

# **6.790 Notes**

**Lecturer:**

ANDREW LIU

Fall 2023

Last updated on Saturday 28<sup>th</sup> October, 2023.

# Contents

<b>1</b>	<b>September 8, 2023</b>	<b>3</b>
<b>2</b>	<b>October 27, 2023</b>	<b>3</b>
2.1	K-Means Clustering . . . . .	3
2.2	PCA Intuition . . . . .	3
2.3	PCA with Encoder/Decoder framework . . . . .	4
2.4	Probabilistic PCA . . . . .	6
2.5	Gaussian Mixture Model Intuition . . . . .	6
2.6	Arbitrarily Bad Local Maxima in GMMs . . . . .	6

# 1 September 8, 2023

# 2 October 27, 2023

## 2.1 K-Means Clustering

Place  $k$  centroids at random  $c_1, \dots, c_k$ . Then, repeat until convergence:

1. For each  $x_i$ , locate the nearest centroid  $c_j = \arg \min_{j=1}^k \mathcal{D}(x_i, c_j)$  for some distance metric  $\mathcal{D}$ .
2. For each cluster  $c_j$ , update the position  $c_j = \sum_{x_i \in c_j} x_i / n$ .

Using  $\mathcal{D} = \|\cdot\|_2^2$ , the objective function being minimized here is

$$C = \sum_{j=1}^K \sum_{x_i \in c_j} \|x_i - c_j\|_2^2$$

There are other ways to choose the centroids. For example, the  $k$ -medioids algorithm places the center of each cluster at

$$x_m = \arg \min_{x \in \{x_1, \dots, x_n\}} \sum_{i=1}^n \|x_i - x\|^2.$$

This is better for outliers.

## 2.2 PCA Intuition

Say we have some data about football players and their different attributes: speed, scoring, height, weight. If we were to plot this data on a graph with each axis representing an attribute, we know that these axes would not all be orthogonal; for example, height and weight are strongly correlated. The idea of PCA is to try to decouple the data space into orthogonal vectors; then, we can more easily select the vectors that “matter”, which allows us to reduce the dimensionality of the data while keeping its most important aspects.

## 2.3 PCA with Encoder/Decoder framework

We have  $x_1, \dots, x_n$  as  $d$ -dimensional data in  $\mathbb{R}^d$ . In PCA, we want to reduce the dimensionality to  $m < d$ , which we can do with an encoder matrix  $W \in \mathbb{R}^{m \times d}$ . We can also use a decoder matrix  $U \in \mathbb{R}^{d \times m}$  to try to retrieve the original data from its low-dimensional representation. Our goal is to minimize the distance between the approximated output  $x' = UWx$ , and its original vector, i.e., we want to find the “best” encoder/decoder pair

$$W^*, U^* = \underset{W \in \mathbb{R}^{m \times d}, U \in \mathbb{R}^{d \times m}}{\operatorname{argmin}} \sum_{i=1}^n \|x_i - UWx_i\|_2^2.$$

First, fix  $(U, W)$  and consider the set of all condensed representations  $R = \{UWx : x \in \mathbb{R}^d\}$ . This is an  $m$ -dimensional linear subspace of  $\mathbb{R}^d$ , and there exists some  $V \in \mathbb{R}^{d \times m}$  whose columns form an orthonormal basis of  $R$ . So, we may use  $V^T V = I_m$  (but not the other way around).

