LUT University

12.03.2026

Student's name: _____ Student number: _____

BM20A6100 Advanced Data Analysis and Machine Learning

Only writing instruments are allowed in the exam.

Exam questions cover topics from Data Analysis (period 1) and Machine Learning (period 2).

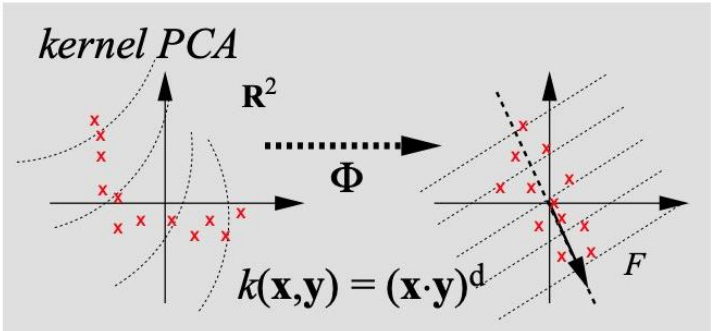
Questions (3) related to **period 1:**

1. Fill in the following sentences:
   a) SVD: $X = U S V^T$. Matrices are called _____, _____, _____, and _____.
   b) Economy SVD is different from full SVD because _____.
   c) PCA: The first principal component extracts _____.
   d) R^2 is called _____ and it describes _____ .
   e) PLS and _____ are the two most applied multivariate regression methods.

2. PLS regression:
   a) Give examples of features in X or Y data that encourage utilization of PLS regression.
   b) Inner validation utilizes calibration samples for defining the number of latent variables. Describe the procedure for autocorrelated time series.
   c) Reporting: list max 5 details that you see the important ones to report from the modelling procedure (PLS models).

3. Choose either robust PCA *OR* kernel PCA and answer the questions:
   a) How is this extension different from PCA?
   b) Give two examples on cases (with different objective), when the method should be utilized.
   c) Describe the theory.
   d) Any consideration or metrics that should be used in validation or during the modelling process.

   These figures are from the lecture materials and might help you. Choose only one method (RPCA or KPCA).



LUT University

LUT Computer Vision and Pattern Recognition Laboratory 2026-03-04

BM20A6100 Advanced Data Analysis and Machine Learning

Lasse Lensu

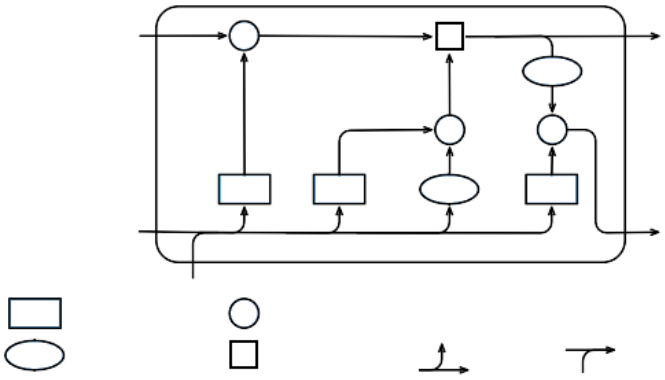
Exam 2026-03-12: Period 2 Student number: _____ Student name: _____

Instructions: i) **Only writing instruments are allowed in the exam**. ii) Justify your answers well, but follow possible task-specific instructions about the length of the answer. iii) For tasks requiring a method implementation in principle, pseudocode or natural language may be used to describe an algorithm. iv) **If you return this paper with your answers, fill in the student number and name above**.

1. Understanding of concepts (5x2 p): Give **compact explanations** for the following terms. In the case of more than one term (... *vs.* ...), explain also the differences of the mentioned terms. The maximum length of the answer is **1 page for the whole task**.
   (a) attention mechanism
   (b) data augmentation *vs.* data imputation
   (c) generative adversarial network
   (d) one-hot encoding *vs.* word embedding
   (e) semi-supervised learning

2. Understanding of methodology (10 p): Return this paper with your other exam answers (or redraw the pictures).

   (a) What does the model shown in the picture represent?
   (b) Label all the inputs, components and outputs of the model.
   (c) Which are the relevant variables of the model and how can they be determined?



A model modified from [1].

(Continues on the other side.)

PO BOX 20, FI-53851 LAPPEENRANTA, FINLAND
E-mail: lasse.lensu@lut.fi

3. Troubleshooting (5x2 p): For each presented challenge, provide a **short description** which options are there to solve the challenge. The maximum length of the answer is **1 page for the whole task**.
   (a) Amount of variables: A data set contains too many variables for ... to be efficient.
   (b) Class imbalance: The number of samples in each data set class varies significantly.
   (c) Discriminative classifier: The model is overly confident when classifying out-of-distribution data.
   (d) Model training: Training a neural network with randomly initialised parameters takes too much time.
   (e) Observability: All the relevant variables cannot be directly observed for training a machine learning model.

References

[1] Aston Zhang, Zachary C Lipton, Mu Li, and Alexander J Smola. *Dive into Deep Learning*. 2023.