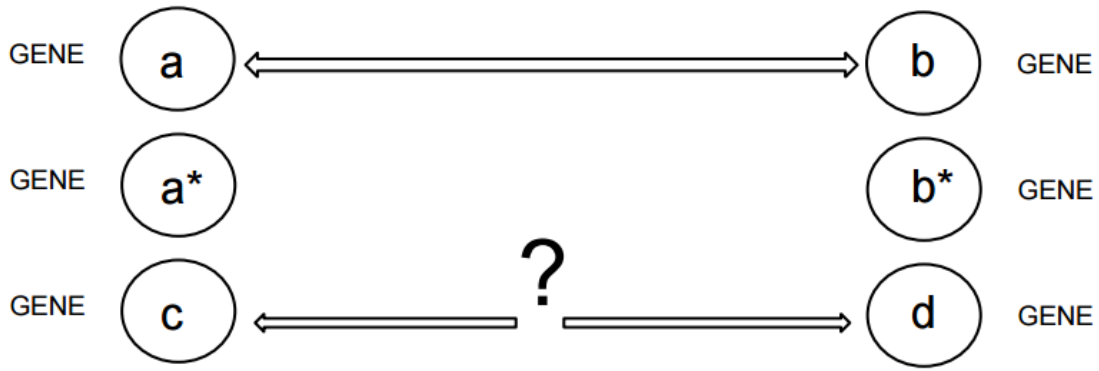
Questions?

How does the similarity-based classification (Kernel approach) work?

Link Prediction - Similarity-based classificiation (Kernel approaches)



For some pairs of genes (a,b) we know that they interact.
For other pairs (a*,b*) we know that they do not interact.
For new pairs (c,d) we do not know whether they interact, and would like to predict whether they interact.

General idea in similarity-based classification approaches:
Measure whether (c,d) is more similar to known interactions (a,b) than to known non-interactions (a*,b*) via a similarity measure, e.g. a kernel function.

in our case, (c,d) might be probable to have an interaction based on our similarity based classification

kernel is in this context something like a translation

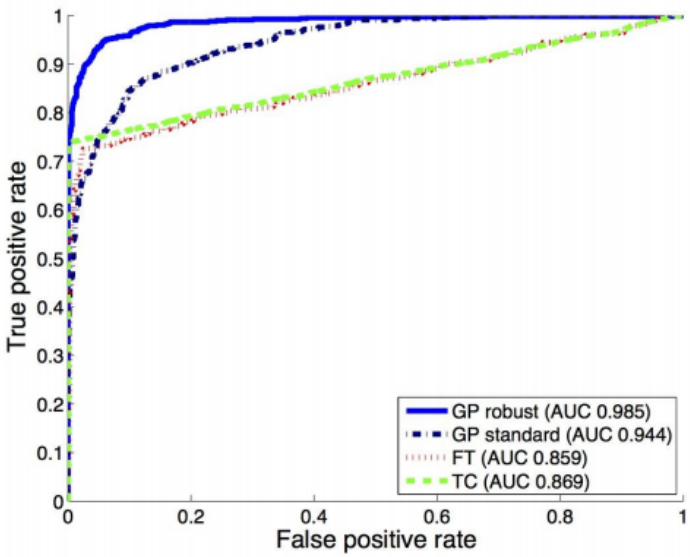
kernels are functions (see LA)

How do you calculate AUC?
How high is AUC if no errors occur?

computational integration: simposns rule (summation)
simply, since we have discrete datapoints, so it's nice

if surface is 1
then perfect

Link Prediction - Evaluation Criteria - ROC



Receiver **O**perating **C**haracteristic **(ROC)** curve plots

False Positive Rate vs. True Positive Rate.

AUC = **A**rea **u**nder the ROC **c**urve

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Systems Biology FS17
Link Prediction: Karsten Borgwardt

34

Could you explain me the disadvantage of the Lasso-model?

important

biologically we find correlations
with lasso we only find 1 correlation
lasso wants most weights beta to be 0
we only get 1 feature with lasso
it is computationally advantegous, I think

Multivariate Feature Selection: Lasso Model

Lasso Model (Tibshirani, 1996)

$$\arg\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda_1 ||\beta||_1$$

Idea: Reward solutions (β) in which few entries of β are non-zero.

Concept: This is achieved by minimizing the L1-norm of β:

$$||\beta||_1 = \sum_{i=1}^{d} |\beta_i|$$

Disadvantage: If there are groups of correlated features, the Lasso often picks just one feature from a group.