For all  $x_i$ , there exists  $y_i \in \mathbb{R}^m$  s.t.  $Vy_i = UWx_i$ , so our objective function becomes

$$\begin{aligned} \sum_{i=1}^n \|x_i - Vy_i\|_2^2 &= \sum_{i=1}^n (x_i - Vy_i)^T (x_i - Vy_i) \\ &= \sum_{i=1}^n x_i^T x_i + y_i^T V^T V y_i - 2y_i^T (V^T x_i) \\ &= \sum_{i=1}^n \|x_i\|^2 + \|y_i\|^2 - 2y_i^T (V^T x_i) \end{aligned}$$

since  $V$  is orthonormal. Differentiating each summed term, we have

$$0 = \frac{d}{dy_i} (\|x_i\|^2 + \|y_i\|^2 - 2y_i^T (V^T x_i)) = -2V^T x_i + 2y_i \implies y_i = V^T x_i.$$

This bound is achieved when  $U = V, W = V^T$ , so we can now reformulate our

objective:

$$\begin{aligned}
& \arg \min_{U \in \mathbb{R}^{d \times m}, U^T U = I_m} \|x_i - U U^T x_i\|^2 \\
&= \arg \min_{U \in \mathbb{R}^{d \times m}, U^T U = I_m} x_i^T x_i + x_i^T U U^T U U^T x_i - 2x_i^T U U^T x_i \\
&= \arg \max_{U \in \mathbb{R}^{d \times m}, U^T U = I_m} x_i^T U U^T x_i
\end{aligned}$$

Recall that this expression is not the same as  $x_i^T x_i$ , since we can't say that  $V V^T$  is the identity. Finally, summing over all elements,

$$U^* = \arg \max_{U \in \mathbb{R}^{d \times m}, U^T U = I_m} \text{Tr} \left( U^T \sum_{i=1}^n x_i x_i^T U \right)$$

Define  $S = \sum_{i=1}^n x_i x_i^T \in \mathbb{R}^{d \times d}$ ; we want to maximize  $\text{Tr}(U^T S U)$ .  $S$  is positive semi-definite and symmetric, so by the spectral decomposition theorem we have  $S = P \Lambda P^T$ , where  $\Lambda$  is diagonal with elements the eigenvalues of  $S$  and  $P$  is the corresponding orthonormal eigenbasis. Choose  $P, \Lambda$  so that  $\Lambda_{i,i} \geq \Lambda_{j,j}$  for all  $i \leq j$ , i.e., order eigenvalues decreasing from left to right.

Now,

$$\text{Tr}(U^T S U) = \text{Tr}(U^T P \Lambda P^T U) = \text{Tr}((P^T U)^T \Lambda (P^T U)).$$

Denoting  $B = P^T U$ , then

$$\text{Tr}(U^T S U) = \sum_{i=1}^m \sum_{j=1}^d B_{ji}^2 \Lambda_{j,j} = \sum_{i=1}^d \Lambda_{i,i} \sum_{j=1}^m B_{ij}^2 \leq \sum_{i=1}^m \Lambda_{i,i}.$$

This bound is tight when  $U$  has columns equal to the first  $m$  eigenvectors in the eigenbasis of  $S$ , since these leads to  $B$  with identity values along the diagonal. More explicitly,

$$B_{ij} = \sum_{k=1}^d P_{ik}^T U_{kj} = P_i \cdot P_j,$$

where  $P_i$  is the  $i$ th eigenvector in  $S$ , i.e., the  $i$ th column of  $P$ . By orthonormality, this is 0 when  $i \neq j$  and 1 when  $i = j$ , so  $B = I_{d \times m}$ . Finally, plugging this construction into the above trace formula gives the exact bound, so this works. This shows that the optimal  $U^*, W^*$  is given by the eigenbasis of  $S$ , and the corresponding error for

this construction is

$$\text{Tr} \left( \sum_{i=1}^n x_i x_i^T + U^T \sum_{i=1}^n x_i x_i^T U \right) = \sum_{i=1}^d \Lambda_{i,i} - \sum_{i=1}^m \Lambda_{i,i} = \sum_{i=m+1}^d \Lambda_{i,i}.$$

