INF554 Machine Learning I

Lab 3: Unsupervised Classification, Gaussian Mixture Models

Notations

In the following, capital letters always designate random variables A. Tiny letters designate realizations a. Bold tiny letters designate the set of realizations $\boldsymbol{a} = (a_1, \dots, a_N)$. "cst" always designates a contextual constant. Finally, to simplify we are going to abuse the $\mathbb{P}$ notations. $\mathbb{P}(a)$ will designate $\mathbb{P}(A = a)$, $\mathbb{P}(\boldsymbol{a})$ designates $\mathbb{P}((A_1 = a_1, \dots A_N = a_N))$, and in the case of contuous distributions $\mathbb{P}(a)$ designates the density of A evaluated at a.

1 K-Means

Let $\boldsymbol{x}$ be a set of N data points. The **K-Means** algorithm aims at clustering those points into K clusters. One of the particularity of this algorithm is that it is model free, there is no probabilistic assumption on the data. The K-Means objective can ve written:

$$\underset{(S_1, \dots, S_K)}{\arg\min} \sum_{k=1}^{K} \sum_{x \in S_k} \|x - \mu_k\|^2 \tag{1}$$

Idea: Given $\boldsymbol{x}$ iteratiely update the means vectors $(\mu_1, \dots, \mu_K)$ and the clusters $(S_1, \dots, S_K)$.
Let's denote $n_k = \#S_k$.

Algorithm 1: K-Means

Data: $(x_1, \dots, x_N)$
Result: $(\mu_1, \dots, \mu_K), (S_1, \dots, S_K)$
Initialize: $(\mu_{0,1}, \dots, \mu_{0,K})$
while *An update is made* **do**
  Assignment: $S_{tk} = \{x_i / k = \arg\min_l \|x_i = \mu_l\|^2\}$
  Update: $\mu_{t,k+1} = \frac{1}{n_k} \sum_{x \in S_{tk}} x$
end

Selecting the best k is finding the "natural" number of clusters in the data. Most of methods aim at evaluating the quality of the proposed clustering. We are going to investigate one of them. Since k-means searchs for clusters that minizes the intra-clusters variances, the evolution of this objective with respect to the number of clusters feels like a good indicator. Let,

$$V_K = \sum_{k=1}^{K} V_{kK}, \text{ where } V_{kK} = \sum_{x \in S_k} \|x - \mu_k\|^2$$

Question 1
(Bonus) Show that $\mathbb{V}(\boldsymbol{x}) = \frac{1}{N}V_K + \mathbb{V}(\boldsymbol{\mu})$, with $\boldsymbol{\mu} = (\mu_k)_{k=1,\ldots,K}$ i.e. the algorithm aims at maximizing the part of the variance explained by clusters.

Question 2
Based on this observation propose a heuristic to find the best number of clusters

Question 3
Think of some situations where k-means will fail to identify satisfying clusters.

Task 1
(Bonus) Implement the k-means algorithm in Python.

Task 2
(Bonus) Implement the Elbow method.

2 Gaussian Mixture Models

2.1 Preliminaries

For this section, let X be a random variable from a parametrized family, and let θ be the associated parameter. The goal is to estimate θ with respect to N observations $\boldsymbol{x} = (x_1, \ldots, x_N)$.

2.1.1 Likelihood function

To evaluate the parameters we want to maximize the **likelihood** corresponding to this problen.

