CS-E3210- Machine Learning Basic Principles
Home Assignment 5 - "Clustering"

Your solutions to the following problems should be submitted as one single pdf which does not contain any personal information (student ID or name). The only rule for the layout of your submission is that for each problem there has to be exactly one separate page containing the answer to the problem. You are welcome to use the LaTeX-file underlying this pdf, available under `https://version.aalto.fi/gitlab/junga1/MLBP2017Public`, and fill in your solutions there.

Problem 1: Hard Clustering

Consider $N = 20$ snapshots, available at `https://version.aalto.fi/gitlab/junga1/MLBP2017Public/tree/master/Clustering/images`, which are named according to the season when they have been taken, i.e., either "winter??.jpeg" or "summer??.jpeg". We represent the ith snapshot, with $i = 1, \ldots, N$, by the feature vector $\mathbf{x}^{(i)} = (x_{\rm r}^{(i)}, x_{\rm g}^{(i)})^T \in \mathbb{R}^2$ with the total image redness $x_{\rm r}$ and greenness $x_{\rm g}$. Thus, the overall dataset is given by the feature vectors $\{\mathbf{x}^{(i)}\}_{i=1}^N$, which are divided into two subsets $\mathbb{X}^{(\rm summer)}$ and $\mathbb{X}^{(\rm winter)}$, which contain only the feature vectors of summer or winter snapshots, respectively. Apply the k-means algorithm, using a fixed number of M iterations, for clustering the dataset $\mathbb{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ into two non-overlapping clusters $\mathcal{C}_0$, $\mathcal{C}_1$ such that each snapshot belongs exactly to one of the clusters $\mathcal{C}_0$ and $\mathcal{C}_1$. Let us characterize the "quality" of the clusters by how well they separate winter from summer images. To this end, we define the "purity" measure $P_{\rm w} = h\big(\frac{|\mathcal{C}_1 \cap \mathbb{X}^{(\rm winter)}|}{|\mathbb{X}^{(\rm winter)}|}\big)$ and $P_{\rm s} = h\big(\frac{|\mathcal{C}_1 \cap \mathbb{X}^{(\rm summer)}|}{|\mathbb{X}^{(\rm summer)}|}\big)$ with the function $h(p) = 1 + p\log_2 p + (1-p)\log_2(1-p)$. The average purity obtained from the k-means output is then $\bar{P} = (1/2)(P_{\rm w} + P_{\rm s})$. Implement the k-means algorithm using different numbers M of iterations and plot the average purity $\bar{P}$ obtained for different values of M. For each choice of M, repeat the application of k-means several (say 10) times (runs), and use for each run two (independently) randomly selected feature vectors $\mathbf{x}^{(i)}, \mathbf{x}^{(j)} \in \mathbb{X}$ as the initial choices for the cluster means $\mathbf{m}_0$ and $\mathbf{m}_1$. Average the results of the different runs to get one single estimate of $\bar{P}$ for each M.

Answer.

Iteration numbers ranging from 1 to 15 was used and for each iteration number ten different random initialization points for two cluster means was used. For each random initialization the purity measure was calculated and then averaged over all the ten random initializations. The resulting plot showing the average purity measure as a function of number of iterations is shown in figure 1.

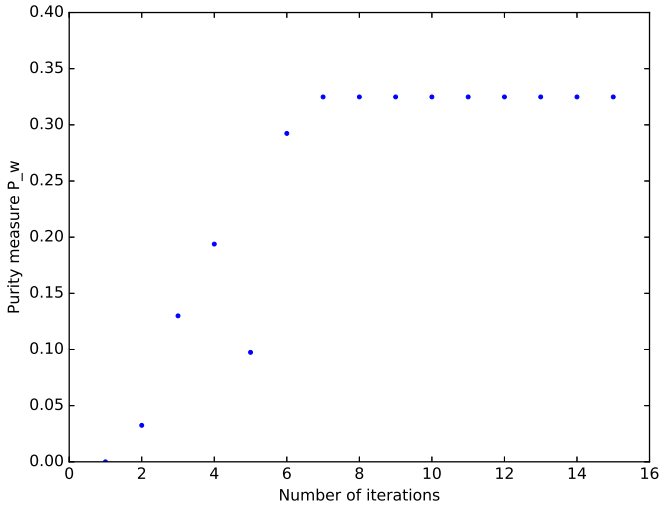

Figure 1: Average of purity measure with respect to number of iterations.

