HW4: K-Means and Gaussian Mixture Models

Last modified: 2023-03-29 21:48

Status: **RELEASED.**

Due date: Thu Apr 06 at 11:59pm ET

How to turn in: Submit PDF to https://www.gradescope.com/courses/496674/assignments/2782332

Jump to: Problem 1 Problem 2 Problem 3 Problem 4

Questions?: Post to the **hw4** topic on the Piazza discussion forums.

Instructions for Preparing your PDF Report

What to turn in: PDF of typeset answers via LaTeX. No handwritten solutions will be accepted, so that grading can be speedy and you get prompt feedback.

Please use provided LaTeX Template: https://github.com/tufts-ml-courses/cs136-23s-assignments/blob/main/unit4_HW/hw4_template.tex

Your PDF should include (in order):

- Cover page with your full name, estimate of hours spent, and Collaboration statement
- Problem 1a, 1b, 1c, 1d, 1e
- Problem 2a, 2b
- Problem 3a
- Problem 4a is OPTIONAL. Worth up to 6 points back on other parts (cannot go higher than 100%).

When you turn in the PDF to gradescope, mark each part via the in-browser Gradescope annotation tool)

Problem 1: K-means walk-through

Recall that K-means minimizes the following cost function:

$$J(x_{1:N}, r_{1:N}, \mu_{1:K}) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}(x_n - \mu_k)^T (x_n - \mu_k)$$

where each assignment variable r_n is a one-hot vector of size K.

The K-means algorithm is specified in pseudocode as:

Inputs:

- $x_1, \dots x_N$: Training dataset
- $\mu^0_{1:K}$: Initial guess of cluster center locations

Procedure:

For iter t in 1, 2, ... until converged:

1. $r^t_{1:N} \leftarrow \arg\min_{r_{1:N}} J(x_{1:N}, r_{1:N}, \mu^{t-1}_{1:K})$
2. $\mu^t_{1:K} \leftarrow \arg\min_{\mu_{1:K}} J(x_{1:N}, r^t_{1:N}, \mu_{1:K})$

Consider running K-means on the following dataset of N=7 examples, in this (N, D)-shaped array

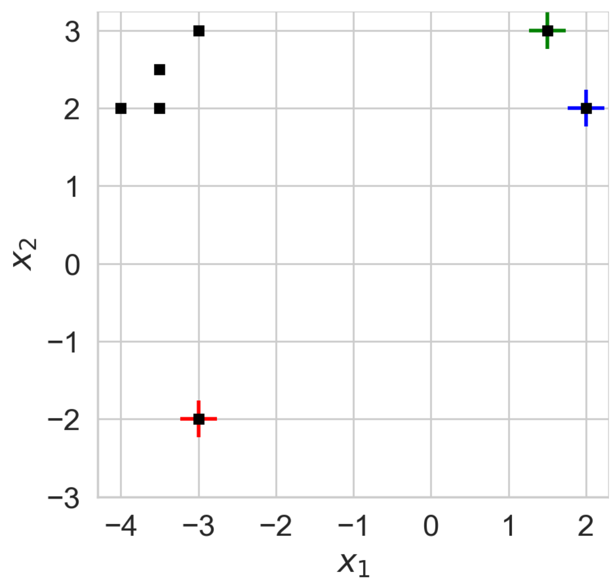
```
x_ND = array([
  [-3.0,-2.0],
  [-4.0, 2.0],
  [-3.5, 2.5],
  [-3.5, 2.0],
  [-3.0, 3.0],
  [ 1.5, 3.0],
  [ 2.0, 2.0]])
```

When the initial cluster locations are given by the following K=3 cluster locations, in this (K, D)-shaped array:

```
mu_KD = array([
  [-3.0,-2.0],
  [ 1.5, 3.0],
  [ 2.0, 2.0]])
```

We'll denote these initial locations mathematically as μ^0. Here and below, we'll use *superscripts* to indicate the specific *iteration* of the algorithm at which we ask for the value, and we'll assume that iteration 0 corresponds to the initial configuration.

