Model Based Clustering

Mixture Models and EM Algorithm

Model Based Clustering

- In order to understand our data, we will assume that there is a generative process (a model) that creates/describes the data, and we will try to find the model that best fits the data.
  - Models of different complexity can be defined, but we will assume that our model is a distribution from which data points are sampled
  - Example: the data is the weight of all people in US

- In most cases, a single distribution is not good enough to describe all data points: different parts of the data follow a different distribution
  - Example: the data is the weight of all male vs all female
  - We need a mixture model
  - Different distributions correspond to different clusters in the data.

Gaussian Distribution

- Example: the data is the weight of all people in US
  - Experience has shown that this data follows a Gaussian (Normal) distribution
  - Normal distribution:

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ Is the mean, σ is the standard deviation

Gaussian Model

- What is a model?
  - A Gaussian distribution is fully defined with mean μ and the standard deviation σ.
  - We define our model as the pair of parameters:

    $\theta = (\mu, \sigma)$
- This is a general principle: a model is defined as a vector of parameters θ

Fitting the Model

- We want to find the normal distribution that best fits our data
  - Find the best values for μ and σ
  - But what does best fit mean?

Maximum Likelihood Estimation (MLE)

- Suppose that we have a vector $X = (x_1, .., x_n)$ of values
- And we want to fit a Gaussian $N(\mu, \sigma)$ model to the data
- Probability of observing point x_i : $P(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Probability of observing all points (assuming independence): $P(X) = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- We want to find the parameters $\theta = (\mu, \sigma)$ that maximize the probability $P(X|\theta)$

Maximum Likelihood Estimation (MLE)

- The probability $P(X|\theta)$ as a function of θ is called the Likelihood function $L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- It is usually easier to work with the Log-Likelihood function

$$LL(\theta) = -\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{1}{2} n \log 2\pi - n \log \sigma$$

- Maximum Likelihood Estimation
  - Find parameters μ, σ that maximize $LL(\theta)$

Sample mean: $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample variance: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i-\mu)^2$

MLE

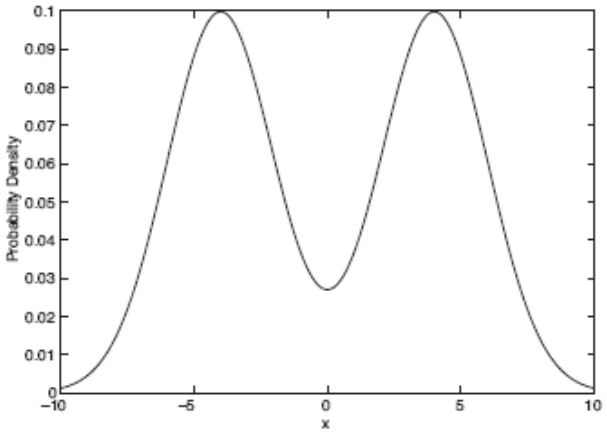
- Note: these are also the most likely parameters given the data

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

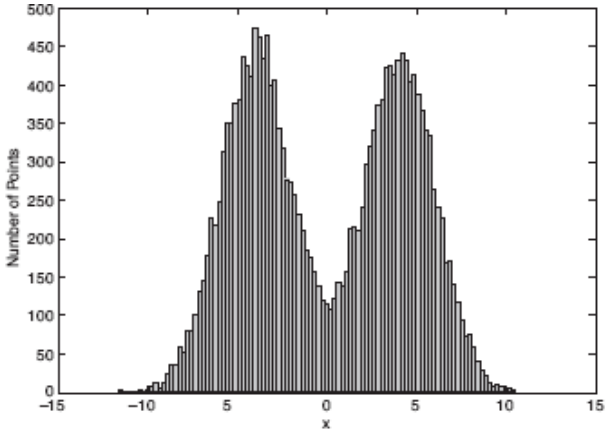
- If we have no prior information about, θ or X, then maximizing $p(X \mid \theta)$ is the same as maximizing $p(\theta|X)$

Mixture of Gaussian

- Suppose that you have the weights of male vs female and the distribution looks like the figure below (dramatization)



(a) Probability density function for the mixture model.

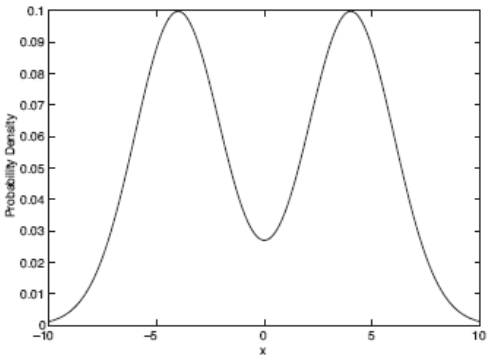


(b) 20,000 points generated from the mixture model.

