EPFL, Autumn 2024
Yoav Zemel
Guillaume Obozinski

Ho Yun
Shivang Sachar

Statistical Machine Learning

Exercise sheet 11

Exercise 11.1 (K-means as alternating minimization) In class, we have seen that the problem that K-means tries to optimize is

$$\min_{(\boldsymbol{\mu}_k)_k} \quad \sum_{i=1}^{n} \min_{1\le k\le K} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

(a) Show that the optimization problem of K-means is equivalent to solving the following optimization problem

$$\begin{aligned} \min_{(\boldsymbol{\mu}_k)_k,(z_{ik})_{i,k}} \quad & \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \\ \text{s.t.} \quad & z_{ik} \in \{0,1\},\ \forall i,k \\ & \sum_{k=1}^{K} z_{ik} = 1,\ \forall i. \end{aligned}$$

In particular, prove that the partial minimization w.r.t. all z_{ik} recovers the objective of K-means from the slides.

Solution: We can focus on the optimization problem

$$\begin{aligned} \min_{(z_{ik})_k} \quad & \sum_{k=1}^{K} z_{ik}\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \\ \text{s.t.} \quad & z_{ik} \in \{0,1\},\ \forall k \\ & \sum_{k=1}^{K} z_{ik} = 1. \end{aligned}$$

where is clear that the minimizing $\mathbf{z}_i$ picks the value of k such that $\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$ is minimum. The value of the objective after minimizing w.r.t. $\mathbf{z}_i$ is thus equal to $\min_{1\le k\le K}\|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$.

(b) Prove that if we let $C_k = \{i \mid z_{ik} = 1\}$, then minimizing w.r.t. all z_{ik} for fixed $\boldsymbol{\mu}_k$s corresponds to the cluster update step in K-means; symmetrically, show that minimizing w.r.t. all $\boldsymbol{\mu}_k$ for fixed z_{ik} produces the centroid update step in K-means. After minimizing w.r.t. to the z_{ik}s , the set $\{i \mid z_{ik} = 1\}$ is exactly the set of indices of the points that are closer to $\boldsymbol{\mu}_k$ than any other centroid. On the other hand minimizing w.r.t. to $\boldsymbol{\mu}_k$ yields

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} z_{ik}\mathbf{x}_i}{\sum_{i=1}^{n} z_{ik}} = \frac{1}{|C_k|}\sum_{i\in C_k}\mathbf{x}_i,$$

which is indeed the centroid update.

(c) Deduce from the last question that K-means can be interpreted as an alternating optimization algorithm.

If we alternatingly minimize w.r.t. all the zs and then w.r.t. to all the $\boldsymbol{\mu}_k$, we recovers the K-means algorithm

Exercise 11.2 (Variance decomposition in clustering) In this exercise, we consider a situation in which data living in $\mathbb{R}^p$ has been partitioned in a number of clusters, and where the clusters centroids are set to be the empirical means (or barycenters) of the data in each cluster. The goal of the exercise is to show that there is a nice relationship between the (co)variance of the data in each cluster, the total (co)variance of the data, and the (co)variance of the centroids. More precisely, we shall show that the total (co)variance is the sum of these two (co)variances.

Let $\{\mathbf{x}_i\}_{i=1}^N$ denote i.i.d. samples of a $\mathbb{R}^d$-valued random variable X. Let $\overline{\mathbf{x}}$ and $\widehat{\Sigma}$ denote the empirical mean and empirical covariance of the sample, respectively. For each $k = 1, \dots, K$, let $\widehat{\Sigma}_k$ denote the empirical covariance matrix of the kth cluster and $\widehat{\pi}_k$ denote the proportion of the sample in the kth cluster.

