Machine Learning Exercise Sheet 12

Clustering

Exercise sheets consist of two parts: In-class exercises and homework. The in-class exercises will be solved and discussed during the tutorial. The homework is for you to solve at home and further engage with the lecture content. There is no grade bonus and you do not have to upload any solutions. Note that the order of some exercises might have changed compared to last year's recordings.

In-class Exercises

K-Medians

Problem 1: Consider a modified version of the K-means objective, where we use L_1 distance instead.

$$\mathcal{J}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \boldsymbol{z}_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||_1$$

This variation of the algorithm is called *K-medians*. Derive the Lloyd's algorithm for this model.

Gaussian Mixture Model

Problem 2: Derive the E-step update for the Gaussian mixture model.

Problem 3: Derive the M-step update for the Gaussian mixture model.

Expectation Maximization Algorithm

Problem 4: Consider a mixture model where the components are given by independent Bernoulli variables. This is useful when modelling, e.g., binary images, where each of the D dimensions of the image $\boldsymbol{x}$ corresponds to a different pixel that is either black or white. More formally, we have

$$\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{z} = k) = \prod_{d=1}^{D} \boldsymbol{\theta}_{kd}^{x_d} (1 - \boldsymbol{\theta}_{kd})^{1 - x_d}.$$

That is, for a given mixture index $\boldsymbol{z} = k$, we have a product of independent Bernoullis, where $\boldsymbol{\theta}_{kd}$ denotes the Bernoulli parameter for component k at pixel d.

Derive the EM algorithm for the parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{kd} \mid k = 1, \ldots, K, d = 1, \ldots, D\}$ of a mixture of Bernoullis.

Assume here for simplicity, that the distribution of components $\mathrm{p}(\boldsymbol{z})$ is uniform: $\mathrm{p}(\boldsymbol{z}) = \prod_{k=1}^{K} \boldsymbol{\pi}_k^{z_k} = \prod_{k=1}^{K} \left(\frac{1}{K}\right)^{z_k}$.

P1 $J(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \|x_i - \mu_k\|_1$

$\min_{z} J(X, z, \mu)$ $\quad z_{ik} = \begin{cases} 1, & \text{if } k = \arg\min_j \|x_i - \mu_k\|_1 \\ 0, & \text{else} \end{cases}$

$\min_{\mu} J(X, z, \mu)$ $\quad J(X, z, \mu) = \sum_{i=1}^{N} z_{ik} \sum_{d=1}^{p} |x_{id} - \mu_{kd}|$

convex function

$$\frac{\partial |x_{id} - \mu_{kd}|}{\partial \mu_{kd}} = \begin{cases} \frac{\partial \mu_{kd} - x_{id}}{\partial \mu_{kd}} = 1 & \mu_{kd} > x_{id} \\ \frac{\partial x_{id} - \mu_{kd}}{\partial \mu_{kd}} = -1 & \mu_{kd} < x_{id} \\ 0 & \mu_{kd} = x_{id} \end{cases}$$

$$\frac{\partial J(X, Z, \mu)}{\partial \mu_{kd}} \propto \sum_{i=1}^{N} z_{ik} \cdot \mathbb{I}(\mu_{kd} > x_{id}) - \sum_{i=1}^{N} z_{ik} \mathbb{I}(\mu_{kd} < x_{id}) \doteq 0$$

$\frac{N_k}{2}$ points $=$ $\frac{N_k}{2}$ points

$N_k = \sum_{i=1}^{N} z_{ik}$

$\mu_{kd} =$ median $\{x_{id}$ such that $z_{ik} = 1\}$

P2 $\gamma_0(z) = p(z \mid X, \pi^{(0)}, \mu^{(0)}, \Sigma^{(0)})$

$$\gamma_t(z = z_{ik}) = p(z_{ik} \mid x_i, \pi^{(t)}, \mu^{(t)}, \Sigma^{(t)}) = \frac{p(x_i \mid z_{ik}, \mu^{(t)}, \Sigma^{(t)}) \cdot p(z_{ik} \mid \pi^{(t)})}{p(x_i \mid \pi^{(t)}, \mu^{(t)}, z^{(t)})}$$

$$= \frac{\pi_k^{(t)} \cdot \mathcal{N}(x_i \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum p(x_i, z_{ik} \mid \pi^{(t)}, \mu^{(t)}, \Sigma^{(t)})}$$

$$= \frac{\pi_k^{(t)} \cdot \mathcal{N}(x_i \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum p(x_i \mid z = z_{ik}, \pi^{(t)}, \mu^{(t)}, \Sigma^{(t)}) \cdot p(z = z_{ik} \mid \pi^{(t)})}$$

$$= \frac{\pi_k^{(t)} \mathcal{N}(x_i \mid \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_j^K \pi_j^{(t)} \mathcal{N}(x_i \mid \mu_j^{(t)}, \Sigma_j^{(t)})}$$

