NBE-4070 : Basics of Biomedical Data Analysis

Stéphane Deny
Prof. in Neuroscience and Biomedial Engineering and Computer Science
Aalto University

Lecture 8: Preparation for oral exam / Guest lecture

Outline of the course

1. Mean, Standard Deviation, Standard Error, Confidence Intervals, T-test
2. Fourier Transform, Wavelet Transforms, Spectrograms, High-pass, Low-pass filters
3. Covariance and Principal Component Analysis (PCA)
4. Clustering Methods
5. Pearson Correlation, PCA and SVD
6. Linear Regression / Logistic Regression
7. Non-linear Methods: k-NN, random forest, t-SNE, deep nets
8. Oral exam preparation / Invited lecture from the biomedical industry

Learning Objectives

After completing this course, you:

- understand the fundamental linear and non-linear methods used for biomedical data analysis, and know how to explain them.

- Given a question and biomedical measurement, you know how to select and apply the suitable data analysis methods for this problem.

- have all the keys to avoid overinterpreting or misinterpreting the results of different analyses.

Oral exam: modalities (30% of final grade)

- Pick a subject from a box: you are given a dataset description and a problem (5 minutes of preparation alone, and then 10 minutes of interview).

- Outline of the interview:
  - “What analysis would you perform on this dataset to address this problem?” (1 point)
  - “Explain the analysis proposed to someone who wouldn’t know what it is.” (1 point)
  - “Your analysis does not provide a satisfying result for [insert reason]. What solution can you think of to mitigate this problem?” (1 point)
  - “Here is the result of your analysis. What is your interpretation of this result? What are the potential risks of over- and mis-interpretation?” (1 point)

Supervised & unsupervised learning: definitions

- Supervised learning methods use labeled datasets to train algorithms to classify data or predict outcomes.

- Unsupervised learning methods discover hidden patterns or data clusters in unlabeled datasets.

Signal processing & machine learning: definitions

- Signal processing methods transform data in a way that allows us to see things in it that are not possible via direct observation. They do not require training on a sample data.

- Machine learning methods build a model based on sample data, known as training data, in order to make predictions or decisions on a testing set.

Linear & non-linear methods: definitions

- Linear methods rely on linear functions of the input features. They are not able to exploit complex non-linear structure present in the data.

- Non-linear methods rely on non-linear functions of the input features. They can uncover sophisticated patterns present in the data. They are prone to over-fitting because they typically have more parameters than linear methods.

Classification of different methods for data analysis

Fourier Transform, Low-pass / High-pass filtering, Wavelet transform, Spectrum, PCA / SVD, Linear/Logistic regression, t-SNE, K-means clustering, Hierarchical clustering, K-nearest neighbor, Random Forest, Deep networks

	Signal processing (no learning)	*Unsupervised learning method*	*Supervised learning method*
Linear method			
Non-Linear method			

Classification of different methods for data analysis

	Signal processing (no learning)	*Unsupervised learning method*	*Supervised learning method*
Linear method	Fourier Transform Low-pass / High-pass filtering Wavelet transform	PCA / SVD	Linear regression Logistic regression
Non-Linear method	Spectrum	K-means clustering Hierarchical clustering t-SNE	K-nearest neighbor Random Forest Deep networks

+ Statistical tests (confidence intervals, t-tests)

Problem 4

You work in a biomedical company which goal is to predict the onset of diabetes from blood samples. You are given the results of laboratory analyses of blood samples for a cohort of 1000 patients. The analysis measures 100 different biomarkers from the blood sample. You are also given the level of a biomarker that is indicative of diabetes measured 3 years after the initial blood sample was performed. This biomarker is a real number related to the severity of diabetes. You would like to build a tool which predicts the level of this biomarker from the blood samples.

What is the simplest method you could try to build a predictor of diabetes from a blood sample?

Explain how this method works in simple terms. You can use the whiteboard to help you.

Problem 4

You work in a biomedical company which goal is to predict the onset of diabetes from blood samples. You are given the results of laboratory analyses of blood samples for a cohort of 1000 patients. The analysis measures 100 different biomarkers from the blood sample. You are also given the level of a biomarker that is indicative of diabetes measured 3 years after the initial blood sample was performed. This biomarker is a real number related to the severity of diabetes. You would like to build a tool which predicts the level of this biomarker from the blood samples.

What is the simplest method you could try to build a predictor of diabetes from a blood sample?

Explain how this method works in simple terms. You can use the whiteboard to help you.

Follow-up questions:

Your method predicts diabetes accurately on the training set, but not on the testing set. How could you try to improve on this result?

After regularization, you find that now performance is weak on both training and testing sets. What might the problem? And what could you do to fix it?

After the break

Guest lecture:

Dr. Karita Salo – Biostatistician at Nordic Healthcare Group

Personalized medicine and drug development

Next Lecture(s): To be defined