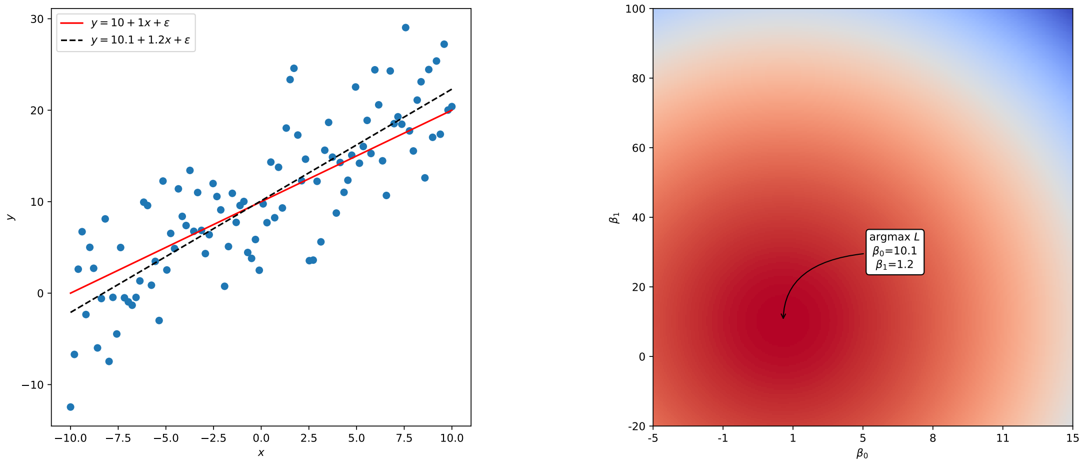
$$L(\boldsymbol{x}; \theta) = \mathbb{P}(\boldsymbol{x}|\theta) \tag{2}$$

This function gives the probability of this dataset being generated given parameters. It can be viewed as a confidence in the parameter. The higher the likelihood, the more probable it is that such parameter under such statistical model, generated this dataset. Then the θ can be estimated:

$$\hat{\theta} = \arg\max_{\theta} L(\boldsymbol{x}; \theta) \tag{3}$$

Example: Linear Regression
Let's consider the linear regression model: $Y = \beta_0 + \beta_1 X + \epsilon$ where $\epsilon \sim \mathcal{N}(0,1)$. Here $\theta = (\beta_0, \beta_1)$. Note that in this model, X is supposed to be given.

$$\begin{align}
L(\boldsymbol{y}, \boldsymbol{x}; \theta) &= P(\boldsymbol{y}|\boldsymbol{x}, \theta) \tag{4}\\
&= \prod_{i=1}^{N} P(y_i|x_i, \theta) \tag{5}\\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp[y_i - \beta_0 - \beta_1 x_i] \tag{6}\\
&\propto \prod_{i=1}^{N} \exp\left[\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2}\right] \tag{7}\\
\log L(\boldsymbol{y}, \boldsymbol{x}; \theta) &= \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 + \text{cst} \tag{8}
\end{align}$$



a: Data

b: log-likelihood

Figure 1: Example of likelihood maximazation on a linear regression, true model $y = 10 + x + \epsilon$.

Where cst is independant of θ. For computation purposes it is easier to use the **log-likelihood** $l(\boldsymbol{y}, \boldsymbol{x}; \theta)$. Then:

$$\nabla_\theta l(\boldsymbol{y}, \boldsymbol{x}; \theta) = 0 \Leftrightarrow \begin{cases} \frac{\partial l(\boldsymbol{y}, \boldsymbol{x}; \theta)}{\partial \beta_1} \propto \sum_{i=1}^{N} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial l(\boldsymbol{y}, \boldsymbol{x}; \theta)}{\partial \beta_0} \propto \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\beta}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\boldsymbol{x}})(y_i - \bar{\boldsymbol{y}})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\boldsymbol{x}})^2} \\ \hat{\beta}_0 = \bar{\boldsymbol{y}} - \hat{\beta}_1 \bar{\boldsymbol{x}} \end{cases} \tag{9}$$

Figure 1 shows a visualization of the log-likelihood function in the case of a linear regression model.

2.1.2 Expectation-Maximization

Now let Z be an unobserved, or hidden, variable. This variable could be hidden from us, the interest (like in GMM), or artificially added to make the likelihood tracktable. If there exists an hypothesis on the distribution of $Z|\theta$, this new information can be used to better maximize the likelihood.