P3 $\mu^{t+1}, \Sigma^{t+1}, \pi^{t+1} = \underset{\mu \Sigma \pi}{\arg\max}\ E\left[\log p(x, z \mid \pi, \mu, \Sigma)\right]$

$$= \underset{\mu \Sigma \pi}{\arg\max} \sum \gamma_t(z) \cdot \log p(x, z \mid \pi, \mu, \Sigma)$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_t(z_i) \cdot \log p(x_i \mid z_i = k, \pi, \mu, \Sigma) \cdot p(z_i = k \mid \pi)$$

$$= \underbrace{\sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_t(z_i) \ln p(x_i \mid z_i = k, \pi, \mu, \Sigma)}_{\sim \mu, \Sigma} + \underbrace{\sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_t(z_i) \cdot \ln p(z_i = k \mid \pi)}_{\sim \pi}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_t(z_i) \cdot \ln \pi_k$$

$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$ μ_k Σ_k.

$$\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_t(z_i)\ln p(x_i|z_i=k,\pi,\mu,\Sigma)$$

$$=\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_t(z_i=k)\cdot\left[\ln\frac{1}{\sqrt{2\pi}}+\ln\frac{1}{\sqrt{|\Sigma_k|}}-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right]$$

$$\left[-\frac{1}{2}\ln 2\pi-\frac{1}{2}\ln|\Sigma_k|-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right]$$

$$(x_i^T\Sigma^{-1}-\mu_k^T\Sigma_k^{-1})(x_i-\mu_k)$$

$$\frac{\partial\mathcal{L}}{\partial\mu_k}=-\frac{1}{2}\sum_{i=1}^{N}\gamma_t(z_i=k)\left(-2x_i^T\Sigma_k^{-1}+2\Sigma_k^{-1}\mu_k\right)\quad x_i^T\Sigma^{-1}x_i-2x_i^T\Sigma_k^{-1}\mu_k+\mu_k^T\Sigma_k^{-1}\mu_k$$

$$=\sum_{i=1}^{N}\gamma_t(z_i=k)\Sigma_k^{-1}(x_i^T-\mu_k)\doteq 0.$$

$$\sum_{i=1}^{N}\gamma_t(z_i=k)\Sigma_k^{-1}x_i^T=N_k\Sigma_k^{-1}\mu_k$$

$$\mu_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma_t(z_i=k)x_i^T$$

$$\mathcal{L}=\sum_{k=1}^{K}N_k\cdot\ln\pi_k+\lambda\left(1-\sum_{i=1}^{K}\pi_k\right)$$

$$\frac{\partial\cdot\cdot}{\partial\pi_k}=\frac{N_k}{\pi_k}-\lambda\doteq 0$$

$$\pi_k=\frac{1}{\lambda}N_k$$

$$g(x)=\sum_{k=1}^{K}N_k\cdot\ln\frac{N_k}{\lambda}+\lambda\left(1-\sum_{k=1}^{K}\frac{N_k}{\lambda}\right)$$

$$=\sum_{k=1}^{K}N_k\cdot\ln\frac{N_k}{\lambda}+\lambda-N$$

$$\frac{\partial g}{\partial\lambda}=-\sum_{k=1}^{K}N_k\cdot\frac{1}{\lambda}+1\doteq 0.$$

$$=-\frac{N}{\lambda}=1.$$

$$\lambda=N\qquad\pi_k^{(t+1)}=\frac{N_k}{N}$$

$$\frac{\partial\mathcal{L}}{\partial\Sigma_k}=-\frac{1}{2}\sum_{i=1}^{N}\gamma_t(z_i=k)\left[\Sigma_k^{-T}+\left(-\Sigma_k^{-T}x_ix_i^T\Sigma^{-T}\right)-2\left(-\Sigma_k^{-T}\mu_k x_i^T\Sigma^{-T}\right)+\left(-\Sigma_k^{-T}\mu_k\mu_k^T\Sigma^{-T}\right)\right.$$

$$\Sigma_k^{-T}-\Sigma_k^{-T}\left[x_ix_i^T-2\mu_kx_i^T+\mu_k\mu_k^T\right]$$

$$\Sigma_k^{-T}-\Sigma_k^{-T}\left[(x_i-\mu_k)(x_i-\mu_k)^T\right]\Sigma_k^{-T}$$