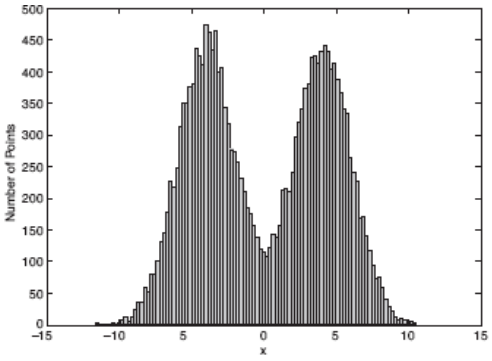
Mixture model consisting of two normal distributions with means of -4 and 4, respectively. Both distributions have a standard deviation of 2.

Mixture of Gaussians

- In this case the data is the result of the mixture of two Gaussians
  - One for male, and one for female
  - Identifying for each value which Gaussian is most likely to have generated it will give us a clustering



(a) Probability density function for the mixture model.



(b) 20,000 points generated from the mixture model.

Mixture model consisting of two normal distributions with means of -4 and 4, respectively. Both distributions have a standard deviation of 2.

Mixture Model

- A value x_i is generated according to the following process:
  - First select the gender
  - With probability π_M select male, with probability π_F select female
    ($\pi_M + \pi_F = 1$)

- Given the nationality, generating the points from the corresponding Gaussian:
  - $P(x_i \mid \theta_M) \sim N(\mu_M, \sigma_M)$ if male
  - $P(x_i \mid \theta_F) \sim N(\mu_F, \sigma_F)$ if female

Mixture Model

- Our model has the following parameters

$$\theta = (\pi_{M,} \pi_{F,} \mu_{M}, \sigma_{M,} \mu_{F}, \sigma_{F})$$

- For value x_i, we have:

$$P(x_i \mid \theta) = \pi_M^* P(x_i \mid \theta_M) + \pi_F^* P(x_i \mid \theta_F)$$

- For all values $X = x_1, . , x_n$

$$P(X \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta)$$

- We want to estimate the parameters that maximize the Likelihood of the data

Mixture Models

- Once we have the parameters $\theta = (p_M, p_F, \mu_M, \sigma_M, \mu_F, \sigma_F)$, we can estimate the membership probabilities $P(M \mid x_i)$ and $P(F \mid x_i)$ for each point x_i:
  - This is the probability that point x_i belongs to Male or Female population (cluster)

$$p(M \mid x_i) = \frac{p(x_i \mid M)p(M)}{p(x_i \mid M)p(M) + p(x_i \mid F)p(F)}$$

$$= \frac{p(x_i \mid M)\,\pi_M}{p(x_i \mid M)\,\pi_M + p(x_i \mid F)\,\pi_F}$$

EM (Expectation and Maximization) Algorithm

- Initialize the values of the parameters in θ to some random values
- Repeat until convergence:
  - E-Step: Given the parameters estimate the membership probabilities $P(\theta_M | x_i)$ and $P(\theta_F | x_i)$
  - M-Step: Compute the parameter values that (in expectation) maximize the data likelihood

$$\pi_M = \frac{1}{n}\sum_{i=1}^{n} P(M|xi) \qquad \pi_F = \frac{1}{n}\sum_{i=1}^{n} P(F|xi)$$

Fraction of population in M or F

$$\mu_M = \frac{1}{n}\sum_{i=1}^{n} \frac{P(M|xi)}{n*\pi_M} x_i \qquad \mu_F = \frac{1}{n}\sum_{i=1}^{n} \frac{P(F|xi)}{n*\pi_F} x_i$$

