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.

(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.

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

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

(b) Show that,

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

(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?

(d) Why can $\sum_{k=1}^{K}\widehat{\pi}_k\,[\widehat{\boldsymbol{\mu}}_k-\overline{\mathbf{x}}]\,[\widehat{\boldsymbol{\mu}}_k-\overline{\mathbf{x}}]^\top$ 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?

(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.

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

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

for a scalar-valued random variable Y and

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

for a vector-valued random variable Y.

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

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

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

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 $\boldsymbol{\Sigma}_1=\cdots=\boldsymbol{\Sigma}_K=\sigma^2\mathbf{I}_q$ for the K classes (you don't need to do all calculations, just try to understand the formulas).

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