## 2.4 Probabilistic PCA

We have observation  $x \in \mathbb{R}^d$  generated in latent space by some  $z \in \mathbb{R}^m$ ,  $m < d$ , via

$$x = Wz + \mu + \varepsilon,$$

where  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$  is noise.  $\mu$  represents the “mean” latent space representation of  $x$ . If we let  $z \sim \mathcal{N}(0, I)$ , then  $Wz \sim \mathcal{N}(0, WW^T)$ , and  $x \sim \mathcal{N}(\mu, WW^T + \sigma^2 I)$ . Thus, the log likelihood of a full dataset is

$$NLL(\mathcal{X}|\mu, W, \sigma^2) = -\frac{nd}{2} \ln 2\pi - \frac{n}{2} \ln \det \Sigma - \frac{1}{2} \sum_{i=1}^n (x_n - \mu)^T \Sigma^{-1} (x_n - \mu).$$

The maximum likelihood PCA satisfies

$$0 = \frac{\partial NLL(\mathcal{X}|\mu, W, \sigma^2)}{\partial \mu} = \sum_{i=1}^n \Sigma^{-1} (x_n - \mu) \implies \mu = \bar{x}.$$

## 2.5 Gaussian Mixture Model Intuition

Gaussian mixture models are an extension of  $K$ -means clustering. In these models, the model is parameters by some  $M$  components, each of which generates a distribution of points according to some model, e.g., multivariate normal. The likelihood of the model can then be evaluated according to some priors placed on how likely each point is to be in each of the  $M$  components, and then a max likelihood can be solved for using techniques like gradient ascent. We will show in the next problem that this unfortunately does not always work as we expect it to.

## 2.6 Arbitrarily Bad Local Maxima in GMMs

(This problem was sourced from Chi Jin, et. al. 2016).

Consider a GMM with  $M = 3$  on the number line and true centers generated by  $\mu_1^* = -R$ ,  $\mu_2^* = R$ , and  $\mu_3^* = \gamma R$ , where  $R > 0$  and  $\gamma \gg 1$ , and all with variance 1. The

idea here is for  $\mu_1^*, \mu_2^*$  to be close to the origin and  $\mu_3^*$  to be very far away.

Assuming uniform prior on each of the three clusters, the likelihood of a model  $\mu = (\mu_1, \mu_2, \mu_3)$  on  $n$  observations is given by

$$\mathcal{L}_n(\mu) = \frac{1}{n} \sum_{i=1}^n \log \left( \frac{1}{3} \sum_{j=1}^3 \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu_j)^2/2} \right).$$

With infinite data, we can take the loss as an expected value over all possible data:

$$\begin{aligned} \mathcal{L}(\mu) &= \mathbb{E}_x \left[ \log \left( \sum_{i=1}^3 e^{-(x - \mu_i)^2/2} \right) - \log(3\sqrt{2\pi}) \right] \\ &= \sum_{V \in \{A, B, C\}} \frac{1}{3} \mathbb{E}_V \left[ \log \left( \sum_{i=1}^3 e^{-(V - \mu_i)^2/2} \right) \right] - \log(3\sqrt{2\pi}), \end{aligned}$$

where the expected value is decomposed (by the law of total expectation) based on the assumption of uniform priors between the three clusters.  $A$ ,  $B$ , and  $C$  are random variables that follow the true generating distribution  $\mu^*$ , since this is what we should observe with infinite data.

### Example 2.1

Consider the set of points  $\mathcal{D} = \{(\mu_1, \mu_2, \mu_3) \in \mathbb{R}^3 \mid \mu_1 \leq \gamma R/3, \mu_2 \geq 2\gamma R/3, \mu_3 \geq 2\gamma R/3\}$ . In the limit as  $\gamma \rightarrow \infty$ , compute:

- the loss on the three borders of this region, i.e., when  $\mu_1 = \gamma R/3$ , when  $\mu_2 = 2\gamma R/3$ , or when  $\mu_3 = 2\gamma R/3$ .
- the loss on the model  $\tilde{\mu} = (0, \gamma R, \gamma R)$ .

