Assignment 4: Unsupervised and Generative Models

31 March 2024 20:08

Mainak Biswas
PhD, student

[20 points - 3 hrs]

Coding :-

1) Principal Component Analysis, K-means clustering [10 points]

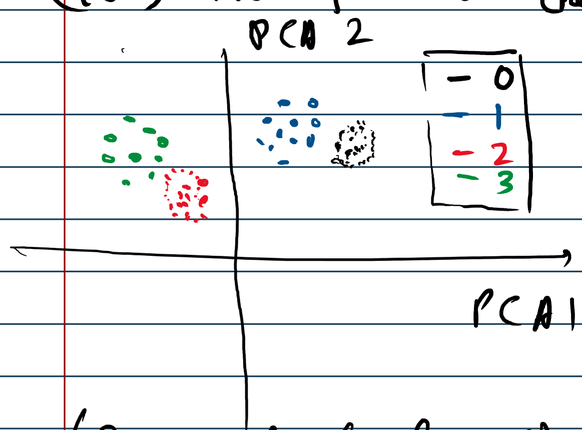
Import the MNIST dataset {Like assignment 3, Q.5}.

(i) Randomly Subsample the dataset to get 1000 samples per digit.
{i.e. 1000 × 10 = 10,000 data points} [0.5]

(ii) Now perform PCA on the (10,000 × 784) to reduce it to 2 features. → (10,000 × 2) → after PCA.

Note: (you can use numpy to function to get eigenvectors and values} [4]

(iii) Now plot the data as a scatter plot with the PCA coordinates being the axes. Also give different colours to all the different digits. Do the same digits look to be clustered together?



[1]

(Sample display format)

(iv) Now use the (10,000 × 2) as your data {after PCA}

$X = \{x_i\}_{i=1}^{10,000}$, where $x_i \in \mathbb{R}^2$.

Write K-means algorithm and use it to cluster the dataset X in k = 10 clusters. [3]

(v) Visualize the clusters like in (iii). But note that they don't have labels. So just name them Cluster 1, Cluster 2 ... Cluster 10.

{Comment on the fact — if K-means is able to identify the clusters: i.e. does the configuration look similar to (iii)} [1.5]

Theory:

2) Soft-clustering/probabilistic clustering — [6 points]

Let $\{x_i\}_{i=1}^{N}$, $x \in \mathbb{R}^d$ be given. We want to cluster the data into $(C_1, C_2, \ldots C_K)$ clusters with their centroids at $(\mu_1, \mu_2, \ldots \mu_K)$. But unlike k-means we want to assign probability of a data point x_i to be in cluster C_j s.t. $P(C_j|x_i) = \gamma_{ij}$.

It turns out that it can be solved by solving the following objective:

$$\gamma_{ij}^*, \mu_j^* = \underset{\mu_j, \gamma_{ij}}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{K} \gamma_{ij}^2 \|x_i - \mu_j\|^2$$

$$\text{s.t.} \quad \sum_{j=1}^{K} \gamma_{ij} = 1, \quad \gamma_{ij} \geq 0 \quad \forall\, i = 1, 2 \ldots N,\; j = 1, 2 \ldots K$$

Hence answer:

(i) Define the unconstrained problem using Lagrangian multipliers [1]

(i) Define the unconstrained problem using Lagrangian multipliers [1]

[Hint: ignore the contraints $\gamma_{ij} \geq 0$ and check if it is satisfied at the end]

(ii) Hence solve for μ_j, γ_{ij}, and Lagrange multipliers [3]

(iii) Then, come up with a iterative algorithm to get optimal - μ_i^*, γ_{ij}^*
{It will look like the k-means algorithm} [1]

(iv) How will you quantize γ_{ij}'s so that it turns out to be k-means algorithm?

3) Expectation Maximization [4 points]

Data: $\{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}$ (scalar)

Defn: Now, we want to model this using a mixture of k-exponential distributions, as:

$$P_\theta(x | Z = i) = \mathrm{Exp}(x ; \lambda_i) = \lambda_i e^{-\lambda_i x} \mathbb{1}_{\{x \geq 0\}}$$

$$P_\theta(Z) = \mathrm{Multinomial}(\mu_1, \mu_2 \cdots \mu_K)$$

Mixture model: $\theta^* = \underset{\theta}{\mathrm{argmin}} \log \sum_{i=1}^{K} \underbrace{P_\theta(x|Z=i) P_\theta(Z=i)}_{P_\theta(x, Z=i)}$

(MLE) →

Hence:

1) Formulate the Expectation step [2]

1) Formulate the Expectation step [2]

2) Formulate the maximization step [2]