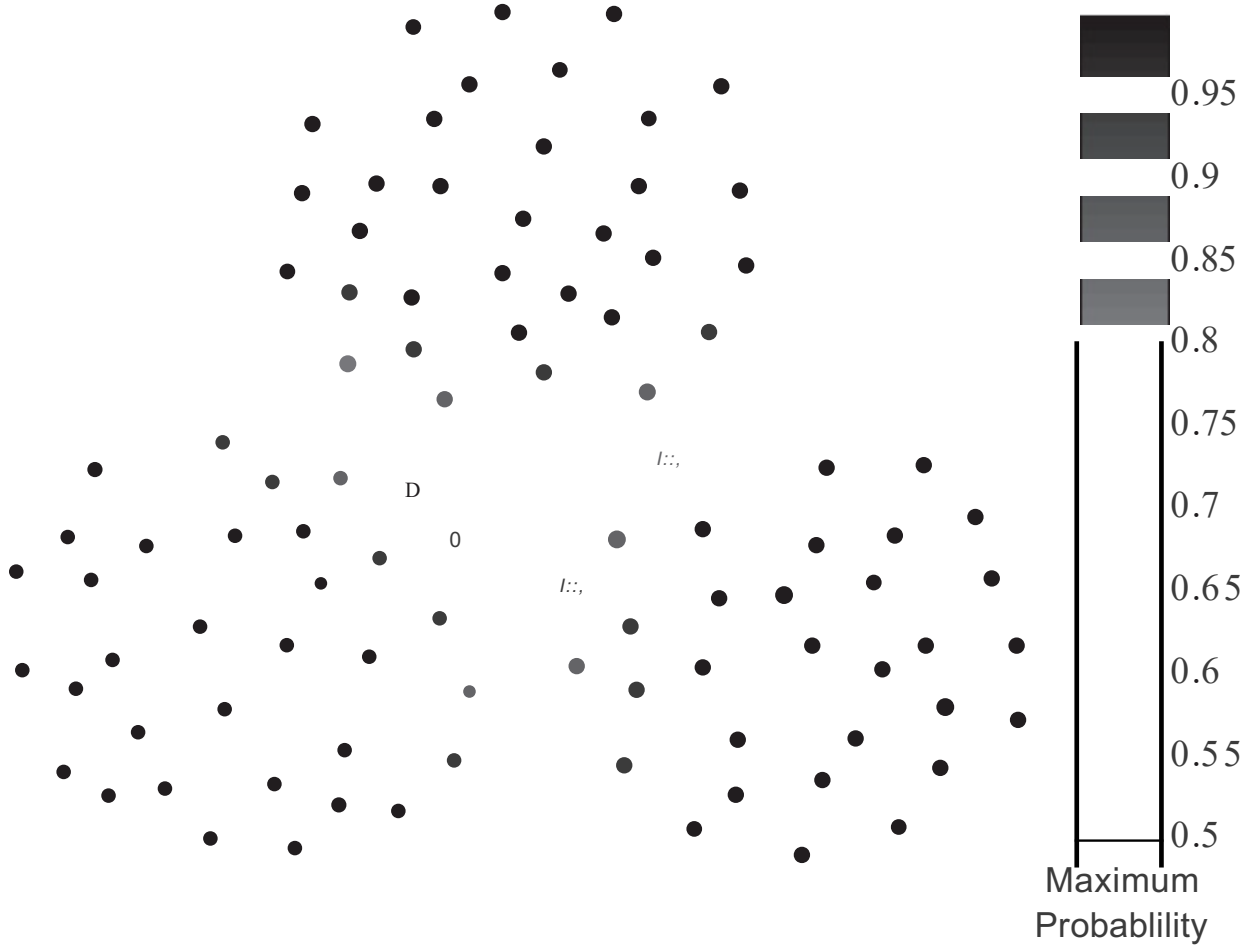
MLE Estimates if πs were fixed

$$\sigma_M^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{P(M|x_i)}{n*\pi_M} (x_i - \mu_M)^2 \qquad \sigma_F^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{P(F|x_i)}{n*\pi_F} (x_i - \mu_F)^2$$