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

Systems Biology FS17
Feature Selection: Karsten Borgwardt

21

If you once separate test-/training data and perform [feature selection] multiple times, is this also a problem over over-fitting?

principally yes. vlaidation from training set are put into feature selection if process is repeated often

Common Pitfalls: Overfitting

if questions to over-fitting occur, it's most often yes and I have to explain why in my own words

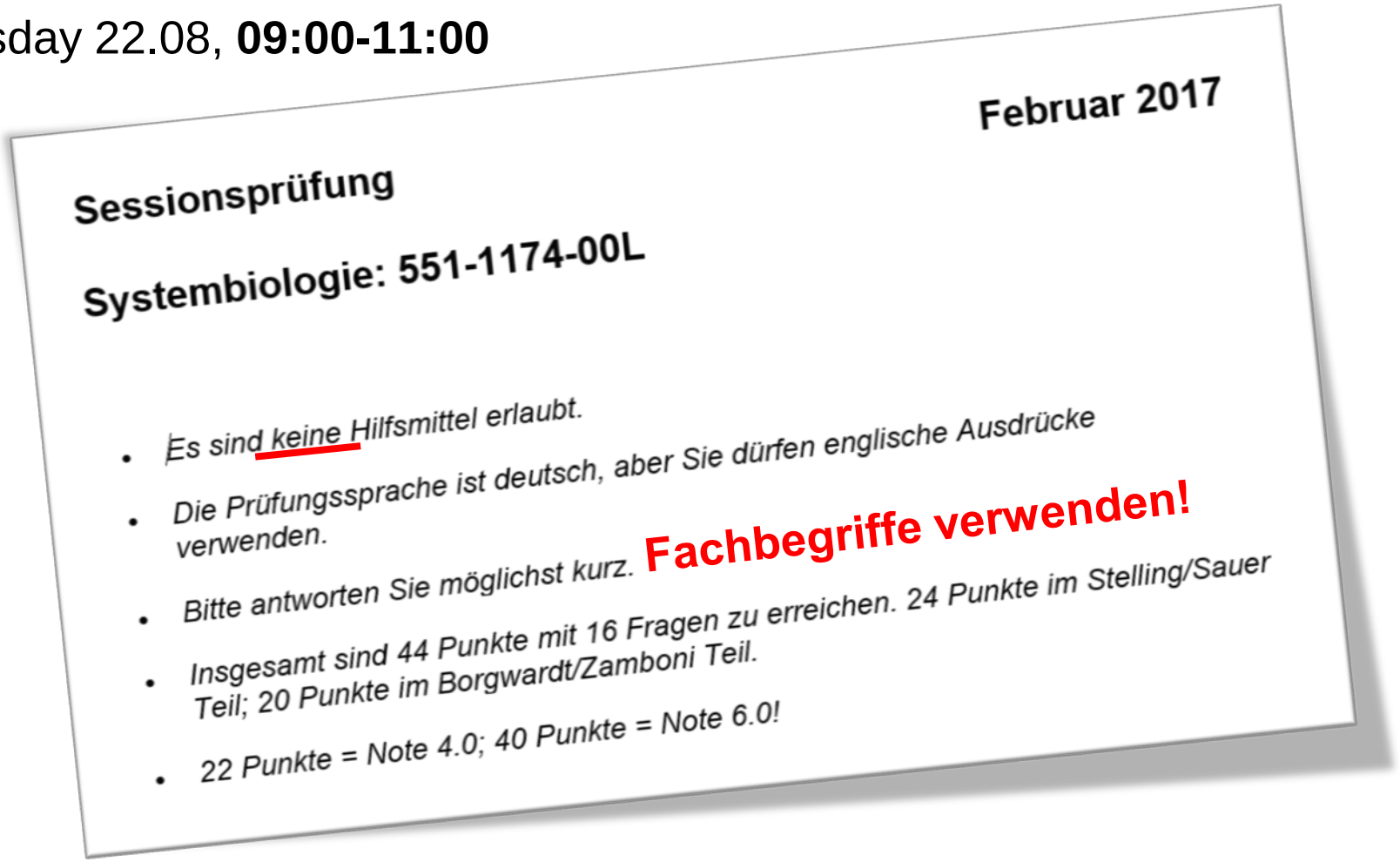
How to avoid the selection bias?

