Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNTI Winter Term 2020/2021

Exercise Sheet 2

Principal Component Analysis

Deadline: 24.11.2020, 23:59

Instructions

This assignment will introduce you to Principal Component Analysis. A good resource to get a basic understanding of how PCA works is this paper, though it uses a slightly different notation as the lecture slides.

Exercise 2.1 - Covariance Matrix (1 + 1 points)

Calculating a covariance matrix is an important step in many machine learning algorithms. Given a matrix of observed data points $\boldsymbol{X}$ (with the rows representing all measurements from a measurement type and the columns single data points), the covariance matrix is defined as $\boldsymbol{C_X} = \boldsymbol{X}\boldsymbol{X^T}$.

a) Which parts of a covariance matrix capture the variance in the dataset? Which capture the covariance? Motivate your answer.

b) Prove or motivate whether any covariance matrix has the following properties or not:

- square (motivate)
- symmetric (motivate)
- positive semidefinite, i. e. $\boldsymbol{v^T C v} \geq 0$ (prove)
- has only positive eigenvalues and orthogonal eigenvectors (motivate)

Exercise 2.2 - Principal Component Analysis (2 + 2 + 1 points)

Consider the following dataset consisting of $n = 5$ data points and $m = 2$ observations each:

$$x^{(1)} = (0, 0),\ x^{(2)} = (1, -1),\ x^{(3)} = (2, -2),\ x^{(4)} = (3, -3),\ x^{(5)} = (4, -4).$$

a) Reduce the dimensionality of the dataset from *2D* to *1D* by performing PCA on it as described in the paper (you should do the calculations by hand):

1) Arranging the dataset into a matrix $\boldsymbol{X}^{m \times n}$
2) Centering the matrix by subtracting row means
3) Calculating the covariance matrix $\boldsymbol{C_X} = \boldsymbol{X}\boldsymbol{X^T}$

4) Derive its eigenvalues and eigenvectors

6) Choose the principal component that catches most of the variance

7) Use it to reduce the dimensionality of the data!

What do the eigenvalues of the covariance matrix tell you?

b) Answer the following questions and provide an explanation for your answer. Write at most 4-5 sentences.

i) The purpose of PCA is to reduce spurious dimensions in the data. Why do we need dimensionality reduction, and when should it be applied? When not?

ii) Can PCA be used to reduce the dimensionality of a non-linear data?

iii) Ordering the 'principle components' by the eigenvalues of the covariance matrix means that PCA relies only on the mean and the variance of a dataset, meaning that the data can be described by a Gaussian (probability) distribution. What happens if the data is not Gaussian distributed?

c) Read this blog post and discuss how PCA can be seen as an auto-encoder.

Exercise 2.3 - Principal Component Analysis with PyTorch (1 + 1 + 1 points)

See ipython notebook

Submission instructions

The following instructions are mandatory. If you are not following them, tutors can decide to not correct your exercise.

- You can and are encouraged to submit the assignment as a team of two students. Submitting as a team will be mandatory for the next assignment.
- Hand in zip file containing the PDF with your solutions and the completed ipython notebook.
- Therefore Make sure to write the Microsoft Teams user name, student id and the name of each member of your team on your submission.
- Your assignment solution must be uploaded by only **one** of your team members to the 'Assignments' tab of the tutorial team (in **Microsoft Teams**).
- If you have any trouble with the submission, contact your tutor **before** the deadline.