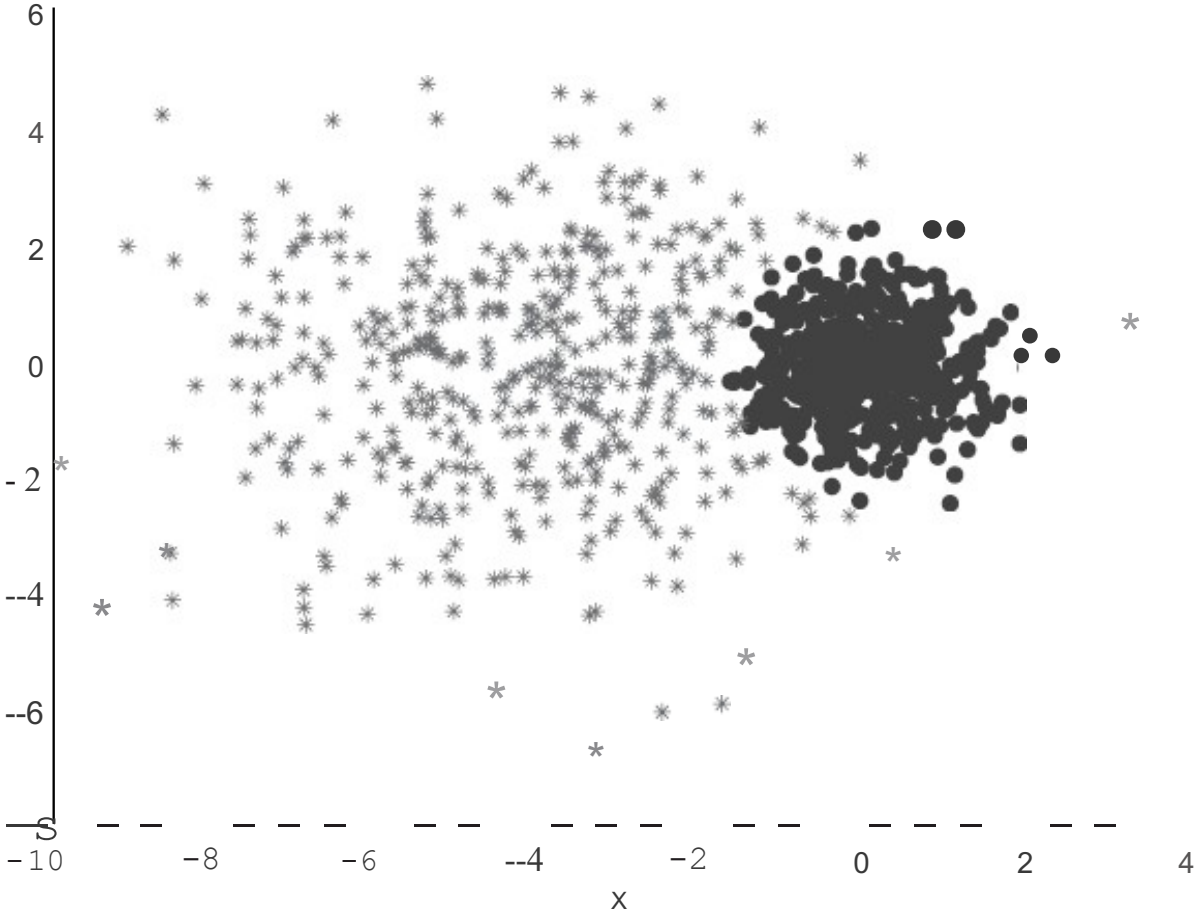
Relationship to K-means

- E-Step: Assignment of points to clusters
  - K-means: hard assignment, EM: soft assignment
- M-Step: Computation of centroids
  - K-means assumes common fixed variance (spherical clusters)
  - EM: can change the variance for different clusters or different dimensions (elipsoid clusters)
- If the variance is fixed then both minimize the same error function

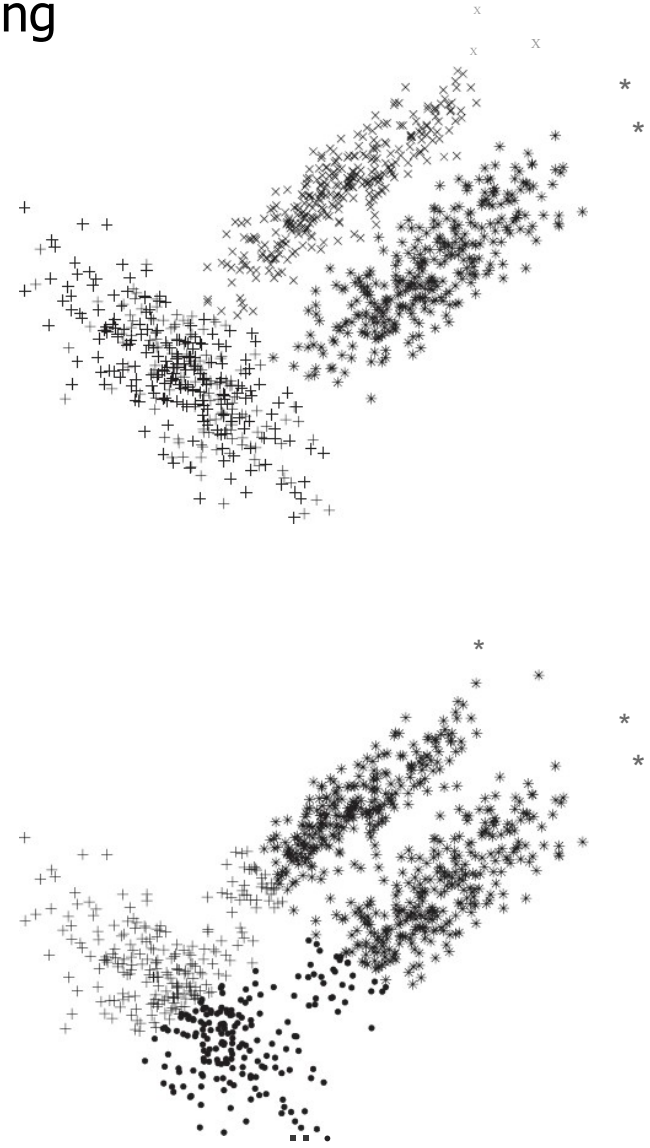
Mixture model clustering with three clusters of spherical shape



