UEF SUMMER SCHOOL 2017
Machine Learning Applied to Bioinformatics and Speech Technology

Your task **is to write a short summary of what you learned during the lecture with answers to** each of the questions below, save your answer sheet as PDF and send it to course TAs.

General machine learning sessions (mandatory to all course students):

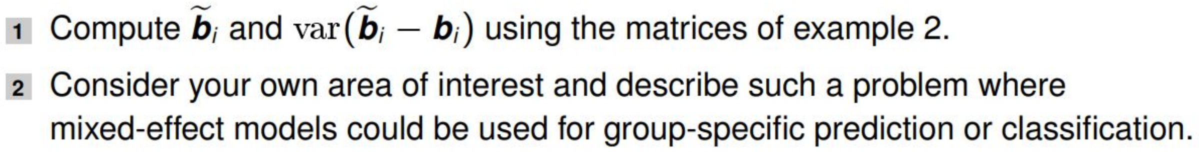
ML1. Introduction to biomedical data

Is biological data big data? In what ways?
Suggested further reading:
http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

ML2. Introduction to speech data

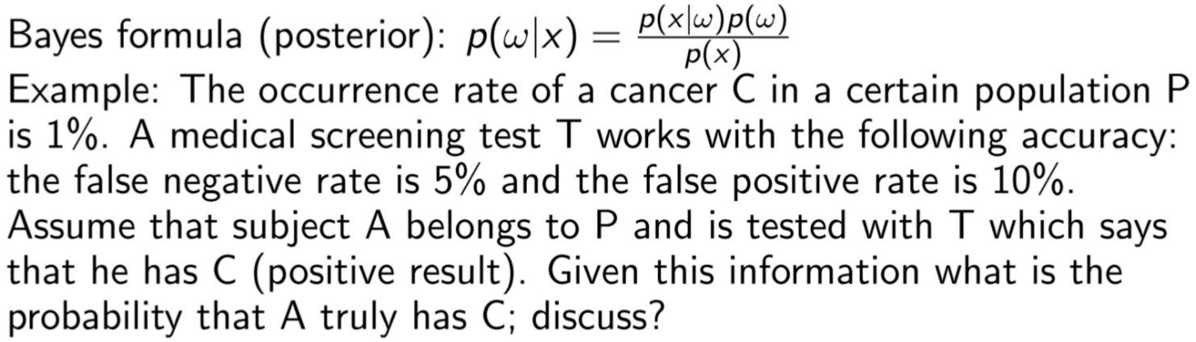
Find out what is the working principle of mel-frequency cepstral coefficient (MFCC) feature extraction and what are its main applications.

ML3. Statistics



1. Compute $\widetilde{\boldsymbol{b}}_i$ and $\mathrm{var}(\widetilde{\boldsymbol{b}}_i - \boldsymbol{b}_i)$ using the matrices of example 2.
2. Consider your own area of interest and describe such a problem where mixed-effect models could be used for group-specific prediction or classification.

ML4. Basics of Machine Learning



Bayes formula (posterior): $p(\omega|x) = \frac{p(x|\omega)p(\omega)}{p(x)}$
Example: The occurrence rate of a cancer C in a certain population P is 1%. A medical screening test T works with the following accuracy: the false negative rate is 5% and the false positive rate is 10%. Assume that subject A belongs to P and is tested with T which says that he has C (positive result). Given this information what is the probability that A truly has C; discuss?

ML5. Deep Neural Networks

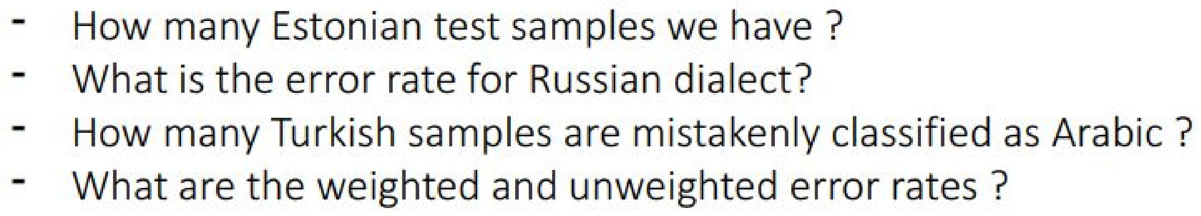
Neural networks (especially deep networks) can overfit very easily. What happens in this phenomenon and what you can do to avoid it?

