TAM 598 Lecture 17:

Unsupervised Learning - Clustering & Density Estimation

Announcements:

- HW 4 covers lectures 13-16; due on Fri Apr 4
- HW 5 covers lectures 17-20; due on Fri Apr 18
- No class next Weds Apr 8

UNSUPERVISED LEARNING

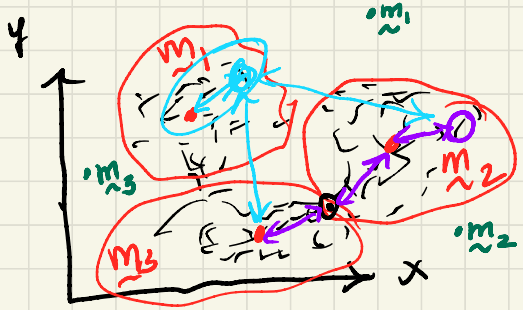
– you are given observations $\underset{\sim}{X}_{1:n} = (\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n)$ and you want to find some structure in the data. (No labels, targets, outputs)

Common approaches:

1) clustering – split observations into K distinct clusters
2) dimensionality reduction – reduce dimensionality of the data
3) density estimation – learn the probability density that gave rise to the data (ie how to generate new samples with similar features as the observations)
4) etc

Clustering using K-means (Chapter 20.1, MacKay 2003)

- define the K clusters by their centroids $\underset{\sim}{m}_{1:K}$, which are the means of the data points assigned to the cluster
- each observation $\underset{\sim}{x}_i$ is assigned to the cluster with the closest centroid, indicated as a one-hot encoding $\underset{\sim}{z}_i$



$$\underset{\sim}{z}_i = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^K$$

where

$$z_{ik} = \begin{cases} 1 & \text{if } k = \underset{k'}{\operatorname{argmin}} \; \| \underset{\sim}{x}_i - \underset{\sim}{m}_{k'} \|^2 \\ 0 & \text{otherwise} \end{cases}$$

↳ the centroids are what we try to learn, by minimizing the sum of squared distances between the data points and their assigned centroids

$$\min_{\underset{\sim}{m}_{1:K}} \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ik} \| \underset{\sim}{x}_i - \underset{\sim}{m}_k \|^2$$

↳ Algorithm: start by initializing the centroids randomly, and iterate until converged:

(1) assign each data point to the cluster with the closest centroid

(2) update the centroids to the mean of the data points assigned to the cluster

Density Estimation via Gaussian Mixtures (Chapter 9, Bishop, 2006)

given: a set of observations $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$

learn: a model $p(\underset{\sim}{x})$ that allows you to generate examples similar to your observations

A model of form: $p(\underset{\sim}{x}) = \sum_{k=1}^{K} \pi_k \underbrace{N(\mu_k, \Sigma_k)}$ is a Gaussian Mixture Model

Each is a multivariate gaussian

parameters to find:

π_k, μ_k, Σ_k

solving the density estimation problem also solves the clustering problem since you can think of the K different gaussians as clusters

Find the parameters by maximizing log likelihood

algorithm to maximize log likelihood is the expectation-maximization (EM) algorithm

iterate between E|M:

- E - compute expected value of log likelihood w/r/t the conditional distribution of latent variables, given the observed data and the current estimate of the parameters (latent variables → cluster assignments)
- M - maximize the expected value w/r/t parameters

so: $\underset{\sim}{x}$ = observations

$\underset{\sim}{\theta}$ = parameters (π_k, μ_k, Σ_k)

$\underset{\sim}{z}$ = latent variables, cluster assignments

MAXIMIZE:

$$\underset{\theta}{\operatorname{argmax}}\ p(\underset{\sim}{x}|\underset{\sim}{\theta}) = \int p(\underset{\sim}{x}, \underset{\sim}{z}|\underset{\sim}{\theta})\, d\underset{\sim}{z}$$

$$= \int p(\underset{\sim}{x}|\underset{\sim}{z}, \underset{\sim}{\theta})\, p(\underset{\sim}{z}|\underset{\sim}{\theta})\, d\underset{\sim}{z}$$

is intractable because z is not observed