We've visualized the 7 data examples (black squares) and the 3 initial cluster locations (crosses) in this figure:



Plot of the N=7 toy data examples (squares) and K=3 initial cluster locations (crosses).

Problem 1a: Find the optimal one-hot assignment vectors r^1 for all $N = 7$ examples, when given the initial cluster locations μ^0. This corresponds to executing step 1 of K-means algorithm. Report the value of the cost function $J(x, r^1, \mu^0)$.

Problem 1b: Find the optimal cluster locations μ^1 for all $K = 3$ clusters, using the optimal assignments r^1 you found in 2a. This corresponds to executing step 2 of K-means algorithm. Report the value of the cost function $J(x, r^1, \mu^1)$.

Problem 1c: Find the optimal one-hot assignment vectors r^2 for all $N = 7$ examples, using the cluster locations μ^1 from 2b. Report the value of the cost function $J(x, r^2, \mu^1)$.

Problem 1d: Find the optimal cluster locations μ^2 for all $K = 3$ clusters, using the optimal assignments r^2 you found in 2c. Report the value of the cost function $J(x, r^2, \mu^2)$.

Problem 1e: What interesting phenomenon do you see happening in this example regarding cluster 2? How could you set cluster 2's location after part d above to better fulfill the goals of K-means (find K clusters that reduce cost the most)?

Problem 2: Relationship between GMM and K-means

Bishop's PRML textbook Sec. 9.3.2 describes a technical argument for how a GMM can be related to the K-means algorithm. In this problem, we'll try to make this argument concrete for the same toy dataset as in Problem 1.

To begin, given *any* GMM parameters, we can use Bishop PRML Eq. 9.23 to compute the *posterior probability* of assigning each example x_n to cluster k via the formula:

$$\begin{aligned}\gamma_{nk} &\triangleq p(z_{nk} = 1 | x_n) \\ &= \frac{\pi_k \quad (x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \quad (x_n | \mu_j, \Sigma_j)}\end{aligned}$$

Now, imagine a GMM with the following *concrete* parameters:

- mixture weights $\pi_{1:K}$ set to the uniform distribution over $K = 3$ clusters
- covariances $\Sigma_{1:K}$ set to ϵI_D for all clusters, for some $\epsilon > 0$

We can leave the locations $\mu_{1:K}$ at any valid values.

Problem 2a: Show (with math) that using the parameter settings defined above, the general formula for γ_{nk} will simplify to the following (inspired by PRML Eq. 9.42):

$$\gamma_{nk} = \frac{\exp(-\frac{1}{2\epsilon}(x_n - \mu_k)^T(x_n - \mu_k))}{\sum_{j=1}^{K} \exp(-\frac{1}{2\epsilon}(x_n - \mu_k)^T(x_n - \mu_k))}$$

Problem 2b: What will happen to the vector γ_n as $\epsilon \rightarrow 0$? How is this related to K-means?

Hint 2(i): Try it out concretely on the toy data from Problem 1 above.

Hint 2(ii): No need for a formal proof here. Just show you understand what happens in the limit, not why.

Problem 3: Covariances of mixtures

Background: Consider a continuous random variable x which is a vector in D-dimensional space: $x \in \mathbb{R}^D$.

We assume that x follows a mixture distribution with PDF p^{mix}, using K components indexed by integer k:

$$p^{mix}(x|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k f_k(x|\mu_k, \Sigma_k)$$

The k-th component has a mixture "weight" probability of π_k. Across all K components, we have a parameter $\pi = [\pi_1 \; \pi_2 \; \ldots \; \pi_K]$, whose entries are non-negative and sum to one.

