Exam Data Science

(14X026)

Below is a list of questions whose number will be drawn randomly at the exam. You will have to answer 2 questions (one for part I, one for part II) in the time slot allocated. **No preparation time** will be allowed. Please be ready with your material directly and **have this list of questions available** on your side when connecting.
When answering do not hesitate to use knowledge also gathered with the TPs.

Part I:

1. Explain the "sparsity" and "concentration" aspects of the Curse of Dimensionality and how they are theoretically justified and their practical impact. Provide examples of Blessing of Dimensionality and how they are used in practice (you may seek other examples than that in the lectures or labs).

2. Explain in detail the method of Principal Component Analysis (PCA) including assumptions, calculation, and interpretation of the technique.

3. When performing Principal Component Analysis (PCA), how should one choose the number of eigenvectors to keep? Explain why and discuss the implication on the reconstruction.

4. Explain how the model of Linear Discriminant Analysis (LDA) relate to conditional model for classification.

5. Explain the method of Linear Discriminant Analysis (LDA) including assumptions, calculation, and interpretation of the technique. How is this technique used for classification?

6. Explain the Gaussian Mixture Model including assumptions, calculation, and interpretation of the model. What are its parameters? Explain how to optimize them given some data.

7. What is the EM algorithm and how is it related to the Gaussian Mixture Model? In particular, explain the concept of responsibility.

8. Explain the problem of clustering and describe the k-means clustering algorithm. In particular, discuss the function it minimizes and its relation to a Gaussian model.

Part II:

1. Considering multiple random variables, explain the chain rule decomposition for the joint probability. Provide examples of different statistical relationships between random variables. Explain the Markov chain of the first and second orders.

2. Give the definition of entropy for discrete random variables. Explain the main properties of entropy. Explain the entropy of binary random variables.

3. Explain the joint and conditional entropy for discrete random variables. Explain the chain rule for entropy.

4. Explain the properties of entropy on the examples of conditional entropy H(X|Y) and joint entropy H(X,Y). Use Venn diagrams.

5. Explain the relative entropy, its properties and provide some examples of its usage.

6. Explain the cross-entropy and provide some examples of its usage. Show a link between the cross-entropy and relative entropy.

7. Give the definition of mutual information. Explain the Venn diagram. Demonstrate different expressions for mutual information via entropies and relative entropy.

8. Explain the chain rules for probability and entropy. Explain the development of mutual information I(X; Y1,Y2,Y3).

10. Explain the difference between the characteristic function and moment generating function.

11. Explain 3 main concepts of feature extraction based on supervised learning, auto-encoding and self-supervised learning. Give the examples of usage of these features.

12. Explain the difference between the Neyman-Pearson and Bayesian threshold selection for binary hypothesis testing.