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):

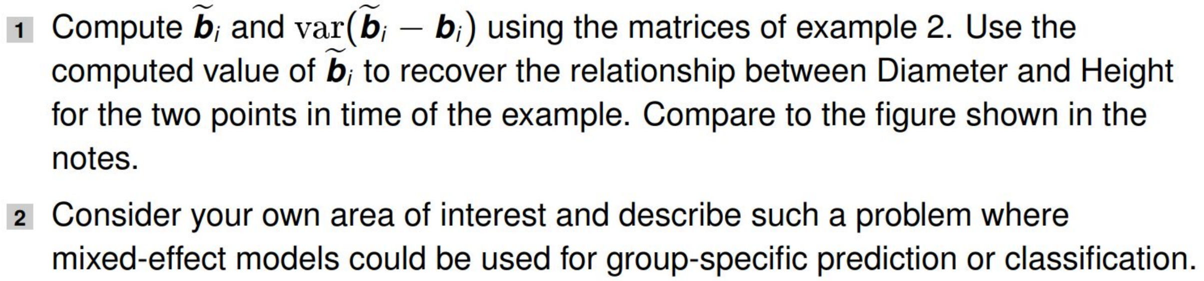
ML1. Introduction to biomedical data

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

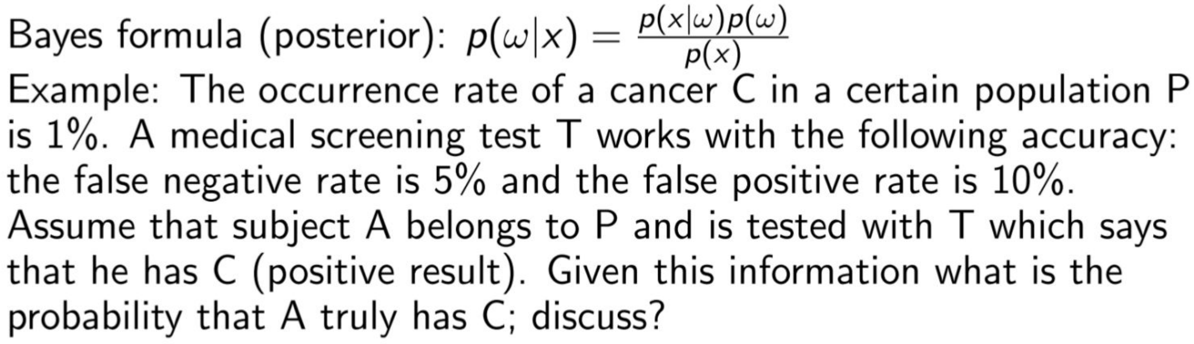
Find out what is the working principle of mel-frequency cepstral coefficient (MFCC) feature extraction and what are its main applications.

ML3. Statistics



1. Compute $\widetilde{\boldsymbol{b}}_i$ and $\mathrm{var}(\widetilde{\boldsymbol{b}}_i - \boldsymbol{b}_i)$ using the matrices of example 2. Use the computed value of $\widetilde{\boldsymbol{b}}_i$ to recover the relationship between Diameter and Height for the two points in time of the example. Compare to the figure shown in the notes.
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

ML8. CSC

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.

SPEECH track sessions (mandatory to SPEECH track students):

SPEECH1: Introduction to speech data

SPEECH2: Speech synthesis

SPEECH3: Factor analysis for speaker recognition

SPEECH4: Prosody prediction

SPEECH5: ML tools for speech data

SPEECH6: Spoofing

SPEECH7: SIDEKIT - a tool for speaker recognition

SPEECH8: Speech enhancement