The k-th component has a specific data-generating PDF f_k. We don't know the functional form of this PDF (it could be Gaussian, or something else), and the form could be different for every k. However, we do know that this PDF f_k takes two parameters, a vector $\mu_k \in \mathbb{R}^D$ and a matrix Σ_k which is a $D \times D$ symmetric, positive definite matrix. We further know that these parameters represent the mean and covariance of vector x under the pdf f_k:

$$\begin{aligned} \mathbb{E}_{(f_k(x|\mu_k,\Sigma_k))}[x] &= \mu_k \\ \mathrm{Cov}_{(f_k(x|\mu_k,\Sigma_k))}[x] &= \Sigma_k \end{aligned}$$

Problem 3a: Prove that the covariance of vector x under the mixture distribution is given by:

$$\mathrm{Cov}_{p^{mix}(x)}[x] = \sum_{k=1}^{K} \pi_k(\Sigma_k + \mu_k\mu_k^T) - mm^T$$

where we define $m = \mathbb{E}_{p^{mix}(x)}[x]$.

Hint 3(i): We know a closed-form for m: $m = \sum_{k=1}^{K} \pi_k \mu_k$.

Hint 3(ii): For any random vector x*, we know*: $\mathbb{E}[xx^T] = \mathrm{Cov}(x) + \mathbb{E}[x]\mathbb{E}[x]^T$

Problem 4: Jensen's Inequality and KL Divergence

Optional. Not required.

Background reading

Skim Bishop PRML's Sec. 1.6 ("Information Theory"), which introduces several key concepts useful for the EM algorithm, including:

- Entropy
- Jensen's inequality
- KL divergence

Background: Negative logarithms are convex

Consider the negative logarithm function: $f(a) = -\log a$, for inputs $a > 0$. Recall that $f(a)$ is a *convex* function, because its second derivative is always positive:

$$\begin{aligned} f(a) &= -\log a, \\ f'(a) &= -a^{-1}, \\ f''(a) &= a^{-2}, \qquad \text{therefore: } f''(a) > 0 \text{ for all } a > 0. \end{aligned}$$

Background: Jensen's inequality for negative logarithms

Now, suppose we have a random variable A that takes one of K possible values.

Define each candidate value $a_k > 0$, and let its associated probability be $r_k \in [0, 1]$. Writing the probabilities as a vector $\mathbf{r} = [r_1, \dots r_K]$, we know these non-negative values must sum to one: $\mathbf{r} \in \Delta^K$.

We are interested in the expected value of $f(A)$, where f is the negative logarithm. We can derive the following bound using **Jensen's inequality** (see PRML textbook Eq. 1.115),

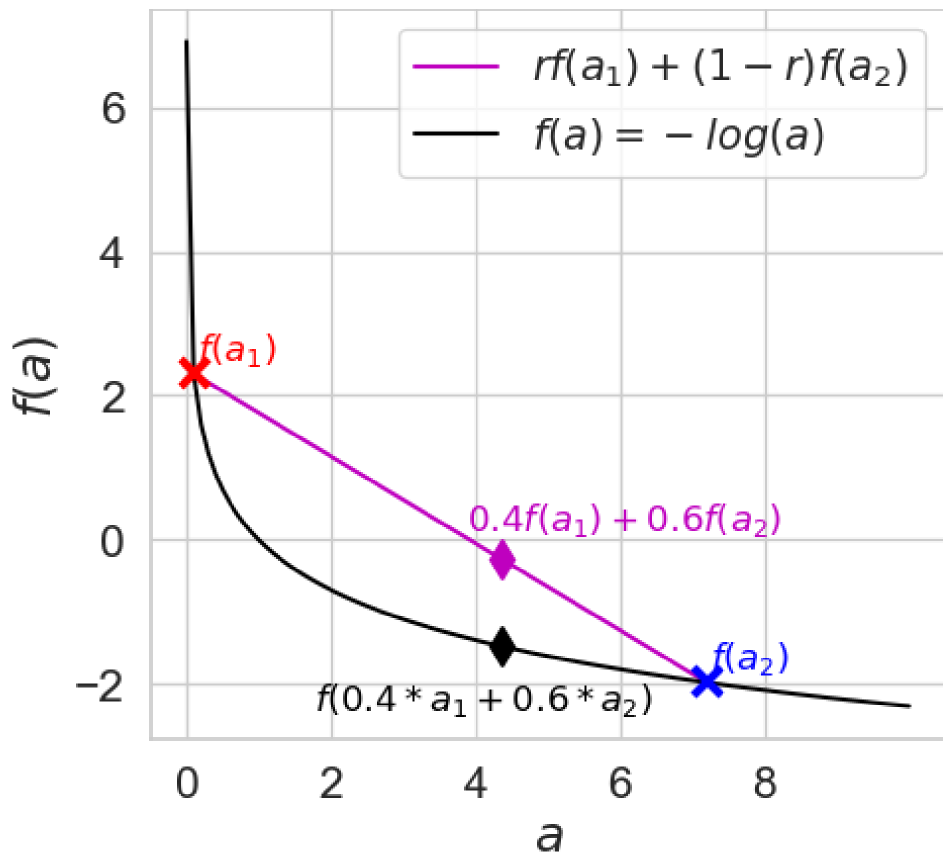
$$\mathbb{E}[f(A)] \geq f(\mathbb{E}[A])$$