Idea: If θ is fixed, it is possible to generate Z according to $Z|\theta$. Conversly, if Z was available, θ would be easier to infer. With this in mind, at each iteration t, one could sample from $Z|\theta_t$, update θ accordingly, then ressample at $t+1$, and so on and so on.... In fact, sampling from $Z|\theta$ is not needed, only the knowledge on its distribution is sufficient. Instead of sampling from $Z|\theta_t$, more information is used when computing $\Delta(\theta, \theta_t) \equiv \mathbb{E}_{Z|\boldsymbol{x}, \theta_t} l(\boldsymbol{x}, Z; \theta)$ and then choosing θ_{t+1} that maximizes $\Delta(\theta, \theta_t)$.

For more detail about the construction and convergence of this algorithm refer to Borman, 2004

Algorithm 2: Expectation-Maximization

Data: $(x_1, \ldots, x_N)$
Result: θ
Initialize: θ_0
while $t < T$ **do**
  Expectation: $\Delta(\theta, \theta_t) = \mathbb{E}_{Z|\boldsymbol{x}, \theta_t} \ln \mathbb{P}(\boldsymbol{x}, Z|\theta)$
  Maximization: $\theta_{t+1} = \arg\max_\theta \Delta(\theta, \theta_t)$
end

2.2 The Gaussian Mixture Model

The goal is to cluster N data points $\boldsymbol{x} = (x_1, \dots, x_N), x_i \in \mathbb{R}^d$ into K clusters. The **Gaussian Mixture** model assumes that each data point was generated by a gaussian distribution corresponding to it's cluster. We thus assume the following generation rule:

$$X|Z, \theta \sim \mathcal{N}(\mu_k, \Sigma_k) \tag{10}$$

$$\begin{aligned} &Z|\theta \sim p(\pi_i, \dots, \pi_K) \\ &\quad \text{s.a.} \quad \mathbb{P}(Z_j = k|\theta) = \pi_j \end{aligned} \tag{11}$$

Where Z is an <u>hidden variable</u>, and $\theta = (\mu_1 \dots, \mu_K, \Sigma_1, \dots, \Sigma_K, \pi_1, \dots, \pi_K)$. Z represents the clusters, and takes is realisations in $(1, \dots, K)$.

2.2.1 EM in the case of Gaussian Mixture Models

Recall that the density of a Gaussian distribution with d dimensions, of mean μ, and variance matrix Σ is:

$$f(x|\mu, \Sigma) = \frac{1}{\sqrt{2\pi}^d \det(\Sigma)^{1/2}} \exp\Big[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big] \tag{12}$$

Estimation Step

Question 4

Show that

$$l(\boldsymbol{x}, Z; \theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}\{Z_i = k\}\Big[\ln \pi_k + \ln f(x_i|\mu_k, \Sigma_k)\Big] \tag{13}$$

Hint: Start by decomposing $\mathbb{P}(\boldsymbol{x}, Z|\theta)$ using the conditional probability and compute each element independently.

Question 5

Using independance of x_i's, Bayes theorem and the law of total probabilities, show that:

$$\mathbb{P}(Z_i = k|\boldsymbol{x}, \theta) = \gamma_{ik}(\theta) \equiv \frac{\pi_k f(x_i|\mu_k.\Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} f(x_i|\mu_{k'}.\Sigma_{k'})} \tag{14}$$

Recall: Bayes theorem $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$, Law of total probabilities $\mathbb{P}(A) = \mathbb{E}_B \mathbb{P}(A|B)$

Question 6

Using the results of the two previous questions, show that

$$\Delta(\theta, \theta_t) = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}(\theta_t)\Big[\ln \pi_k + \ln f(x_i|\mu_k, \Sigma_k)\Big] \tag{15}$$

