

# Assignment 3: PCA, Kmeans and Kernel Methods

Started: 4 Oct at 15:34

## Quiz instructions

Please read the following instruction carefully before you complete assignment 3

### Instructions and submission guidelines:

- This is not a group assignment. Everyone is asked to complete the assignment individually.
- The assignment consists of five sub-tasks. You need to provide your solutions to the corresponding sections in this quiz.
- You need to also upload your code. We will check the code to ensure that your results can be generated from your code. If your result does not match with your code, you will get 0 mark for the relevant sections.
- You should use Python to finish this assignment. Jupyter notebook is allowed and encouraged.

### Data:

Please download the data that will be used in this assignment here: [mnist.csv](https://myuni.adelaide.edu.au/courses/75049/files/11310714/download?download_frd=1) ↓  
([https://myuni.adelaide.edu.au/courses/75049/files/11310714/download?download\\_frd=1](https://myuni.adelaide.edu.au/courses/75049/files/11310714/download?download_frd=1))

The above data is a subsampled version of the MNIST dataset. It contains images for 10 digits (10 classes). The dataset contains 6,000 samples. The images from the data set have the size 28 x 28. They are saved in the csv data files. Every line of these files consists of an image, i.e. 785 numbers between 0 and 1. The first number of each line is the label, i.e. the digit which is depicted in the image. The following 784 numbers are the pixels of the 28 x 28 image.

### Third-party Libraries,

You can use any third-party libraries to read and process data.

You can use numpy in any question.

Please follow the instruction of each question to complete the relevant parts.

**Question 1****25 pts**

Perform PCA on the dataset to reduce each sample into a 10-dimensional feature vector. Show the covariance matrix of the transformed data. Please also copy your code snippet here.

Edit View Insert Format Tools Table

12pt ▾ Paragraph ▾ | **B** *I* U A ▾  ▾  $T^2$  ▾ | ⋮

p



0 words





**Question 2****25 pts**

Perform k-means clustering to cluster the dataset (without applying PCA) into 10 groups. Please copy your code snippet here.

Edit View Insert Format Tools Table

12pt ▾ Paragraph ▾ | **B** *I* U A ▾  ▾  $T^2$  ▾ | ⋮

p



 | 0 words |   

Question 3






10 pts

Please plot the loss curve, that is, the change of loss value of the k-means algorithm with respect to the number of iterations

EditViewInsertFormatToolsTable

12pt ▾ Paragraph ▾ | **B** *I* U A ▾  ▾  $T^2$  ▾ | 

p

 | 0 words |   

Question 4

10 pts

Please use the first 4000 samples as the training set and remaining 2000 samples as the validation set, and design a way to choose the best k in k-means algorithm. Please copy your code snippet here.

Edit View Insert Format Tools Table

12pt ▾ Paragraph ▾ | **B** *I* U A ▾  ▾  $\text{T}^2$  ▾ | ⋮

p



0 words



## Question 5

30 pts

Please implement kernel k-means algorithm with RBF-kernel, that is,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right).$$

The hyper-parameter can be empirically set to


$$2\sigma^2 = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$$

Please only use the first 500 samples and cluster them into 5 groups. This is for reducing the running time of your code.





Please copy your code snippet here.

TIPS: If you can use matrix operations to replace summations, your code will be more efficient. However, this is just optional.

EditViewInsertFormatToolsTable

12pt ▾Paragraph ▾|**B***I*UA ▾ ▾ $\text{T}^2$  ▾|⋮

p

|0 words|⋮

Question 6

0 pts

Please upload your code for this assignment here.

Upload

Choose a file

Saved at 10:49

Submit quiz