Machine Learning	Prof. Dr. Bernt Schiele
	Dr. Paul Swoboda & Dr. Gerard Pons-Moll
Exercise Sheet 10 - 07.01.2019	due: 14.01.2019, 14:15

Exercise 18 - Clustering

Please implement the clustering algorithms discussed in the lecture.

- **(2 Points)** Complete the function `kmeans_diy` for the Llyod's algorithm for k-means clustering.
- **(3 Points)** Complete the function `spectralclustering_diy` for the spectral clustering algorithm with unnormalized graph Laplacian:

  $$L = D - W$$

  where D is the `degree matrix` and W is the `weighted adjacency matrix` of the graph. The similarity graph here is set to be the fully connected type, i.e., we construct all points with positive similarity with each other and we weight all edges by

  $$s(x_i, x_j) \ = exp(-\|x_i - x_j\|^2/\gamma)$$

  where x_i, x_j denote the i-th and j-th data points in the dataset, γ is fixed to be 1 in this exercise.
- **(2 Points)** Now apply your clustering code (k-means and spectral clustering) to the provided dataset `ex9_data` and visualize your results by running the script `ex9_main.m`. The data file contains two datasets `data1` and `data2`, where each includes two groups of data points X_1, X_2in R^2, you need to apply the clustering on `data1` and `data2` separately. What's your observation and conclusion?

Hints:

- For spectral clustering, you can use *eig* function to compute the eigen vector and eigen value of a matrix.
- You are recommended to implement a function `compute_pairwise_distance` to compute pairwise distance between two datasets which can be shared by both of your clustering algorithms' code.

Exercise 19 - Distances in High Dimensions

Let X and Y be two independent $\mathbb{R}^d$-valued random variables with expectations

$$\mathbb{E}[X] = \mu_X = (\mu_{X,1} \dots \mu_{X,d})^T, \qquad \mathbb{E}[Y] = \mu_Y = (\mu_{Y,1} \dots \mu_{Y,d})^T,$$

and covariances

$$\mathbb{E}[(X - \mu_X)(X - \mu_X)^T] = \sigma_X^2 \mathbb{1}, \qquad \mathbb{E}[(Y - \mu_Y)(Y - \mu_Y)^T] = \sigma_Y^2 \mathbb{1}, \qquad \sigma_X^2, \sigma_Y^2 > 0,$$

with $\mathbb{1}$ denoting the identity matrix in $\mathbb{R}^d$.

a. **(3 Points)** Show that the expected squared Euclidean distance of X and Y is given by

$$\mathbb{E}\left[\|X - Y\|^2\right] = d\sigma_X^2 + d\sigma_Y^2 + \|\mu_X - \mu_Y\|^2.$$

b. **(1 Point)** Suppose $\sigma_X^2 > \sigma_Y^2$ and consider the limit $d \to \infty$, assuming that $\|\mu_X - \mu_Y\|^2 = C$, where C is a positive constant independent of d. Let Z be a random variable with the same distribution as X which is independent of X. Using the result in b., compare $\mathbb{E}\left[\|X - Y\|^2\right]$ to $\mathbb{E}\left[\|X - Z\|^2\right]$.

Hints:

a. Use that $\|X - Y\|^2 = \|(X - \mu_X) - (Y - \mu_Y) + (\mu_X - \mu_Y)\|^2$.

b. Exploit the linearity of $\mathbb{E}$.

Submission instructions

- We accept both handwritten and electronic submissions. So you can choose what is more convenient for you. In any case, you should specify full names and immatriculation IDs of all team members. Obviously, programming tasks you can submit only electronically.
- Handwritten submissions should be submitted in the lecture hall of Monday's lecture (before the lecture starts).
- Electronic submissions should be zipped, containing the m-files (`Basis` etc.), your plots (png files) and the matlab data files (.mat) and emailed to the corresponding tutor:

  a. Apratim Bhattacharyya (Wednesday 8-10): abhattac@mpi-inf.mpg.de

  b. Maksym Andriushchenko (Thursday 8-10): s8mmandr@stud.uni-saarland.de

  c. Max Losch (Friday 16-18): mlosch@mpi-inf.mpg.de

  If not all 3 students belong to the same tutorial group, then you should email your submission to **only** one tutor (e.g. to the tutor of the first author of your homework), so please do not put other tutors in copy of the email.

  The email subject must have the following form: "[ML18/19 Exercise] Sheet X", where X is the number of the current exercise sheet. Then please specify in the email full names and immatriculation IDs of all team members. Then please attach all your files as a single zip archive, which consists of your immatriculation IDs, e.g. "2561234_2561235_2561236.zip".

- Reminder: you should submit in groups of 3. Otherwise, we will later on merge the groups smaller than 3 students.