(a) Show that

$$\widehat{\Sigma} = \sum_{k=1}^{K} \widehat{\pi}_k \widehat{\Sigma}_k + \sum_{k=1}^{K} \widehat{\pi}_k \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right] \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right]^\top$$

Solution: We proceed as follows:

$$\begin{aligned}
\widehat{\Sigma} &= \tfrac{1}{n} \sum_{i=1}^{n} \left[\mathbf{x}_i - \overline{\mathbf{x}}\right] \left[\mathbf{x}_i - \overline{\mathbf{x}}\right]^\top \\
&= \tfrac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \left[\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right] \left[\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right]^\top \\
&= \tfrac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \left[\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k\right] \left[\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k\right]^\top + \tfrac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right] \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right]^\top \\
&+ \tfrac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \left(\left[\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k\right] \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right]^\top + \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right] \left[\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k\right]^\top\right) \\
&= \tfrac{1}{n} \sum_{k=1}^{K} |C_k| \widehat{\Sigma}_k + \tfrac{1}{n} \sum_{k=1}^{K} |C_k| \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right] \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right]^\top + 0 \\
&= \sum_{k=1}^{K} \widehat{\pi}_k \widehat{\Sigma}_k + \sum_{k=1}^{K} \widehat{\pi}_k \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right] \left[\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right]^\top
\end{aligned}$$

since $\sum_{i \in C_k} \left[\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k\right] = 0$ and $\widehat{\pi}_k = |C_k|/n$.

(b) Show that,

$$\tfrac{1}{n} \sum_{i=1}^{n} \left\|\mathbf{x}_i - \overline{\mathbf{x}}\right\|^2 = \sum_{k=1}^{K} \widehat{\pi}_k \left[\sum_{i \in C_k} \left\|\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k\right\|^2\right] + \sum_{k=1}^{K} \widehat{\pi}_k \left\|\widehat{\boldsymbol{\mu}}_k - \overline{\mathbf{x}}\right\|^2$$

Solution: Simply take the trace of the equation in (a).

(c) Why can $\widehat{\Sigma}_k$ and $\frac{1}{|C_k|}\sum_{i\in C_k}\|\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k\|^2$ be considered as a measure of *intra-cluster variance* of the kth cluster?

Solution: By definition, $\widehat{\Sigma}_k$ is the covariance of the data if we only have the data from the kth cluster and $\frac{1}{|C_k|}\sum_{i\in C_k}\|\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k\|^2$ is the average quadratic error made by approximating $\mathbf{x}_i$ with the cluster center $\widehat{\boldsymbol{\mu}}_k$. The latter is also the sum of the variances $\widehat{\sigma}^2_{k,j}=\frac{1}{|C_k|}\sum_{i\in C_k}(x_{ij}-\mu_{kj})^2$ of all coordinates inside of the kth cluster.

(d) Why can $\sum_{k=1}^K\widehat{\pi}_k\left[\widehat{\boldsymbol{\mu}}_k-\overline{\mathbf{x}}\right]\left[\widehat{\boldsymbol{\mu}}_k-\overline{\mathbf{x}}\right]^\top$ and $\sum_{k=1}^K\widehat{\pi}_k\|\widehat{\boldsymbol{\mu}}_k-\overline{\mathbf{x}}\|^2$ be considered as measures of *inter-cluster variance*? What do they represent?

Solution: If we were to draw each $\widehat{\boldsymbol{\mu}}_k$ with probability $\widehat{\pi}_k$, that is in proportion to the data contained in the kth cluster, then the expectation of the resulting distribution would be $\overline{\mathbf{x}}=\sum_{k=1}^K\widehat{\pi}_k\widehat{\boldsymbol{\mu}}_k$ and therefore covariance matrix of the resulting distribution would be $\sum_{k=1}^K\widehat{\pi}_k\left[\widehat{\boldsymbol{\mu}}_k-\overline{\mathbf{x}}\right]\left[\widehat{\boldsymbol{\mu}}_k-\overline{\mathbf{x}}\right]^\top$. And $\sum_{k=1}^K\widehat{\pi}_k\|\widehat{\boldsymbol{\mu}}_k-\overline{\mathbf{x}}\|^2$ is the average quadratic error made by replacing $\widehat{\boldsymbol{\mu}}_k$ by $\overline{\mathbf{x}}$.

(e) Explain how the problem of clustering can be thought of as that of grouping the data in such a way that the barycenters of the clusters are as spread out as possible in space.

Solution: In clustering, $C_k=\{i\mid\|\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k\|=\min_j\|\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_j\|\}$.