Mixture model clustering with two clusters of elliptical shape



Mixture model clustering

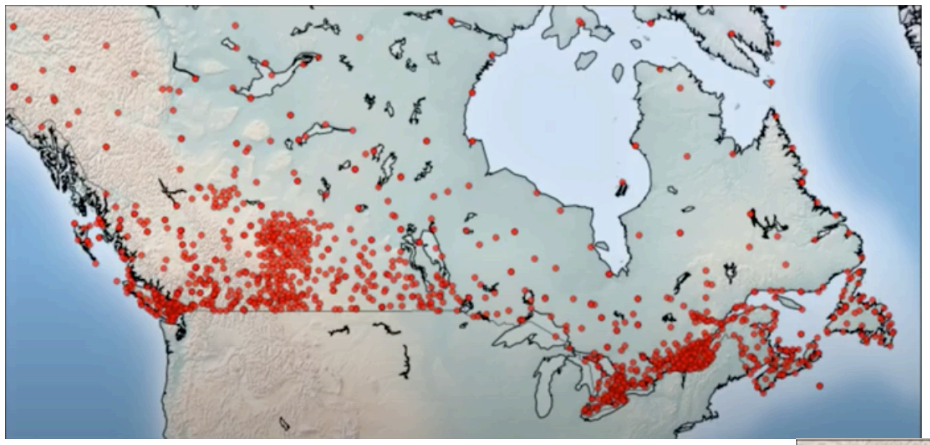


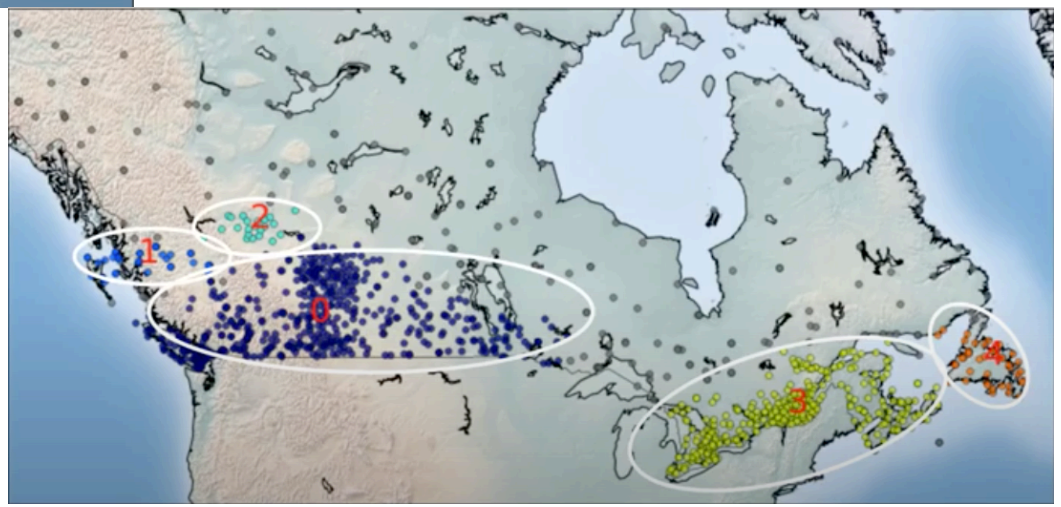
K means clustering

Density Based Clustering

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

- In density based clustering we partition points into dense regions separated by not-so-dense regions.





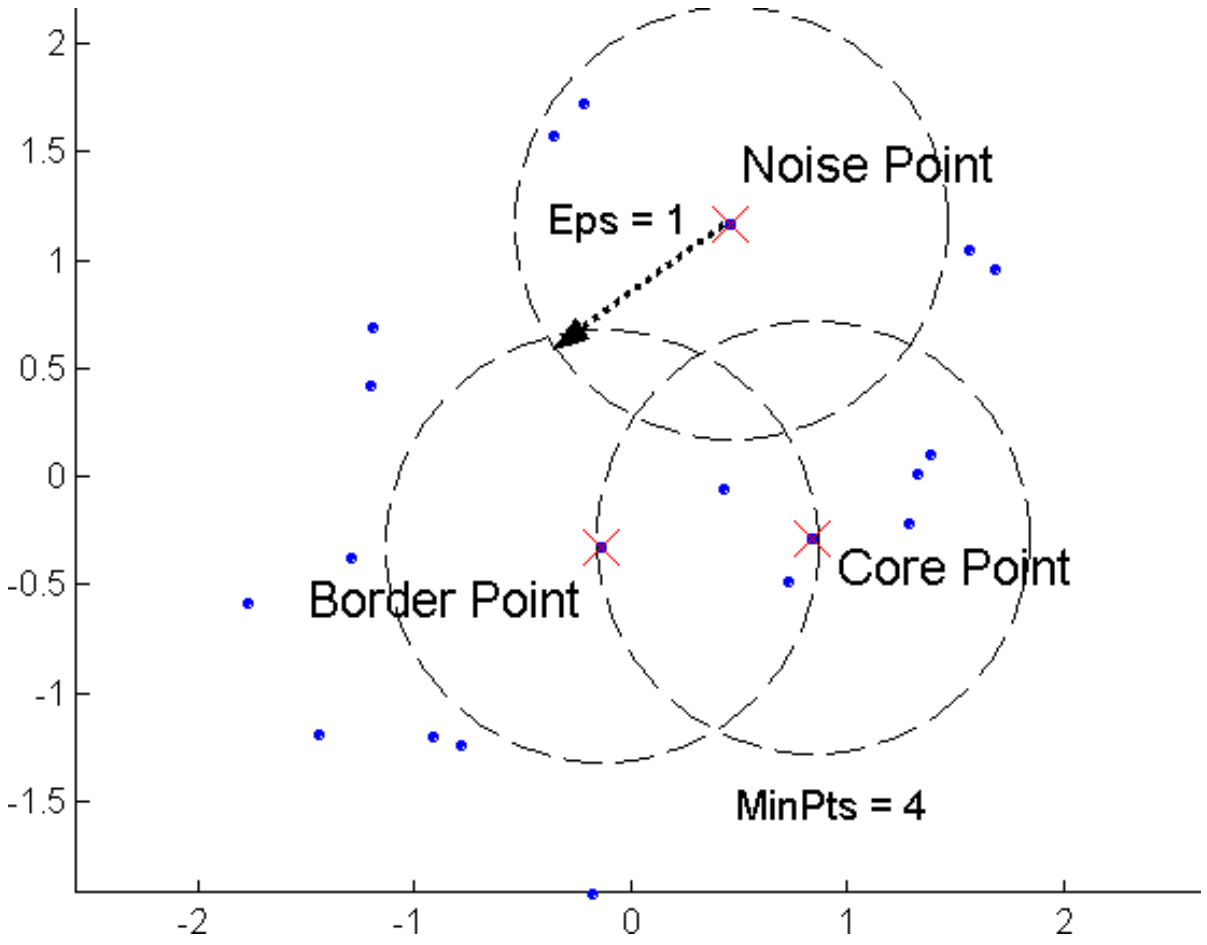
DBSCAN: Density-Based Clustering

- Important Questions:
  - How do we measure density?
  - What is a dense region?
- Two Parameters:
  - Density at point p: number of points within a circle of radius Eps
  - Dense Region: A circle of radius Eps that contains at least MinPts points

