal ROC curve from
continuously distributed test results.

Java translation by John Eng, M.D.
The Russell H. Morgan Department of Radiology and
  Radiological Science
Johns Hopkins University, Baltimore, Maryland, USA
Version 1.0.1, Aug 2006

Original Fortran program LABROC4 by Charles Metz & colleagues
Department of Radiology, University of Chicago
October 1997
------------------------------------------------------------

INPUT DATA
  Scores from the 68 actually negative cases:
   -0.0370     0.2880    -1.6490    -0.0740     0.8330
   -2.0190     0.9760     0.5610     0.4940    -1.6990
    0.9810     0.8080     0.0440    -1.1070     2.1620
    0.0900     0.3040     0.1530    -0.2340    -0.6810
```

ROC Curve



Summary Statistics:

```
Total Cases:       148
Positive Cases:    80
Negative Cases:    68

Fitted ROC Area:   0.773
```

Points for Plotting: (copy & paste to Excel)

FPF	TPF	Lower	Upper
0.0000	0.0000	0.0000	0.0000
0.0050	0.0340	0.0045	0.1503
0.0100	0.0621	0.0121	0.2061
0.0200	0.1105	0.0310	0.2803
0.0300	0.1527	0.0525	0.3341
0.0400	0.1908	0.0752	0.3776
0.0500	0.2257	0.0984	0.4146
0.0600	0.2581	0.1219	0.4471
0.0700	0.2883	0.1452	0.4762
0.0800	0.3167	0.1684	0.5027
0.0900	0.3435	0.1912	0.5270
0.1000	0.3690	0.2137	0.5495
0.1100	0.3931	0.2358	0.5704
0.1200	0.4161	0.2574	0.5901
0.1300	0.4381	0.2785	0.6086

http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html

Outline

1. Classification and performance measures

2. Factorial designs

Rationale and concept

- Problem: Often multiple hypotheses are studied in the same biological context. Investigating these independently is costly.

- Solution: Factorial designs investigating different hypotheses in same experiment.

- Advantage: hidden replication and ability to study interactions between factors

- Statistical test method: Multi-Way ANOVA (online tool e.g. http://vassarstats.net/vsanova.html)

Nomenclature

- Factor: something which might have an effect
- Level: state of a factor (quantitative)
- Treatment: particular combination of one level of one factor plus one level of another factor plus…
- Replicates: number of individuals that experience the same treatment

Example: 2x2 design

Factor 2: antibiotics	**Factor 1: diet**	
	Level 1: high-fat (H)	Level 2: low fat (L)
Level 1: Ciprofloxacin (C)	Treatment CH	Treatment CL
Level 2: Rifampicin (R)	Treatment RH	Treatment RL

Example: 2x2 design

Factor 2: antibiotics	**Factor 1: diet**	
	Level 1: high-fat (H)	Level 2: low fat (L)
Level 1: Ciprofloxacin (C)	5 biological replicates	4 biological replicates
Level 2: Rifampicin (R)	10 biological replicates	12 biological replicates

Determine the number of replications for each treatment and for each factor.