Expanding out these expectations and invoking f's definition as the negative log (which only takes positive inputs), we have:

$$\sum_{k=1}^{K} r_k[-\log a_k] \geq -\log \left[\sum_{k=1}^{K} r_k a_k \right]$$

This bound holds for *any* positive vector $\mathbf{a}$ and any probability vector $\mathbf{r}$.

We can visualize this Jensen bound in the following figure, using two selected points $a_1 = 0.1$ and $a_2 = 7.2$.



Plot of our convex function of interest ("f", black) and its *linear interpolation* (magenta) between outputs that correspond to two inputs "a1" and "a2". Clearly, function f is a *lower bound* of its interpolation (magenta). In terms of probabilities, this means \$\mathbb{E}[f(A)] \geq f(\mathbb{E}[A])\$

Notation setup

We'll use one-hot indicator vectors here. Let e_k denote the one-hot vector of size K where entry k is non-zero.

Define random variable z as a one-hot indicator vector of size K. So, the K possible values of z are $\{e_1, e_2, \dots e_K\}$.

Define two possible Categorical distributions over z, denoted q and p.

$$q(z) = \text{CatPMF}(z|r_1, \dots r_K) = \prod_{k=1}^{K} r_k^{z_k}, \qquad q(z = e_k) = r_k, \qquad r_k > 0 \ \forall k$$

$$p(z) = \text{CatPMF}(z|\pi_1, \dots \pi_K) = \prod_{k=1}^{K} \pi_k^{z_k}, \qquad p(z = e_k) = \pi_k, \qquad \pi_k > 0 \ \forall k$$

Each uses an *all positive* probability vector parameters $\mathbf{r} \in \Delta_+^K$ and $\pi \in \Delta_+^K$. Here Δ_+^K denotes the set of K-length vectors whose sum is one and whose entries are all *strictly positive*.

The KL divergence from q to p is defined as:

$$\text{KL}(q(z)||p(z)) \triangleq \mathbb{E}_{q(z)}[-\log \frac{p(z)}{q(z)}]$$

Problem statement

Problem 4a: Consider any two Categorical distributions $q(z)$ and $p(z)$ that assign *positive* probabilities over the same size-K sample space. Show that their KL divergence is non-negative.

That is, show that $KL(q(z)||p(z)) \geq 0$, or equivalently that

$$KL\left(\mathrm{CatPMF}(z|\mathbf{r})||\mathrm{CatPMF}(z|\pi)\right) \geq 0$$

when $\mathbf{r} \in \Delta_+^K$ and $\pi \in \Delta_+^K$.

Hint: Expand the definition of KL as an expectation out so it is purely an evaluatable function of r and π, then use Jensen's inequality for negative logarithms.

Note: it is possible to prove the KL is non-negative even when some entries in r or π are exactly zero, but this requires taking some limits rather carefully, and we want you to avoid that burdensome detail. Thus, here we consider r and π as having all positive entries.