Introduction to learning and analysis of big data
Exercise 3

Dr. Sivan Sabato

Fall 2017/8

Submission guidelines, **please read and follow carefully**:

- You may submit the exercise in pairs.
- Submit using the submission system.
- The zip file should include **exactly three files in the root - no subdirectories please**.
- The files in the zip file should be:
  1. A file called "answers.pdf" - The answers to the questions
  2. A file called "kmeans.m"
  3. A file called "singlelinkage.m"
- For questions use the course Forum, or if they are not of public interest, send them via the course requests system.

Question 1. (a) (10pts) Implement the k-means heuristic algorithm for Euclidean metric which we learned in class. The function should be implemented in the submitted file called "kmeans.m". The first line in the file (the signature of the function) should be:

```
function C = kmeans(X, k, t)
```

The input parameters are:

- k - the number of clusters
- t - the number of iterations to run
- X - a 2-D matrix of size $m \times d$. Row i in this matrix is a vector with d coordinates that describes example x_i from the training sample.

The function returns the variable C, which is a column vector of length m, where $C(i) \in \{1, \ldots, k\}$ is the identity of the cluster in which x_i has been assigned.

(b) (10pts) Implement the single-linkage algorithm that we learned in class, again using the Euclidean metric. The function should be implemented in the submitted file called "singlelinkage.m". The first line in the file (the signature of the function) should be:

```
function C = singlelinkage(X, k)
```

The input parameters are:

- k - the number of clusters
- X - a 2-D matrix of size $m \times d$. Row i in this matrix is a vector with d coordinates that describes example x_i from the training sample.

The function returns the variable C, which is a column vector of length m, where $C(i) \in \{1, \ldots, k\}$ is the identity of the cluster in which x_i has been assigned.

(c) (5pts) Run your k-means code on an **unlabeled** random sample of size 1000 generated from all the digits in the MNIST data file `mnist_all.mat`, with $k = 10$. Use the resulting clustering and the true labels of the points in the sample, to provide a table showing, for each cluster, (1) what is its size (2) which label is most common in it, and (3) what percentage of the points in the cluster have this label. Report the classification error on the sample, that would result if we classified all the points in each cluster using the cluster's most common label. Explain your calculation.

(d) (5pts) Repeat (c) for your single linkage algorithm, again reporting the table and the classification error. Which clustering algorithm worked better for this problem?

(e) (5pts) Run kmeans and single-linkage again on the same data set from MNIST, this time set $k = 8$. Again provide the table of results and the classification errors. Considering the way the two algorithms work, explain the differences in their results when moving from $k = 10$ to $k = 8$.

Question 2. In an experiment, several measurements were taken at times $t = 1, 2, \ldots, m$. At each time t, the measurements taken were $x_t(1), x_t(2), x_t(3), x_t(4)$. This created a data set $S = x_1, \ldots, x_m$, where x_t is a vector in $\mathbb{R}^4$ which includes all the measurements from time t. PCA was performed on the data set S to reduce its dimensionality from 4 to 2.

(a) (10pts) In one experiment, it turned out that in all times t, $x_t(3) = 2x_t(1) + x_t(2)$, and $x_t(4) = x_t(3) - 4x_t(1)$. What will be the distortion of the PCA in this case? Prove your claim.

(b) (10pts) In another experiment, it turned out that in all times t, $x_t(3) = (x_t(1))^2 + (x_t(2))^3$, and $x_t(4) = (x_t(3) - x_t(1))^2$. Show an example of experiment results that satisfy these equations such that the distortion of the PCA is larger than the distortion you showed for the experiment in (a). You may choose m as you like.

Question 3. Consider the following distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{-1, 1\}^2$ and $\mathcal{Y} = \{-1, 1\}$.

x(1)	x(2)	y	prob.
-1	-1	$+1$	$5/60$
-1	$+1$	$+1$	0
$+1$	-1	$+1$	$11/60$
$+1$	$+1$	$+1$	$4/60$
-1	-1	-1	$6/60$
-1	$+1$	-1	$24/60$
$+1$	-1	-1	$2/60$
$+1$	$+1$	-1	$8/60$

(a) (10pts) Does the Naive-Bayes assumption hold for this distribution? Prove your claim.

(b) (10pts) Suppose we had a sample $S \sim \mathcal{D}^m$, such that the frequencies of the possible (x, y) in the data set was *exactly* the same as in the distribution, and suppose that we then ran the Naive-Bayes algorithm on this data set. What predictor would we get from this algorithm? Prove your claim.

(c) (5pts) Compare the Bayes-optimal predictor $h : \mathcal{X} \to \mathcal{Y}$ for the distribution $\mathcal{D}$ to the one you got in question (b). Explain why your answer does not stand in contradiction to (a).

Question 4. Let $\mathcal{X} = \{0, 1, 2\}$. Let $\Theta \subseteq [0, 1]^3$ such that for $\theta \in \Theta$, $\theta(1) + \theta(2) + \theta(3) = 1$. Define a *Trinomial* distribution $\mathcal{D}_\theta$ for $\theta \in \Theta$ as follows: $\mathbb{P}_{X \sim \mathcal{D}_\theta}[X = i] = \theta(i)$. Assume that we have a sample $S = x_1, \ldots, x_m \sim \mathcal{D}_\theta^m$.

(a) (10pts) Let $\Theta' = \{\theta \in \Theta \mid \theta(1) = 3\theta(2)\}$. Give an explicit formula for the value of the maximum likelihood estimator $\hat{\theta}$ using $x_1, \ldots, x_m$, assuming that $\theta \in \Theta'$. Prove your claim.

(b) (10pts) Consider a distribution which is a mixture of k densities, each density coming from $\{f_\sigma \mid \sigma > 0\}$, where f_σ is the density of a Gaussian random variable $N(1, \sigma^2)$.

- Write down a parametrized expression for the mixture distribution. Define a parameter set Θ which includes all (and only) the possible parameter settings of this mixture distribution.
- Define a multinomial random variable Z over $\{1, \ldots, k\}$. Suppose that we get an augmented sample $(x_1, z_1), \ldots, (x_m, z_m)$, with $S = (x_1, \ldots, x_m)$, $Z = (z_1, \ldots, z_m)$. Write down the augmented log-likelihood $L(S, Z; \theta)$, where $\theta \in \Theta$, and derive the maximum-likelihood estimator for θ, assuming that both S and Z are given.