Midterm Homework 7 - due Wed. Apr. 1st
Mathematical and Computational Foundations of Data Science

Instructor: Mauro Maggioni
Office: 302D Whitehead Hall
Web page: https://mauromaggioni.duckdns.org
E-mail: mauro.maggioni *at* youknowwhat.edu

Homework Policies. Unlike other homework sets, for this set you will work by yourself with no help from anyone else. It is open book, open notes. This homework will count as two homework sets. Make sure your answers are detailed and well-motivated. The code and corresponding plots should be submitted together with everything else. Submission is electronic via e-mail to me, with scanned/imaged pages in one .pdf (or .zip file if multiple files are included) with name 110.445_FirstLastName.pdf, sent to me in an e-mail with the title "Homework 7 for 110.445". Thank you.

Assignment

Exercises

Exercise 1 (30pts). Let $X \in \mathbb{R}^{D\times n}$ be a matrix representing n data points, each a D-dimensional vector. We think of the columns $X_i \in \mathbb{R}^D$ as independent identically distributed samples from some probability measure.

a. Define what the Principal Component Analysis of X is, carefully specifying the properties (and dimensions) of the matrices U, Σ, V involved. Do not forget to center the data first (i.e. subtract the empirical mean $m = \frac{1}{n}\sum_i X_i \in \mathbb{R}^D$ of the columns from each column X_i).

b. How does PCA help in finding a low-rank approximation to the data? What is a geometric interpretation of such low-rank approximation? Which optimization problem can easily be solved with PCA?

c. Recalling that the empirical variance of n i.i.d. samples $z_1, \dots, z_n$ of a random variable is $var[Z] = \frac{1}{n-1}\sum_{i=1}^n (z_i - \overline{z})^2$, where $\overline{z} = \frac{1}{n}\sum_{i=1}^n z_i$ is the empirical mean. Compute the empirical variance of the data X, centered, projected onto each principal component, and show it is equal to σ_k^2, i.e. $(n-1)var[\{\langle X_i - m, U_k\rangle\}_i] = \sigma_k^2$, where σ_k is the k-th singular value of X.

d. Let $Proj_k$ be the projection (to be applied to centered data) onto the first k principal components (the first k columns of U): write an expression for such projection. Then project the centered data onto the first k principal components: what is the mean squared error incurred by such projection, i.e. $\frac{1}{n}\sum_{i=1}^n ||X_i - m - Proj_k(X_i - m)||^2$?

Exercise 2 (20pts). Define the space $L^2([0,\pi])$ of square-integrable functions on $[0,\pi]$, as well as its natural norm and inner product. Then compute the orthogonal projection of $f(x) = \sin(x)$ onto the subspace spanned by $\{1, x, x^2\}$. Finally, compare that orthogonal projection with the second-order Taylor expansion of f centered at $x = 0$: are these two approximations to f by a degree 2 polynomial the same? If yes, why, if not, discuss which one of the two projections is closer, in the L^2-distance, to f.

Exercise 3 (50pts). Consider data sampled from a mixture of 2 Gaussians in $\mathbb{R}^D$: $\mu = wg_1 + (1-w)g_2$ where $0 \le w \le 1$, and $g_i = \mathcal{N}(\mu_i, \Sigma_i)$. A sample is obtained by drawing from the Gaussian g_1 with probability w or from g_2 with probability $(1-w)$.

a. What is the problem of clustering, and what is K-means clustering? Write down the objective functional minimized by K-means, and discuss why finding an optimal solution of this functional is likely very difficult.

b. Write code that samples n points in $\mathbb{R}^D$ according to the distribution above, and runs K-means (with $K = 2$ of course), and evaluates the performance of K-means both in terms of the value of the K-means functional, and in terms of the percentage of points correctly assigned to clusters (you may reuse pieces of your own code from past homework if that helps).

b. Suppose $D = 2$, $w = 1/2$, $\mu_1 = (0, 0)$, $\mu_2 = (m, 0, \ldots, 0)$, $m > 0$ and $\Sigma_1 = \Sigma_2$ is a diagonal matrix with diagonal entries $(1, 2)$. Study numerically the clustering performance of K-means (percentage of correctly clustered points) as $m \to 0^+$. In particular, show a graph of performance as m decreases from 1 down to 0, and discuss the results.

c. If a rotation is applied to the data in example (b), would the results change? If yes, how, if not, why?

d. An affine transformation $A = \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix}$, $\alpha > 0$ is applied to the data in (b): what does this transformation do to points in $\mathbb{R}^2$? to the distribution of the data? Then study how the performance of K-means changes for fixed m (say, $m = 1$, or you could also consider different values of m), as a function of increasing α. Explain and discuss the results.

e. Same questions (b)-(d) above but using Linear Discriminant Analysis instead of K-means. This is a supervised technique, and so the data above is going to be labelled training data, and the labelling performance will be measure on a test set constructed in the same way (with, say, the same size n). Discuss your results for LDA, and compare them to those of K-means. [you do not have to implement LDA yourself if you do with to do: there are functions implementing it in Matlab, Python, R, etc...].