Problem 2: Soft Clustering

Redo Problem 1 using, instead of the hard clustering algorithm k-means, the soft clustering algorithm discussed in Lecture 9 (cf. slide 35 in `https://version.aalto.fi/gitlab/junga1/MLBP2017Public/blob/master/Clustering/mlbp17_Clustering_v1.pdf`). We run this soft clustering algorithm for a fixed number M of iterations to obtain, for each snapshot, the degree $y^{(i)}$ to which the ith snapshot belongs to $\mathcal{C}_1$. A reasonable adaption of the purity measure of Problem 1 to the soft clustering setting is to use[1] is $P_{\rm w} = h\big((2/N)\sum_{i\in\mathbb{X}^{(\rm winter)}} y^{(i)}\big)$ and $P_{\rm s} = h\big((2/N)\sum_{i\in\mathbb{X}^{(\rm summer)}} y^{(i)}\big)$ for computing the average purity $\bar{P} = (1/2)(P_{\rm w} + P_{\rm s})$. Implement the soft clustering algorithm using different numbers M of iterations and plot the average purity $\bar{P}(M)$ as a function of the number of iterations M. For each choice of M, use two (independently) randomly selected feature vectors $\mathbf{x}^{(i)}, \mathbf{x}^{(j)} \in \mathbb{X}$ as the initial cluster means $\mathbf{m}_0$ and $\mathbf{m}_1$. Initialize the covariance matrices with the identity matrix, i.e., $\mathbf{C}_0 = \mathbf{C}_1 = \mathbf{I}$. Average the results obtained from these runs to get one single estimate of $\bar{P}(M)$ for each M.

Answer.

Similar calculations as in previous section were performed, but instead of using hard clusterin a soft clusterin algorithm was used. The resulting plot showing the average purity measure as a function of number of iterations is shown in figure 2. According to figure 2 the cluster means do not proberly converge to any definite value with different initialization and thus the impurity measure fluctuates between different number of iterations used.

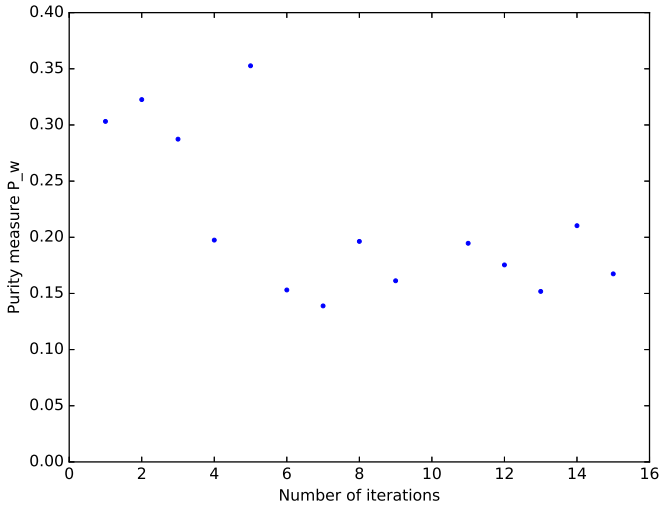

Figure 2: Average of purity measure with respect to number of iterations.

[1]For an index $i \in \{1,\ldots,N\}$, with a slight abuse of notation, we write $i \in \mathbb{X}^{(\rm winter)}$ if the ith feature vector $\mathbf{x}^{(i)}$ represents a winter image, i.e., $\mathbf{x}^{(i)} \in \mathbb{X}^{(\rm winter)}$.