Hint: expectation commutes with finite summations

Maximization Step

Now that the closed form formula of $\Delta(\theta, \theta_t)$ is known, the program we are trying to solve is:

$$\theta_{t+1} = \arg\max_{\theta} \Delta(\theta, \theta_t) \tag{16}$$

$$\text{s.t.} \sum_{k=1}^{K} \pi_k = 1 \tag{17}$$

The Lagrangian associated with this problem is $\mathcal{L}(\theta, \lambda) \equiv \Delta(\theta, \theta_t) - \lambda(\sum_{k=1}^{K} \pi_k - 1)$. Solving the previous optimisation prblem is equivalent to solving:

$$\theta_{t+1} = \arg\min_{\theta,\lambda} \mathcal{L}(\theta, \lambda) \tag{18}$$

Question 7

(Bonus) Show that:

$$\frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \pi_k} = \frac{\sum_{i=1}^{N} \gamma_{ik}(\theta_t)}{\pi_k} - \lambda \tag{19}$$

$$\frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \lambda} = 0 \tag{20}$$

$$\frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \mu_k} \propto \sum_{i=1}^{N} \gamma_{ik}(\theta_t)^{-1}(x_i - \mu_k) \tag{21}$$

$$\frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \Sigma_k} \propto \sum_{i=1}^{N} \gamma_{ik}(\theta_t)\left[\Sigma_k^{-1} - \Sigma_k^{-1}(x_i - \mu_k)(x_i - \mu_k)^\top \Sigma_k^{-1}\right] \tag{22}$$

$$\tag{23}$$

Hint: When computing the gradient with respect to a variable, take out any constant, multipicative or additive.

Hint: $\frac{\partial \ln \det(\Sigma)}{\partial \Sigma} = (\Sigma^{-1})^\top$, *and* $\frac{\partial x^\top \Sigma^{-1} x}{\partial \Sigma} = -(\Sigma^{-1})^\top x x^\top (\Sigma^{-1})^\top$

(Bonus Hard) Prove the hints.

Question 8

Finally, show the following update rules:

$$\pi_{k,t+1} = \frac{\sum_{i=1}^{N} \gamma_{ik}(\theta_t)}{N} \tag{24}$$

$$\mu_{k,t+1} = \frac{\sum_{i=1}^{N} \gamma_{ik}(\theta_t) x_i}{\sum_{i=1}^{N} \gamma_{ik}(\theta_t)} \tag{25}$$

$$\Sigma_{k,t+1} = \frac{\sum_{i=1}^{N} \gamma_{ik}(\theta_t)(x_i - \mu_{k,t})(x_i - \mu_{k,t})^\top}{\sum_{i=1}^{N} \gamma_{ik}(\theta_t)} \tag{26}$$

$$\tag{27}$$

Hint: Start by showing that $\lambda = N$

We can now rewrite the EM-algorithm in the case of GMMs.

Algorithm 3: Expectation-Maximization for Gaussian Mixture Models

Data: $(x_1, \ldots, x_N)$
Result: θ
Initialize: θ_0
while $t < T$ **do**
  Expectation:
  for $k = 1, \ldots, K$ **do**
    $\gamma_{ik}(\theta) \equiv \frac{\pi_k f(x_i|\mu_k.\Sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} f(x_i|\mu_{k'}.\Sigma_{k'})}$
  end
  Maximization:
  for $k = 1, \ldots, K$ **do**
    $\pi_{k,t+1} = \frac{\sum_{i=1}^{N} \gamma_{ik}(\theta_t)}{N}$
    $\mu_{k,t+1} = \frac{\sum_{i=1}^{N} \gamma_{ik}(\theta_t) x_i}{\sum_{i=1}^{N} \gamma_{ik}(\theta_t)}$
    $\Sigma_{k,t+1} = \frac{\sum_{i=1}^{N} \gamma_{ik}(\theta_t)(x_i - \mu_{k,t})(x_i - \mu_{k,t})^\top}{\sum_{i=1}^{N} \gamma_{ik}(\theta_t)}$
  end
end

2.3 Application

Task 3
Follow the instructions of the notebook you were provided with

References

Borman, S. (2004). The expectation maximization algorithm a short tutorial.