- Select the features on the training dataset only
- Use these features for prediction on the separate test dataset

Written examination

General information

Tuesday 22.08, **09:00-11:00**



Februar 2017

Sessionsprüfung

Systembiologie: 551-1174-00L

- *Es sind keine Hilfsmittel erlaubt.*
- *Die Prüfungssprache ist deutsch, aber Sie dürfen englische Ausdrücke verwenden.*
- *Bitte antworten Sie möglichst kurz.* **Fachbegriffe verwenden!**
- *Insgesamt sind 44 Punkte mit 16 Fragen zu erreichen. 24 Punkte im Stelling/Sauer Teil; 20 Punkte im Borgwardt/Zamboni Teil.*
- *22 Punkte = Note 4.0; 40 Punkte = Note 6.0!*

Officially: 2 parts > start from what you prefer.

Exam content on Borgwardt/Zamboni part

Expect 2 kind of questions:

I can answer all these kind of questions for the methods that occurred in the exercises, I think, for the question just below

- **Starting from a concept or method**, explain…
  - what the concept/method is
  - how it works (including variants presented in the lecture)
  - what typical applications are
  - what the pitfalls are

- **Starting from a bio problem**, explain how you would approach it to get the information you want

- **Starting from a picture** (PCA, clustering, …), interpret the results

Key concepts/methods to know

those concepts are central for a successful exam

- Univariate feature selection
- Multivariate feature selection
- Clustering
- Enrichment analysis
- Dimension reduction
- Principal component analysis
- Link prediction
- Contingency table, error types, ROC curve, precision-recall

Exemplary Data Mining questions

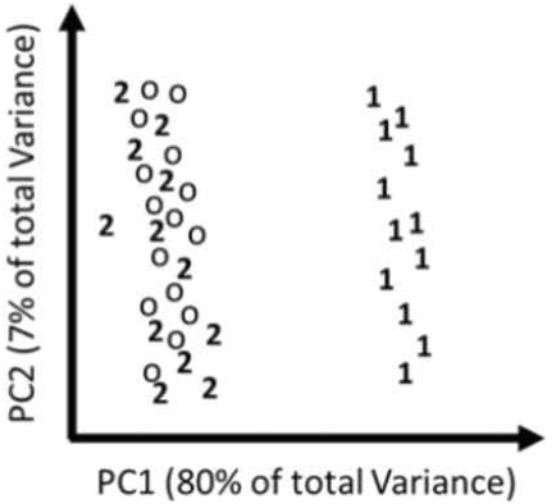
- Out of 100 patients with lung cancer, 30 were found to be resistant to a given drug. Cancer samples from each patients were profiled at proteomics level. How would you use data mining to find the molecular causes of resistance?

  reponsive <-> non-responsive

- About 100 putative transcription factors exist in yeast. How would you experimentally/computationally infer the topology of transcriptional network?

- You are trying to identify genes that convey virulence in a pathogenic *E. coli* strain. You performed a transcriptome analysis of several pathogenic and non-pathogenic strains. You performed a principal component analysis, but the two groups don't cluster when you plot the scores of the first two components. What do you try next?

13) Ein Spital hat eine Proteom-Analyse von Blutproben für drei im voraus bekannten Patientengruppen durchgeführt: Krankheit 1, Krankheit 2, und gesunden Kontrollen (o). Etwa 1000 Proteine wurden in jeder Probe gemessen. Eine Hauptkomponenten Analyse (PCA) liefert dieses Bild:



a) Wie interpretieren sie diese Ergebnisse? **(1 P)**

1: very different from controls, 2 is not so different from control

b) Wie würden Sie vorgehen, um Proteom-Features zu identifieren die zur Erkennung von Krankheit 1 dienen? **(1 P)**

feature selection, maybe PCA etc etc

Method questions…

Ein Forscher versucht, mittels Genexpressionsdaten einen Krankheitsphänotypen vorherzusagen. Er trainiert ein Modell auf den Trainingsdaten und sagt dann auf den Testdaten voraus. Wenn die Vorhersagegenauigkeit nicht zufriedenstellend ist, trainiert er ein neues Modell auf den Trainingsdaten.

Dieses Vorgehen wiederholt er Dutzende Male, bis ein erfolgreiches Modell gefunden wurde.

Auf die Rückfrage, ob er nicht Overfitting betreibe, antwortet der Forscher, er separiere stets Trainings- und Testdatensatz gründlich beim Trainieren und Testen und vermeide damit Overfitting.

yes, because testdata are put in into traningsdata

a) Begründen Sie, warum hier in Wahrheit doch ein Problem mit Overfitting vorliegt. (2 P)

b) Begründen Sie, warum ebenfalls ein Problem mit multiplem Hypothesentesten vorliegt. (2 P)