DBSCAN

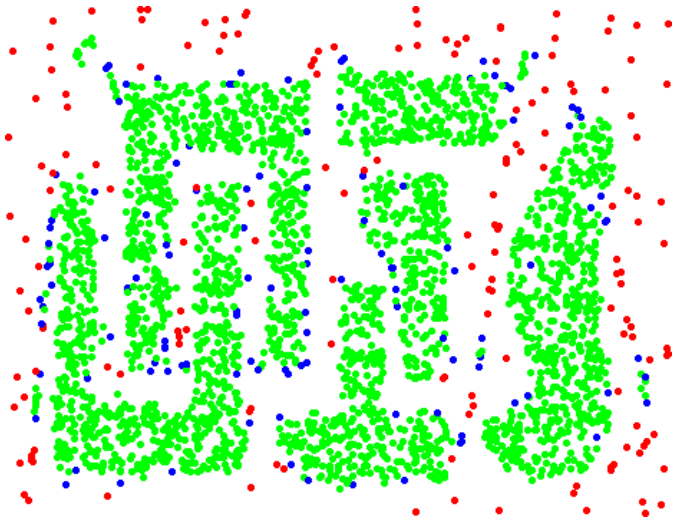
- Characterization of points
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
    - These points belong in a dense region and are at the interior of a cluster
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point.
    - Directly reachable by a core point
  - A noise point is any point that is not a core point or a border point.

DBSCAN: Core, Border, and Noise Points



DBSCAN: Core, Border, and Noise Points

Original Points



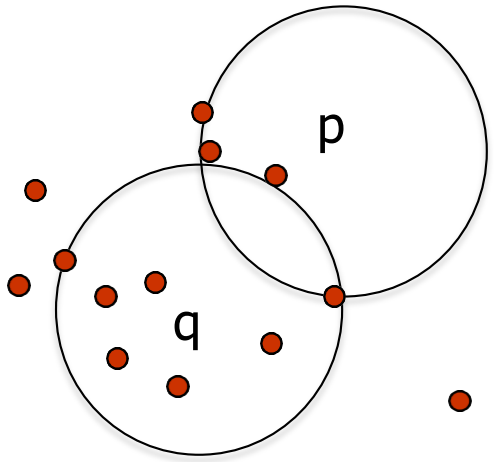
Point types: core, border and noise

Eps = 10, MinPts =4

Density-Reachable Points

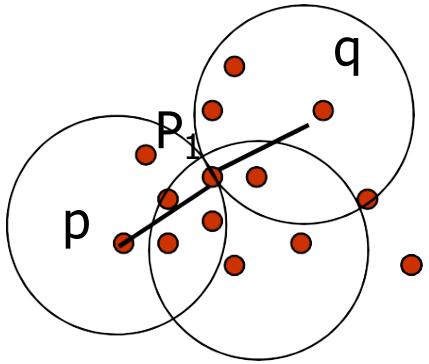
- Directly Density-reachable:
  - A point p is directly density-reachable from a point q w.r.t. Eps, MinPts if:
    - P belongs to $N_{Eps}(q)$
    - Core point condition: $|N_{Eps}(q)| \geq$ MinPts

Eps = 1
MinPts = 5

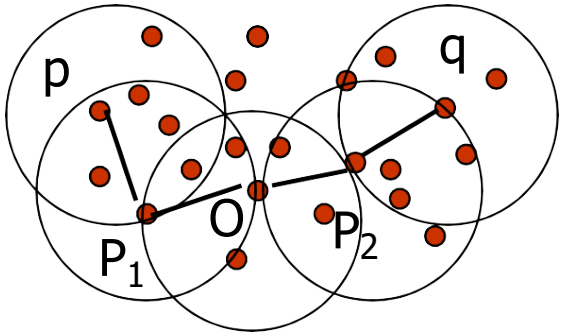


Density-Connected Points

- Density reachable:
  - A point p is density-reachable from a point q w.r.t. Eps, MinPts if there is a chain of points P_1, ..., P_n, P_1=q, P_n=p such that P_{i+1} is directly density-reachable from P_i



- Density-connected:
  - A point p is density-connected to a point q if there is a point O such that both p to q are density-reachable from O w.r.t. Eps and MinPts



DBSCAN Algorithm

- Label points as core, border and noise
- Eliminate noise points
- For every core point p that has not been assigned to a cluster
  - Create a new cluster with the point p and all the points that are density-connected to p.
- Assign border points to the cluster of the closest core point.

