EE378B Statistical Signal Processing

K-Means

Andrea Montanari *Lecture 1-4 - Due on 1/23/2017*

The K-Means algorithm

K-Means is a simple and popular greedy algorithm for data clustering. Nowadays it is most commonly used in conjunction with spectral, or convex optimization methods, as a rounding procedure. For instance, spectral methods are used to construct a low-dimensional embedding and K-Means is then applied to construct the clusters.

The input to K-Means are items $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathcal{X}$ (most often $\mathcal{X} = \mathbb{R}^d$, but the items can also be images, customers, and so on). The output is a partition of these items in k groups, or equivalently a labeling $\boldsymbol{\sigma} \in [k]^n$. For instance, $\sigma_i = 3$ means that item i belongs to cluster 3.

K-Means attempts at at minimizing the following cost function over $\boldsymbol{\sigma}$ and $\boldsymbol{c} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_k)$, $\boldsymbol{c}_\ell \in \mathcal{X}$:

$$\mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{c}) \equiv \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{x}_i, \boldsymbol{c}_{\sigma(i)}) \, . \tag{1}$$

Here $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a measure of distance between items (but does not need to be a distance function) and $\boldsymbol{c}_\sigma$ can be interpreted as the center of cluster σ. The cost function $\mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{c})$ is the average distance between datapoints and the center of their cluster. Notice that this depends implicitly on the data $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$: this dependence is implicit.

In many cases, it is simple to minimize $\mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{c})$ with respect to $\boldsymbol{\sigma}$ (for $\boldsymbol{c}$ given) or with respect to $\boldsymbol{c}$ (for $\boldsymbol{\sigma}$ given). The most classical form of K-Means is Batch K-Means and alternates between these two minimization steps. Here, with a slight abuse of notation, we write $\mathcal{L}(\boldsymbol{\sigma}) \equiv \min_{\boldsymbol{c}} \mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{c})$.

Batch K-Means

Input : Data $\{\boldsymbol{x}_i\}$, initial partition $\boldsymbol{\sigma}^0$

Output : New partition $\boldsymbol{\sigma}^{\text{new}}$

1: $\boldsymbol{\sigma}^{\text{new}} := \boldsymbol{\sigma}^0$;

2: Do

3: $\quad \boldsymbol{\sigma} := \boldsymbol{\sigma}^{\text{new}}$;

4: $\quad$ For $j \in \{1, \dots, k\}$, compute center $\boldsymbol{c}_j$:

$$\boldsymbol{c}_j = \arg\min_{\boldsymbol{c} \in \mathcal{X}} \sum_{i \in [n] : \sigma_i = j} \ell(\boldsymbol{x}_i, \boldsymbol{c})$$

5: $\quad$ For $i \in \{1, \dots, n\}$, compute new label

$$\sigma_i^{\text{new}} = \arg\min_{\sigma \in [k]} \ell(\boldsymbol{x}_i, \boldsymbol{c}_\sigma)$$

6: While $\mathcal{L}(\boldsymbol{\sigma}^{\text{new}}) \le \mathcal{L}(\boldsymbol{\sigma}) - \varepsilon$

7: Return $\boldsymbol{\sigma}^{\text{new}}$

In words, we compute centers of the current clusters, and then update the labels assigning each data point to the closest center. Step 5 is efficient and requires only nk evaluation of the metric $\ell(\,\cdot\,,\,\cdot\,)$. Step 4 is also efficient for many definition of the function $\ell(\,\cdot\,,\,\cdot\,)$. Normally this iteration is initialized with random labels $\boldsymbol{\sigma}^0$.

In this homework, we will consider the most standard example, namely $\boldsymbol{x}_i \in \mathcal{X} = \mathbb{R}^d$ and $\ell(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. In these case, the center of a cluster is just the empirical mean. In other words, step 4 is implemented by

$$\boldsymbol{c}_j = \frac{1}{n_j} \sum_{i \in [n]: \sigma_i = j} \boldsymbol{x}_i \,. \tag{2}$$

where $n_j = |\{i : \sigma_i = j\}|$ (here and below $|S|$ denotes the cardinality of set S).

(*a*) Program the BATCH K-MEANS for $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ and $d(x, y) = \|x - y\|_2^2$.

Test your program on a synthetic dataset, by taking $n = 10^4$, $d = 10$, $k = 3$ and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ i.i.d. with density

$$\rho(\boldsymbol{x}) = \frac{1}{3}\phi(\boldsymbol{x} - s\boldsymbol{e}_1) + \frac{1}{3}\phi(x - s\boldsymbol{e}_2) + \frac{1}{3}\phi(x - s\boldsymbol{e}_3) \,, \tag{3}$$

where $\phi(\boldsymbol{z}) = (2\pi)^{-d/2} \exp\{-\|\boldsymbol{x}\|_2^2/2\}$ is the standard Gaussian density, and $\boldsymbol{e}_i$ is the i-th standard basis vector.

Run 10 trials (with different realizations of the data $\boldsymbol{x}$) for each of $s \in \{0.5, 1, 1.5, \dots, 10\}$, For each value of s, report the mean fraction of misclassified points.

[Note that the three clusters can only be identified up to a permutation of the labels. Propose a suitable definition of misclassification error that overcomes this problem.]

(*b*) Let $\overline{X} \in \mathbb{R}^{n \times d}$ be the matrix whose rows are $\boldsymbol{x}_1^\mathsf{T}, \dots \boldsymbol{x}_n^\mathsf{T}$. Show that (always using the loss function $\ell(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$)

$$\mathcal{L}(\boldsymbol{\sigma}) = \frac{1}{n}\|\overline{X}\|_F^2 - \frac{1}{n}\sum_{a=1}^{k} \frac{1}{n_a} \|\overline{X}^\mathsf{T} \mathbf{1}_{\sigma^{-1}(a)}\|_2^2 \,. \tag{4}$$

where $\sigma^{-1}(a) = \{i \in [n] : \sigma_i = a\}$ and, as above, $n_a = |\sigma^{-1}(a)|$.

Deduce that

$$\min_{\boldsymbol{\sigma}, \mathbf{c}} \mathcal{L}(\boldsymbol{\sigma}, \boldsymbol{x}) \geq \frac{1}{n} \sum_{j=k+1}^{n \vee d} \sigma_j(\overline{X})^2 \,. \tag{5}$$

Compare this lower bound with the value achieved by BATCH K-MEANS in the simulations in point (*a*) above.

(*c*) The dataset `seeds_dataset.txt` from the UC Irvine Machine Learning repository contains measurements of geometrical properties of kernels belonging to three different varieties of wheat. Data are provided for $n = 207$ seeds. Each row contains $d = 7$ attributes and (as last entry) the seed type (a label between 1 and 3.

Use BATCH K-MEANS to cluster the 207 data points in $k = 3$ clusters. Do you think it is useful to pre-process the data in some way? Re-run BATCH K-MEANS with 10 random initializations and report, each time, the value of objective $\mathcal{L}(\cdots)$ achieved, as well as the number of misclassified data-points. (As in point (*a*), mis-classification error has to be defined up to a permutation of the 3 labels.) Also, compare the value of the objective achieved with the lower bound at point (*b*).

Remark Paper submissions are preferred. You can submit the homework in class, or in the box of the second floor in Packard (there is a box for EE378B). You can use any programming languages. Tables and figures are preferred to help to illustrate your results. There are several sub-questions for each question. Make sure that it is easy for readers to figure out which sub-questions you are answering.