It is always important to think about the problem at hand carefully, since there are different ways to model to data to and get a solution. What is the difference between generative and discriminative models? And when (and why) you would select to use either of them?

Excellent and advanced look on the state-of-the-art neural network (and at the same time machine learning in general) results are available in this online book (published 2016):
http://www.deeplearningbook.org/

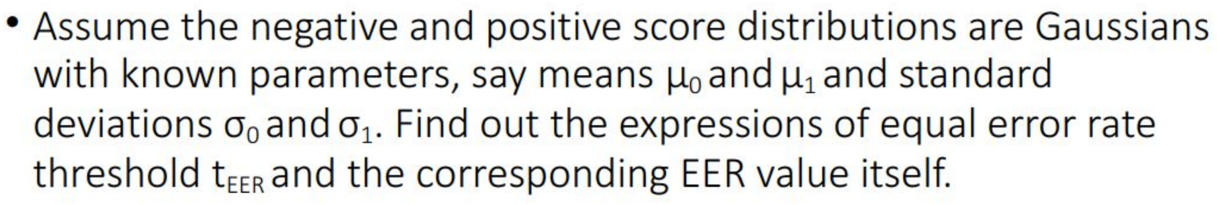
ML6. Sequence Modeling
Describe differences and similarities between stochastic sequence models and neural sequence models.

ML7. Evaluation in machine learning



- How many Estonian test samples we have ?
- What is the error rate for Russian dialect?
- How many Turkish samples are mistakenly classified as Arabic ?
- What are the weighted and unweighted error rates ?

Additional bonus homework:



- Assume the negative and positive score distributions are Gaussians with known parameters, say means μ_0 and μ_1 and standard deviations σ_0 and σ_1. Find out the expressions of equal error rate threshold t_{EER} and the corresponding EER value itself.

BIO track sessions (mandatory to BIO track students):

BIO1. What is the function of DNA, RNA and protein molecules in cells? (LS1)

You may find the tutorial and links useful to learn about the central dogma:
https://www.nobelprize.org/educational/medicine/dna/index.html
This fundamental knowledge was worth several Nobel prizes!

If this is all familiar to you, write a short summary of current high-throughput measurement technologies to measure these three key molecular types.

BIO2. How would we know if the DNA letter change affects synthesis or regulation by key molecules and could therefore also be linked with disease? Can a machine predict which changes in DNA have a functional consequence?
The following articles are useful further reading on the topic

http://www.ncbi.nlm.nih.gov/pubmed/21526222
http://www.nature.com/encode/threads/machine-learning-approaches-to-genomics
http://deepsea.princeton.edu/job/analysis/create/

BIO3. What is gene expression profiling? You may also wish to explore how many (gene) expression profiling experiments are publicly available from one of the main data repositories http://www.ncbi.nlm.nih.gov/geo/summary/

BIO4. Can genome-wide gene expression be monitored from single cells?
Here is an overview of best practices of RNA-sequencing that at the end covers this topic
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8
If an RNA molecule is present in low amounts, would you expect to detect each copy in your experiment? The technology allows capture rates of around 10–50 %. As you may have correctly concluded, this poses a challenge, and calls for analysis methods that can take stochasticity into account.

BIO5. Image part

Doing 1 OR 2 is enough:

1) Analyze and reflect fundamental differences in the design of machine learning systems for a) imaging based prediction models and b) image segmentation?

2) Analyze the variance of counting based classification error estimates: Generate samples for two classes from 1-dimensional Gaussian distribution, N(mu1,sigma) and N(mu2,sigma); use n = 10,20,50, or 100 samples per class. Train a classifier using the nearest centroid method https://en.wikipedia.org/wiki/Nearest_centroid_classifier . Estimate classification errors using Leave-one-out for k = 1000 different samples, sampled from the same distribution. What do you observe? How does the variance of the error estimate behave as the function of sigma (when mu1 and mu2 are fixed; you can use ,mu1 = 0 and mu2 = 1) and sigma = 0.1, 0.2, 0.3, 0.4, 0.5, 1. Discuss.