Algorithm: DBSCAN: a density-based clustering algorithm.

Input:

- D: a data set containing n objects,
- ϵ: the radius parameter, and
- $MinPts$: the neighborhood density threshold.

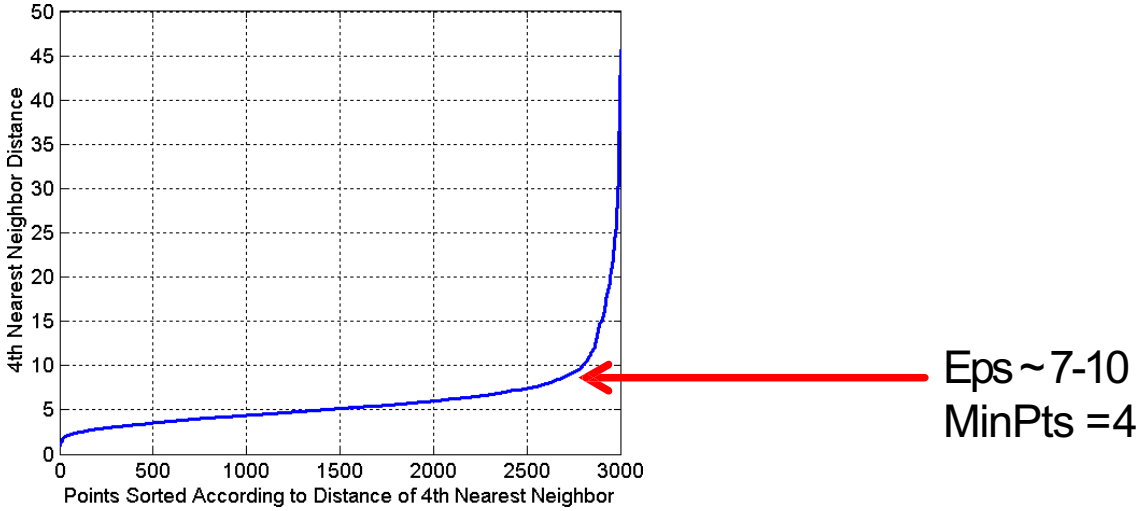
Output: A set of density-based clusters.

Method:

```
(1)  mark all objects as unvisited;
(2)  do
(3)      randomly select an unvisited object p;
(4)      mark p as visited;
(5)      if the ε-neighborhood of p has at least MinPts objects
(6)          create a new cluster C, and add p to C;
(7)          let N be the set of objects in the ε-neighborhood of p;
(8)          for each point p' in N
(9)              if p' is unvisited
(10)                 mark p' as visited;
(11)                 if the ε-neighborhood of p' has at least MinPts points,
                     add those points to N;
(12)             if p' is not yet a member of any cluster, add p' to C;
(13)         end for
(14)         output C;
(15)     else mark p as noise;
(16) until no object is unvisited;
```

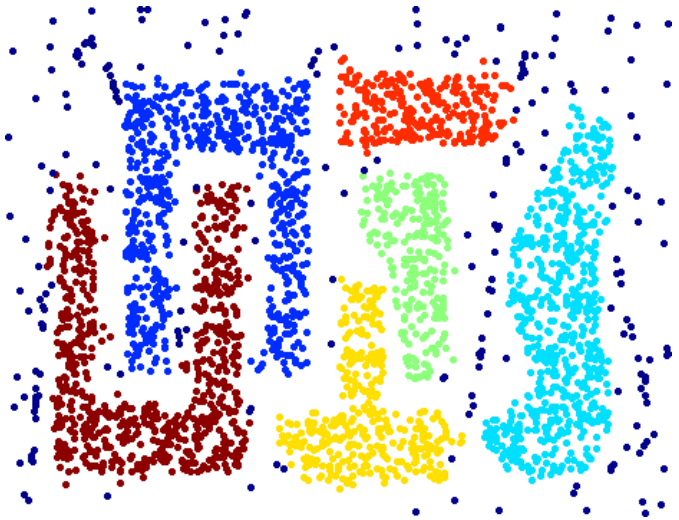
Determining Eps and MinPts

- Idea is that for points in a cluster, their k^{th} nearest neighbors are at roughly the same distance
- Noise points have the k^{th} nearest neighbor at farther distance
- So, plot sorted distance of every point to its k^{th} nearest neighbor
- Find the distance d where there is a "knee" in the curve
  - Eps = d, MinPts = k



