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

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
2) dimensionality reduction
3) density estimation

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 $\underset{\sim}{z}_i$

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

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

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

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

iterate between E IM:

E -

M -

Avoiding overfitting using the Bayesian information criterion