Since the total variance is fixed, it follows that minimizing the *intra-cluster variances* $\sum_{k=1}^K\widehat{\pi}_k\left[\sum_{i\in C_k}\|\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k\|^2\right]=\frac{1}{n}\sum_{i=1}^n\min_k\|\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k\|^2$ is equivalent to maximizing the *inter-cluster variance* $\sum_{k=1}^K\widehat{\pi}_k\|\widehat{\boldsymbol{\mu}}_k-\overline{\mathbf{x}}\|^2$, and in this sense, the barycenters $\widehat{\boldsymbol{\mu}}_k$ should be as spread out in space as possible.

(f) How are the results in (a) and (b) related to the variance decomposition formulas:

$$\operatorname{Var}[Y]=\mathbb{E}[\operatorname{Var}[Y|Z]]+\operatorname{Var}[\mathbb{E}[Y|Z]]$$

for a scalar-valued random variable Y and

$$\operatorname{Cov}[Y]=\mathbb{E}[\operatorname{Cov}[Y|Z]]+\operatorname{Cov}[\mathbb{E}[Y|Z]]$$

for a vector-valued random variable Y.

Solution: If $Y=X$ and Z is the cluster index of X and moreover, the covariances and expectations are evaluated according to the empirical distribution of X (not the population distribution) then the above formula returns the results of (a) and (b) depending on whether X is vector or scalar-valued respectively.

Here is a simple proof of the variance decomposition formula:

$$\begin{aligned}\operatorname{Cov}[Y]&=\mathbb{E}[(Y-\mathbb{E}[Y|Z]+\mathbb{E}[Y|Z]-\mathbb{E}[Y])(Y-\mathbb{E}[Y|Z]+\mathbb{E}[Y|Z]-\mathbb{E}[Y])^\top]\\&=\mathbb{E}[(Y-\mathbb{E}[Y|Z])(Y-\mathbb{E}[Y|Z])^\top]+\mathbb{E}[(\mathbb{E}[Y|Z]-\mathbb{E}[Y])(\mathbb{E}[Y|Z]-\mathbb{E}[Y])^\top]\\&\quad+2\mathbb{E}[(Y-\mathbb{E}[Y|Z])(\mathbb{E}[Y|Z]-\mathbb{E}[Y])^\top]\end{aligned}$$

Now, notice that the first and the second terms can be written as:

$$\begin{aligned}\mathbb{E}[(Y-\mathbb{E}[Y|Z])(Y-\mathbb{E}[Y|Z])^\top]&=\mathbb{E}\left[\mathbb{E}[(Y-\mathbb{E}[Y|Z])(Y-\mathbb{E}[Y|Z])^\top|Z]\right]\\&=\mathbb{E}[\operatorname{Cov}(Y|Z)]\\\mathbb{E}[(\mathbb{E}[Y|Z]-\mathbb{E}[Y])(\mathbb{E}[Y|Z]-\mathbb{E}[Y])^\top]&=\operatorname{Cov}[\mathbb{E}[Y|Z]]\end{aligned}$$

And the third term is zero:

$$\begin{aligned}\mathbb{E}[(Y - \mathbb{E}[Y|Z])(\mathbb{E}[Y|Z] - \mathbb{E}[Y])^\top] &= \mathbb{E}\left[\mathbb{E}[(Y - \mathbb{E}[Y|Z])(\mathbb{E}[Y|Z] - \mathbb{E}[Y])^\top|Z]\right] \\ &= \mathbb{E}\left[(\mathbb{E}[Y|Z] - \mathbb{E}[Y|Z])(\mathbb{E}[Y|Z] - \mathbb{E}[Y])^\top\right] = 0\end{aligned}$$

Exercise 11.3 (Properties of the EM algorithm) Prove the following properties of the EM algorithm. We will use the same notation as the lecture.

(a) Show that $\mathcal{L}(q, \theta) = \log p(x; \theta) - KL(q(z) \,\|\, p(z \,|\, x; \theta))$.

This is straightforward to show with the definition of the Kullback-Leibler divergence.

$$\begin{aligned}\mathcal{L}(q, \theta) &= \int q(z) \log \tfrac{p(x,z;\theta)}{q(z)} \mathrm{d}z \\ &= \int q(z) \log \tfrac{p(z \,|\, x;\theta) p(x;\theta)}{q(z)} \\ &= \log p(x;\, \theta) \int q(z) \mathrm{d}z + \int q(z) \log \tfrac{p(z \,|\, x;\theta)}{q(z)} \mathrm{d}z \\ &= \log p(x;\, \theta) - KL(q(z) \,\|\, p(z \,|\, x;\, \theta))\end{aligned}$$

(b) In the E step of the EM algorithm, show that $\log p(x;\, \theta^{t-1}) = \mathcal{L}(q, \theta^{t-1})$ when $q(z) = p(z \,|\, x;\, \theta^{t-1})$.