Starting with  $\tilde{\mu}$ , first expand the entire definition:

$$\begin{aligned} \mathcal{L}(\tilde{\mu}) &= \frac{1}{3} \mathbb{E}_A \left[ \log \left( e^{-(A-0)^2/2} + e^{-(A-\gamma R)^2/2} + e^{-(A-\gamma R)^2/2} \right) \right] \\ &\quad + \frac{1}{3} \mathbb{E}_B \left[ \log \left( e^{-(B-0)^2/2} + e^{-(B-\gamma R)^2/2} + e^{-(B-\gamma R)^2/2} \right) \right] \\ &\quad + \frac{1}{3} \mathbb{E}_C \left[ \log \left( e^{-(C-0)^2/2} + e^{-(C-\gamma R)^2/2} + e^{-(C-\gamma R)^2/2} \right) \right] - \log(3\sqrt{2\pi}). \end{aligned}$$

Taking the limit,

$$\lim_{\gamma \rightarrow \infty} \mathcal{L}(\tilde{\mu}) = \frac{1}{3} \mathbb{E}_A [\log(e^{-A^2/2})] + \frac{1}{3} \mathbb{E}_B [\log(e^{-B^2/2})] + \frac{1}{3} \mathbb{E}_C [\log(2e^{-(C-\gamma R)^2/2})] - \log(3\sqrt{2\pi}).$$

Doing a bit more algebra,

$$\begin{aligned} \mathbb{E}[A^2] &= \text{Var}[A] + \mathbb{E}[A]^2 = 1 + R^2 \\ \mathbb{E}[B^2] &= \text{Var}[B] + \mathbb{E}[B]^2 = 1 + R^2 \\ \mathbb{E}[(C - \gamma R)^2] &= \mathbb{E}[C^2 - 2C\gamma R + \gamma^2 R^2] = 1, \end{aligned}$$

so

$$\lim_{\gamma \rightarrow \infty} \mathcal{L}(\tilde{\mu}) = -\frac{2R^2 + 3 - 2\log 2}{6} - \log(3\sqrt{2\pi}).$$

Repeating the same algebra for the borders, it can be shown that the loss is strictly less on the borders than it is for  $\tilde{\mu}$ . This is significant because it shows that there is a local maxima in  $\mathcal{D}$ .

### Example 2.2

Show that this local maximum can be arbitrarily bad compared to the global maximum, i.e., that the log likelihood can be arbitrarily worse in the limit  $R \rightarrow \infty$ .

Repeating the same calculation as above, we can compute the global loss by plugging in  $\mu^* = (-R, R, \gamma R)$ :

$$\lim_{R \rightarrow \infty} \mathcal{L}(\mu^*) = \frac{1}{3} \left( -(A+R)^2/2 - (B-R)^2/2 - (C-\gamma R)^2/2 \right) - \log(3\sqrt{2\pi}),$$

where

$$\mathbb{E}[(A+R)^2] = \mathbb{E}[(B-R)^2] = \mathbb{E}[(C-\gamma R)^2] = 1,$$

so

$$\lim_{R \rightarrow \infty} \mathcal{L}(\mu^*) = -\frac{1}{2} - \log(3\sqrt{2\pi}).$$

We can show that  $\lim_{R \rightarrow \infty} \mathcal{L}(\mu') = -\infty$  as follows.  $\mathcal{L}(\mu')$  is bounded above by the  $\mu_2$  term, since  $A$  and  $C$  grow arbitrarily far from the center  $R$ ; but, the contribution of loss from the  $\mu_2$  term is  $-\infty$  since all three of  $\mu'_1, \mu'_2, \mu'_3$  grow arbitrarily far from



center. Therefore, in the limit, global loss is constant, while our local maximum can achieve arbitrarily bad loss. Since loss is continuous in  $R$ , this shows that we can grow the gap between our local and global maxima arbitrarily large by increasing  $R$ . The significance of this result is that gradient ascent does not always work as expected!