Lecture 5-6-7
Generative Models

Ref: Outlier Analysis, Charu C Agrawal
Ref: Outlier Analysis: A Review, Chandola et al.

Limitations of Euclidean distance

$$d_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots (x_n - y_n)^2}$$

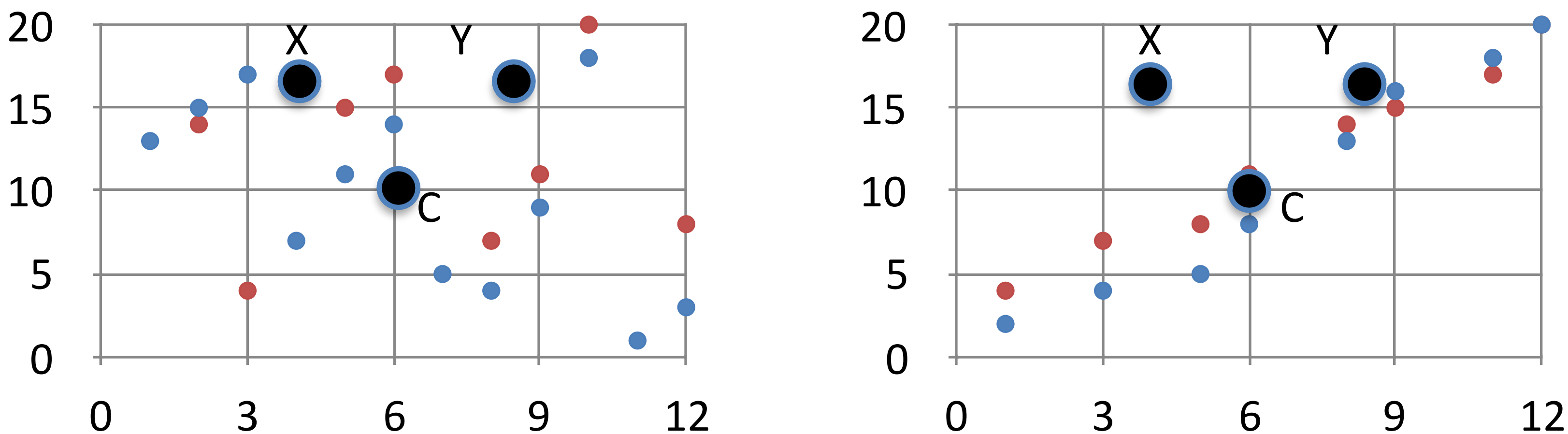
- Gives equal weightage to each dimension
- A feature in lower range will have minimal effect on the score
- Features may be correlated

Range Problem

	Price (INR)	Weight (g)	Price (lakh)	Weight (kg)
Phone 1	36000	400	0.36	0.4
Phone 2	37000	420	0.37	0.42
Phone 3	60000	350	0.60	0.35
Phone 4	20000	510	0.20	0.51

$$d_{12} = 1000 \qquad d'_{12} = 0.022$$

Correlated Features



Euclidean distance of X and Y is the same from C in both the figures!

Mahalanobis Distance

- It is a metric to measure distance between a point and a distribution
- It is very effective for multivariate distributions

Mahalanobis Distance

$$D^2 = (x - c)^T \Sigma^{-1} (x - c)$$

where Σ is the covariance matrix.

Generative Models

Underlying Principle

"An anomaly is an observation which is suspected of being partially or wholly irrelevant because it is not generated by the stochastic model assumed"

Ref: Anscombe and Guttman 1960

Main Assumption

Normal data instances occur in high probability regions of a stochastic model, while anomalies occur in the low probability regions of the stochastic model.

Probabilistic Generative Models

- Train a generative probabilistic model
- Calculate probability (or probability density) of a given data point
- Inverse of this is the anomaly score

How to train a generative model

- Assume an underlying model that lead to generation of the dataset
- The model is generally a mixture of components (e.g. Gaussians)
- The model parameters are learned such that the dataset has maximum likelihood of being generated

Probability Mixture Models

- Probabilistic version of clustering
- Dataset is modelled as mixture of Gaussians
- Inverse of probability density can be used as anomaly score

Two Paradigms

1. Mixture components may only model normal data
2. There can be separate components to model normal as well as abnormal data

Separate Models

$$D = \lambda A + (1 - \lambda)M$$

- Each data point is an anomaly with a prior probability.
- Since we do not know which data is generated by which distribution, we use EM to find A and M.

Gaussian Mixture Model (GMM)

- A probabilistic generative model
- Assumes the data is generated by a mixture of Gaussian distributions
- The mixture components can represent normal data only or both normal data and anomalies

The Gaussian Distribution

- **Univariate density**

$$\mathcal{N}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2 * \pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean

Variance

- **Multivariate density**

$$\mathcal{N}(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{\sqrt{2 * \pi |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}$$

Mean

Covariance

Probability density when the data is represented by a mixture of Gaussians

Number of Gaussians

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x} \,|\, \mu_{\mathbf{k}}, \Sigma_k)$$

Mixing coefficient

$$0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1$$

Data Likelihood: probability of observing data given a GMM

- **Likelihood**

$$p(X/\mu, \Sigma, \pi) = \prod_{n=1}^{N} f(x_n)$$

- **Log Likelihood**

$$ln\ p(X/\mu, \Sigma, \pi) = \sum_{n=1}^{N} f(x_n) = \sum_{n=1}^{N} ln\ \left(\sum_{k=1}^{K} \pi_k\ \mathcal{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k) \right)$$

Parameter Estimation

- Obtain parameters such that the log likelihood is maximised
- No closed form solution is possible for GMM

If we know which data point is generated by which Gaussian distribution, we can easily calculate the parameters!

Expectaation Maximization

Generic EM algorithm

1. Initialise the parameters (randomly of based on prior knowledge)
2. E-Step: estimate the latent variables
3. M-step: update the parameters according to the latent variables estimated in the E step
4. Repeat 2-3 until convergence

How to estimate latent parameter (component for each data point)?

- PDF of being generated by kth component

$$\pi_k \, \mathcal{N}(\mathbf{x} \,|\, \mu_{\mathbf{k}}, \Sigma_k)$$

- Probability of x being generated by kth component (also called responsibility of nth component)

$$\gamma_k(\mathbf{x}) = \frac{\pi_k \, \mathcal{N}(\mathbf{x} \,|\, \mu_{\mathbf{k}}, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}(\mathbf{x} \,|\, \mu_{\mathbf{j}}, \Sigma_j)}$$

Each data point is assigned to all the clusters, also known as soft clustering!

M-Step

- Update the parameters according to the estimated latent variable
- In current case, we have responsibilities of each component for a given data point
- Use the responsibilities as fraction of the data point being generated by that component

Update Weight

$$\pi_j = \frac{1}{N}\sum_{n=1}^{N}\gamma_j(\mathbf{x}_n)$$

Update Mean

$$\mu_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)}$$

Update Covariance

$$\Sigma_j = \frac{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)(\mathbf{x}_n - \mu_j)(\mathbf{x}_n - \mu_j)^T}{\sum_{n=1}^{N} \gamma_j(\mathbf{x}_n)}$$

Calculate the log likelihood again, stop if there is no change!