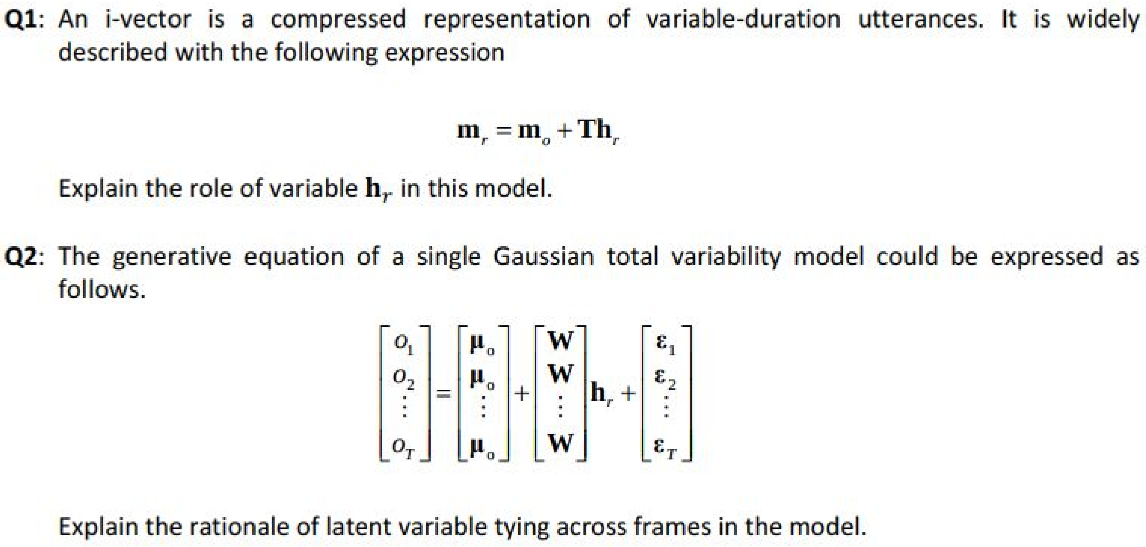
SPEECH track sessions (mandatory to SPEECH track students):

SPEECH1: Introduction to speech data
Describe the differences between high-level features and low-level short-term Fourier transform (STFT) based features.

SPEECH2: Speech synthesis

Describe why the parameters of the source-filter model can be determined by using linear prediction in order to synthesise a speech signal S(n).

SPEECH3: Factor analysis for speaker recognition



Q1: An i-vector is a compressed representation of variable-duration utterances. It is widely described with the following expression

$$\mathbf{m}_r = \mathbf{m}_o + \mathbf{T}\mathbf{h}_r$$

Explain the role of variable $\mathbf{h}_r$ in this model.

Q2: The generative equation of a single Gaussian total variability model could be expressed as follows.

$$\begin{bmatrix} o_1 \\ o_2 \\ \vdots \\ o_T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_o \\ \vdots \\ \boldsymbol{\mu}_o \end{bmatrix} + \begin{bmatrix} \mathbf{W} \\ \mathbf{W} \\ \vdots \\ \mathbf{W} \end{bmatrix} \mathbf{h}_r + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_T \end{bmatrix}$$

Explain the rationale of latent variable tying across frames in the model.

SPEECH4: Prosody prediction

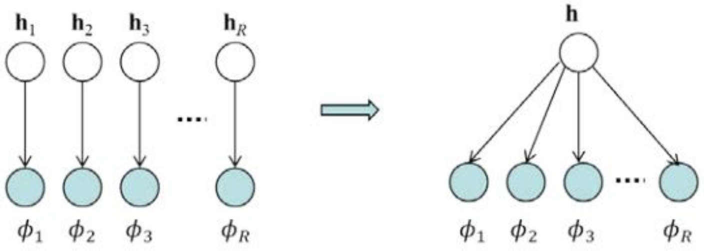
1) Why do we need models to connect between prosodic parameters and information transmitted in the speech waveform?

2) What are the three components of the Fujisaki model and what do they relate to?

SPEECH7: SIDEKIT - a tool for speaker recognition

Q1: Sufficient statistics are pre-whitened prior to i-vector extraction. This pre-whitening step does not change the i-vector as similar transformation is observed by the T matrix. Proof this.

Q2: A probabilistic LDA (PLDA) model is an extension to the classical factor analysis model by tying of observed variables



Explain the rationale of tying multiple observed variables (in the context of i-vector PLDA speaker recognition system) to a single latent variable.

SPEECH8: Speech enhancement

Describe characteristics of filtering-based speech enhancement and model-based speech enhancement.