Fundamentals of Data Analytics
Exercise Sheet 12: K-means, GMMs and Linear Regression

University of Hamburg

Key Concepts

K-means Clustering

- **Algorithm Steps:**
  1. Initialize K cluster centers (centroids) $\boldsymbol{\mu}_k$
  2. **E-Step:** Assign each point to nearest centroid using Euclidean distance
  3. **M-Step:** Update centroids as mean of assigned points
  4. Repeat until convergence
- **Distance Calculation:** $\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 = \sum_d (x_{nd} - \mu_{kd})^2$
- **Objective Function:** Minimize sum of squared distances

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

Gaussian Mixture Models

- **Model Definition:**

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- **Gaussian PDF:** For 1D case

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Expectation step:** Compute the responsibilities given current parameter estimates:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

- **Maximization step:** Update parameter estimates using current responsibilities:
  - Update means: $\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$
  - Update covariances: $\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^T$
  - Update mixing coefficients: $\pi_k^{\text{new}} = N_k/N$, where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

Linear Regression

- **Linear Model:**

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

- **Design Matrix:** $\boldsymbol{\Phi}$ contains basis functions evaluated at all points
- **Normal Equations:**

$$\mathbf{w}_{ML} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t}$$

- **Regularized Solution:**

$$\mathbf{w} = (\lambda \mathbf{I} + \boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t}$$

- **Error Function:**

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n))^2$$

Exercises

Limit your solutions to 3 decimal places.

1.1 Gaussian Mixture Models

Given a one-dimensional Gaussian mixture model with two components:

$$p(x) = 0.7\mathcal{N}(x|2, 1) + 0.3\mathcal{N}(x|5, 2)$$

a) Calculate the responsibilities $\gamma(z_1)$ and $\gamma(z_2)$ for a data point $x = 3.5$ (E-step).

b) Given the data point from the item above and the following additional data points and their responsibilities:

x	$\gamma(z_1)$	$\gamma(z_2)$
2.0	0.9	0.1
4.0	0.4	0.6
5.5	0.1	0.9

Perform one M-step of the EM algorithm to compute:

- New mixing coefficients π_1^{new}, π_2^{new}
- New means μ_1^{new}, μ_2^{new}
- New variances $(\sigma_1^2)^{\text{new}}$, $(\sigma_2^2)^{\text{new}}$

Give the answer in the form $p(x) = \pi_1\mathcal{N}(\mu_1, \sigma_1^2) + \pi_2\mathcal{N}(\mu_2, \sigma_2^2)$.

1.2 Linear Regression

Consider a simple linear regression problem with the following three data points: $(1, 2)$, $(2, 4)$, and $(3, 5)$.

a) Using the normal equations, find the best-fit line $y = w_0 + w_1 x$.

b) Calculate the regularized solution with $\lambda = 1$ using the formula:

$$\mathbf{w} = (\lambda \mathbf{I} + \mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t}$$

c) Draw a plot comparing the regularized and unregularized solutions.

1.3 K-means Clustering (Extra)

Consider a dataset with four 2D points: $(1, 1)$, $(2, 1)$, $(4, 3)$, and $(5, 4)$. Perform one complete iteration of the K-means algorithm with $K = 2$, starting with initial centroids $\mu_1 = (1, 1)$ and $\mu_2 = (5, 4)$.