$$=-\frac{1}{2}\left[N_k\cdot\Sigma_k^{-1}-\sum_{i=1}^{N}\gamma_t(z_i=k)\Sigma_k^{-1}\left[(x_i-\mu_k)(x_i-\mu_k)^T\right]\Sigma_k^{-T}\right]=0$$

$$N_k\cdot I=\sum_{i=1}^{N}\gamma(z_i=k)\Sigma_k^{-1}\left[(x_i-\mu_k)(x_i-\mu_k)^T\right]$$

$$\Sigma_k^{t+1}=\frac{1}{N_k}\sum_{i=1}^{N}\gamma(z_i=k)(x_i-\mu_k)(x_i-\mu_k)^T$$

P4 $p(x|z=k)=\prod_{d=1}^{D}\theta_{kd}^{x_d}(1-\theta_{kd})^{1-x_d}$.

$$p(z)=\prod_{k=1}^{K}\pi_k^{z_k}=\prod_{k=1}^{K}\left(\frac{1}{K}\right)^{z_k}.$$

① $$\gamma(z_{ik})=\frac{\left(\frac{1}{K}\right)^{z_k}\cdot\theta_{kd}^{x_d}(1-\theta_{kd})^{1-x_d}}{\sum_j^K\left(\frac{1}{K}\right)^{z_j}\theta_{jd}^{x_d}(1-\theta_{jd})^{1-x_d}}$$

② $$E(\log p(X,Z|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_t(z_i=k)\ln P(x_i|z_i=k,\theta)\cdot P(z_i=k)$$

$$=\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_t(z_i=k)\left[\sum_{d=1}^{D}x_{id}\ln\theta_{kd}+(1-x_d)\ln(1-\theta_{kd})+\ln\frac{1}{K}\right]$$

$$\frac{\partial\mathcal{L}}{\partial\theta}=\sum_{i=1}^{N}\gamma_t(z_i=k)\left[x_{id}\frac{1}{\theta_{kd}}-(1-x_{id})\cdot\frac{1}{1-\theta_{kd}}\right]=0$$

$$\left[\frac{x_{id}}{\theta_{kd}}-\frac{1}{1-\theta_{kd}}+\frac{x_{id}}{1-\theta_{kd}}\right]$$

$$\left[\frac{x_{id}-x_{id}\theta_{kd}-\theta_{kd}+x_{id}\theta_{kd}}{\theta_{kd}(1-\theta_{kd})}\right]$$

$$\sum_{i=1}^{N}\gamma_t(z_i=k)\frac{x_{id}}{\theta_{kd}(1-\theta_{kd})}=\sum_{i=1}^{N}\gamma_t(z_i=k)\frac{\theta_{kd}}{\theta_{kd}(1-\theta_{kd})}$$

$$\theta_{kd}=\frac{\sum_{i=1}^{N}\gamma_t(z_i=k)\,x_{id}}{\sum_{i=1}^{N}\gamma_t(z_i=k)}$$

Homework

Gaussian Mixture Model

Problem 5: Consider a mixture of K Gaussians

$$\mathrm{p}(\boldsymbol{x}) = \sum_k \boldsymbol{\pi}_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Derive the expected value $\mathbb{E}[\boldsymbol{x}]$ and the covariance $\mathrm{Cov}[\boldsymbol{x}]$.

Hint: it is helpful to remember the identity $\mathrm{Cov}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\,\mathbb{E}[\boldsymbol{x}]^T$.

Problem 6: Consider a mixture of K isotropic Gaussians, all with the same *known* covariances $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$.

Derive the EM algorithm for the case when $\sigma^2 \to 0$, and show that it's equivalent to Lloyd's algorithm for K-means.

Problem 7: Consider two random variables $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^D$ distributed according to two different Gaussian mixture models with $\boldsymbol{\theta}^x = \{\boldsymbol{\pi}^x, \boldsymbol{\mu}^x, \boldsymbol{\Sigma}^x\}$ and $\boldsymbol{\theta}^y = \{\boldsymbol{\pi}^y, \boldsymbol{\mu}^y, \boldsymbol{\Sigma}^y\}$, i.e.

$$\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\theta}^x) = \sum_{k=1}^{K_x} \boldsymbol{\pi}_k^x \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k^x, \boldsymbol{\Sigma}_k^x),$$
$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{\theta}^y) = \sum_{l=1}^{K_y} \boldsymbol{\pi}_l^y \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{\mu}_l^y, \boldsymbol{\Sigma}_l^y),$$

and the random variable $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$.

a) Describe a generative process (process of drawing samples) for $\boldsymbol{z}$.

b) Explain in a few sentences why $\mathrm{p}(\boldsymbol{z} \mid \boldsymbol{\theta}^x, \boldsymbol{\theta}^y)$ is again a mixture of Gaussians.

c) State the probability density function $\mathrm{p}(\boldsymbol{z} \mid \boldsymbol{\theta}^x, \boldsymbol{\theta}^y)$ of $\boldsymbol{z}$.