When $q(z) = p(z \,|\, x;\, \theta^{t-1})$, $KL(q(z) \,\|\, p(z \,|\, x;\, \theta^{t-1})) = 0$ and so $\log p(x;\, \theta^{t-1}) = \mathcal{L}(q, \theta^{t-1})$. In other words, the E step brings the variational bound $\mathcal{L}$ to "touch" the likelihood for $\theta = \theta^{t-1}$.

(c) Show that the EM algorithm never decreases the likelihood.

We have

$$\log p(x; \theta^{(k-1)}) = \mathcal{L}(q^{(k)}, \theta^{(k-1)}) \leq \mathcal{L}(q^{(k)}, \theta^{(k)}) \leq \log p(x; \theta^{(k)}).$$

Indeed,

- The first equality holds because of the E step, which brings the variational bound $\mathcal{L}$ to the likelihood.
- The second inequality holds because of the M step, where $\mathcal{L}(q^{(k)}, \cdot)$ is maximized with respect to θ.
- Finally, the last inequality holds because of the main result shown in class using Jensen's inequality, i.e. that $\mathcal{L}(q^{(k)}, \theta) \leq \log p(x; \theta)$ is true for any θ (this can also be viewed as a consequence of the non-negativity of the Kullback-Leibler divergence in the previous question).

Exercise 11.4 (K-means as isotropic Gaussian mixtures with zero variance)

(a) Using results from the lecture slides, derive the form of the EM algorithm for a Gaussian mixture model in $\mathbb{R}^p$ with K classes and with equal covariance matrices $\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_K = \sigma^2 \mathbf{I}_q$ for the K classes (you don't need to do all calculations, just try to understand the formulas).

Solution: The expectation step is given by:

$$q_{ik}^{(t)} = \frac{\pi_k^{(t-1)} \exp(-\|\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t-1)}\|^2 / 2\sigma^2)}{\sum_{j=1}^{K} \pi_j^{(t-1)} \exp(-\|\mathbf{x}^{(i)} - \boldsymbol{\mu}_j^{(t-1)}\|^2 / 2\sigma^2)}$$

and the maximization step is given by:

$$\begin{aligned}
\boldsymbol{\mu}_k^{(t)} &= \frac{\sum_i \mathbf{x}^{(i)} q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}} \\
\Sigma_k^{(t)} &= \frac{\sum_i [\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}][\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}]^\top q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}} \\
\pi_k^{(t)} &= \frac{\sum_i q_{ik}^{(t)}}{\sum_{i,k'} q_{ik'}^{(t)}}
\end{aligned}$$

(b) Show that the K-means algorithm is a limiting case of the EM algorithm of question (a) when $\sigma^2 \to 0$.

Comparison of the K-means algorithm with the EM algorithm for Gaussian mixtures shows that there is a close similarity. Whereas the K-means algorithm performs a hard assignment of data points to clusters, in which each observation is associated uniquely with one cluster, the EM algorithm makes a soft assignment based on the posterior probabilities (also known as the responsibilities).

In the setup of (a), we have that

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2\right\},$$

so that for a particular point $\boldsymbol{x}_i$, the responsibilities are given by

$$q_{ik} = \frac{\pi_k \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2\right\}}{\sum_j \pi_j \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|^2\right\}}.$$

If we consider the limit $\sigma^2 \to 0$, we see that in the denominator the term for which $\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2$ is smallest will go to zero most slowly, and hence the responsibilities q_{ik} for the data point $\boldsymbol{x}_i$ all go to zero except for term j, for which the responsibility q_{ij} will go to unity. Note that this holds independently of the values of the π_k, so long as none of the π_k is zero. Thus, in this limit, we obtain a hard assignment of data points to clusters, just as in the K-means algorithm, so that each data point is assigned to the cluster having the closest mean.