TECHNISCHE UNIVERSITÄT MÜNCHEN
Fakultät für Elektrotechnik und Informationstechnik
Lehrstuhl für Datenverarbeitung
PD Dr. Martin Kleinsteuber

Information Retrieval in High Dimensional Data

Assignment #2, 10.12.2020

Due date: 07.01.2021, 11 P.M.

Please hand in your solutions via Moodle as an IPYTHON (Jupyter) notebook.

Solutions must be handed in by groups. Please state the names of your group members at a prominent place in your submission. (For example, at the beginning of your provided notebook or in a separate text file.)

Principal Component Analysis

Task 1: [10 points] Let $\mathbf{X} \in \mathbb{R}^{p \times N}$, $p < N$ be a centered matrix of p N-dimensional data samples with the full-size SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Assume that the singular values are sorted in a descending manner, i.e. $\sigma_{1,1} \geq \ldots \geq \sigma_{p,p}$.

- Provide a normalized vector $\hat{\mathbf{s}} \in \mathbb{R}^p$, such that

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s} \text{ s.t. } \|\mathbf{s}\|=1} \mathbf{s}^\top \mathbf{\Sigma}\mathbf{\Sigma}^\top \mathbf{s}.$$

- Show that the empirical variance of the inner products of the columns of $\mathbf{X}$ with a normalized vector $\mathbf{a}$,

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{a}^\top \mathbf{x}_i)^2 = \frac{1}{N}\mathbf{a}^\top \mathbf{X}\mathbf{X}^\top \mathbf{a},$$

  is maximized when $\mathbf{a}$ is set to the first column of $\mathbf{U}$, i.e. $\mathbf{a} = \mathbf{u}_1$ (note that $\|\mathbf{a}\| = 1$).
  Hint: Write $\mathbf{a}$ as a linear combination of the columns of $\mathbf{U}$. Verify that such a representation does not affect the norm constraint.

Task 2: [15 points] For this task, download the modified version of the *Yale Face Database B* provided on Moodle (`task2_data.zip`). The Yale Face Database B consists of single light source images of 10 subjects, each seen in different poses and illumination conditions. In the provided form the database is divided into 5 subsets. In subset 0 the subject is illuminated by an almost frontal light source, while for subsets 1-4 the

light source is gradually moved along the horizon. Subset 0 will serve as the training set, while subsets 1-4 are used for testing.

- Write a function that takes as an input matrix $\mathbf{T}$ of vectorized images from subset 0. The output of this function are the 20 first singular vectors $\mathbf{U}[:,1],\ldots,\mathbf{U}[:,20]$. Display the first 3 vectors as images, i.e., reshape them to size 50×50 and display them.

- Write a function that takes as an input the training set $\mathbf{T}$ (a matrix composed of vectorized pictures from subset 0), a vector containing the labels of the training set (i.e., if the the i-th sample belongs to class j, the i-th entry of the labels vector is j. In this exercise j is an integer between 1 and 10), the test samples S (a matrix composed of vectorized pictures from subsets 1-4) and the corresponding labels (in a separate vector), the 20 singular vectors from the first step, and the parameter k that denotes how many of the PCs are used. Use the Euclidean distance to classify each sample image based on its three nearest neighbors. (This is done by comparing the test samples with the training samples in the reduced space.) As an output give the fraction of images from S that were misclassified, i.e., the error rate. Repeat this for subsets 1 through 4 and for $k = 1,\ldots,20$. Plot the error rate for each subset.

- Repeat the above experiment without using the first three singular vectors, i.e., use $k = 1,\ldots,17$ singular vectors starting from the 4-th. Plot the error rate as before. How do you explain the difference in recognition rate?