Problem 8:

a) Given is the dataset displayed in the figure below. Apply the K-means algorithm to this data using $K = 2$ and using the circled points as initial centroids.

What are the clusters after K-Means converges? Draw your solution in the figure above, i.e. mark the location of the centroids with $\times$'s and show the clusters by drawing two bounding boxes around the points assigned to each cluster.

How many iterations did it take for K-Means to converge in the above problem?

b) Provide a different initialization, for which the algorithm will take **more** iterations to converge to the **same** solution. Make sure that your initialization does not lead to ties. Circle the initial centroids in the figure below.

P5 $E(x) = E\left(E(x|z)\right) = \sum_{k=1}^{k} E(x|z) \cdot p(z) = \sum_{k=1}^{k} \mu_k \cdot \pi_k$.

$E(xx^T) = E\left(E(xx^T|z)\right) = \sum_{k=1}^{k} E(xx^T|z) \cdot p(z) =$

$= \sum_{k=1}^{k} \left(Cov(x|z) + E(x|z) E(x|z)^T \right)$

$= \sum_{k=1}^{k} \left(\Sigma_k + \mu_k \mu_k^T \right) \cdot \pi_k$.

$Cov(x) = E(xx^T) - E(x) \cdot E(x)^T$

$= \sum_{k=1}^{k} \left(\Sigma_k + \mu_k \mu_k^T \right) \pi_k - \sum_{l=1}^{k} \sum_{j=1}^{k} \mu_l \pi_l \mu_j^T \pi_j$

P6. $p(z_{ik} = 1 | x_i, \theta) = \dfrac{1}{\sum_l \frac{\pi_l}{\pi_k} \exp\left(\frac{-\|x_i - \mu_l\|^2 + \|x_i - \mu_k\|^2}{2\sigma^2} \right)}$

if μ_k is close to x_i, for $k = l$

$$\frac{-\|x_i - \mu_l\|^2 + \|x_i - \mu_k\|^2}{2\sigma^2} \leq 0$$

$k = l$ $\exp(\cdot) = 1$ $k \neq l$ exponential increasing negative

if μ_k isn't close to x_i $l \neq k$.

$$0 < \frac{-\|x_i - \mu_l\| + \|x_i - \mu_l\|^2}{2\sigma^2} \to \infty, \quad \sigma \to 0.$$

---

① Draw a sample x from $p(x|\theta^x)$ with the usual GMM

$\sim$ y

add them together

② $p(z|\theta^x, \theta^y) = \sum_{k=1}^{k_x} \sum_{l=1}^{k_y} \pi_k^x \pi_l^y \mathcal{N}(z | \mu_k^x + \mu_l^y, \Sigma_k^x + \Sigma_l^y)$

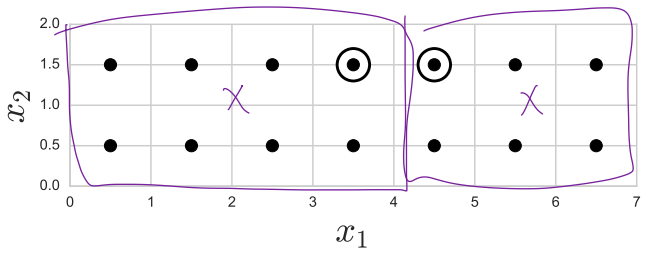

Figure 1: K-Means Dataset

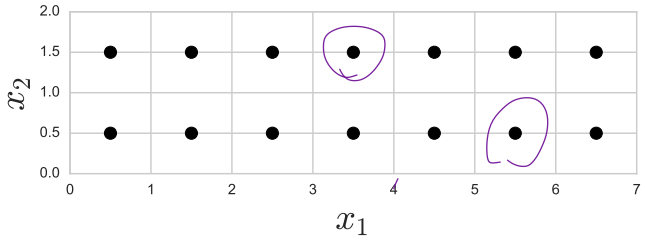

Figure 2: Provide your initialization

Problem 9: Download the notebook `exercise_12_clustering.ipynb` from Moodle. Fill in the missing code and run the notebook.