When DBSCAN Works Well

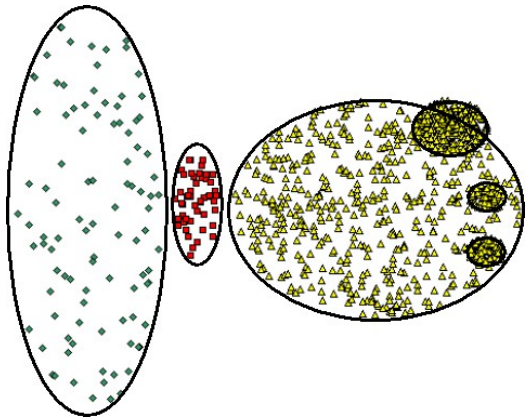
Original Points



Clusters

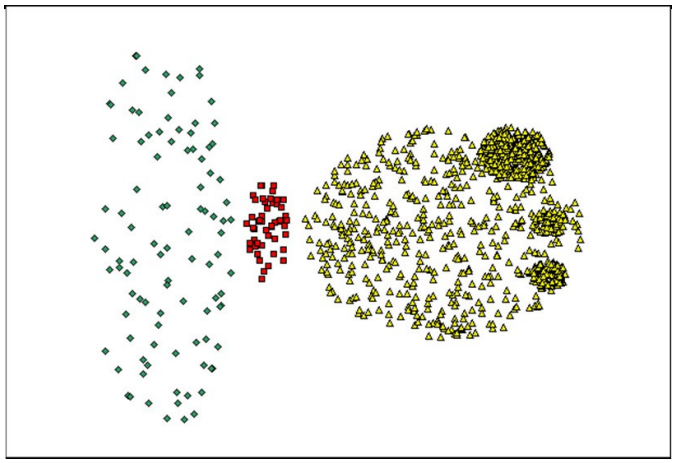
- Resistant to Noise
- Can handle clusters of different shapes and sizes

When DBSCAN Does not Work Well

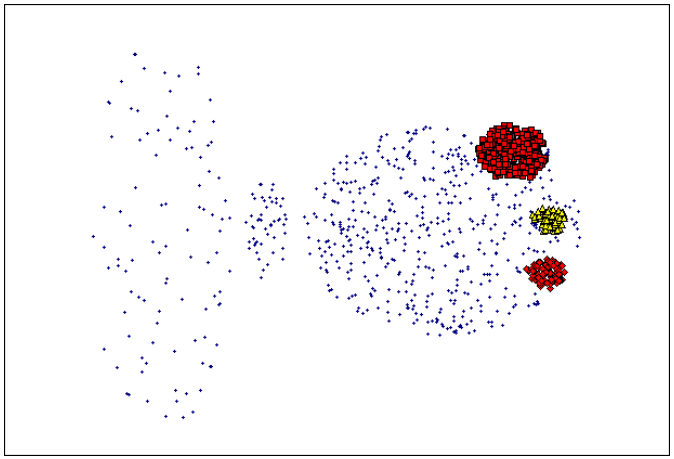


Original Points

- Varying densities
- High-dimensional data



(MinPts=4, Eps=9.75).



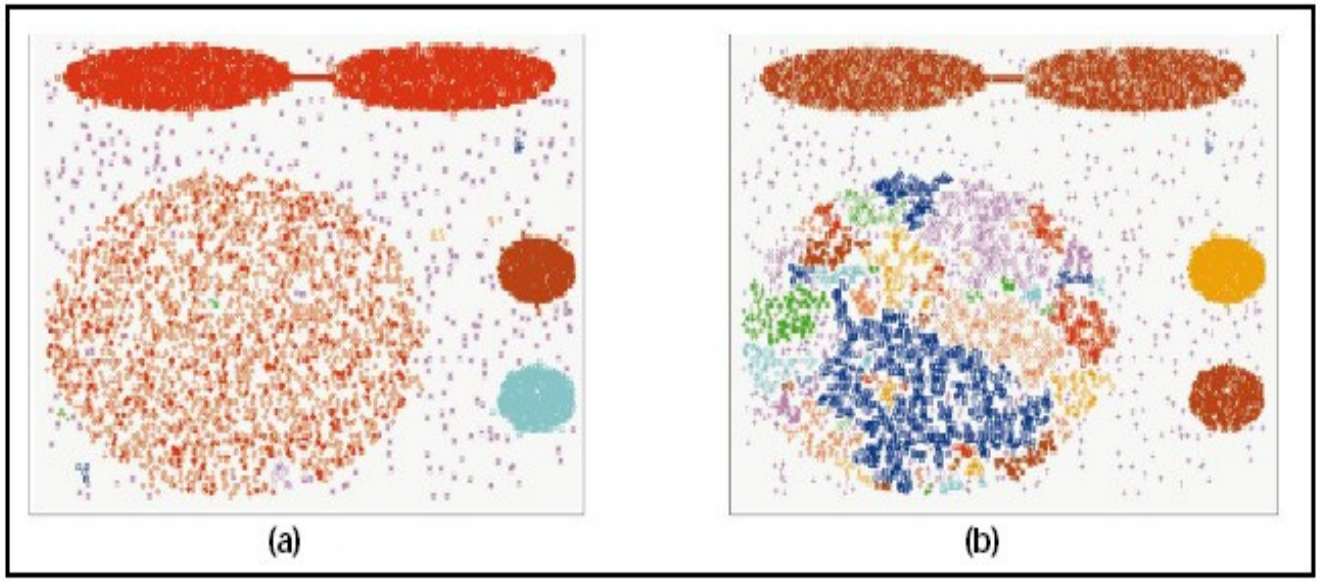
(MinPts=4, Eps=9.92)

DBSCAN Sensitive to Parameters

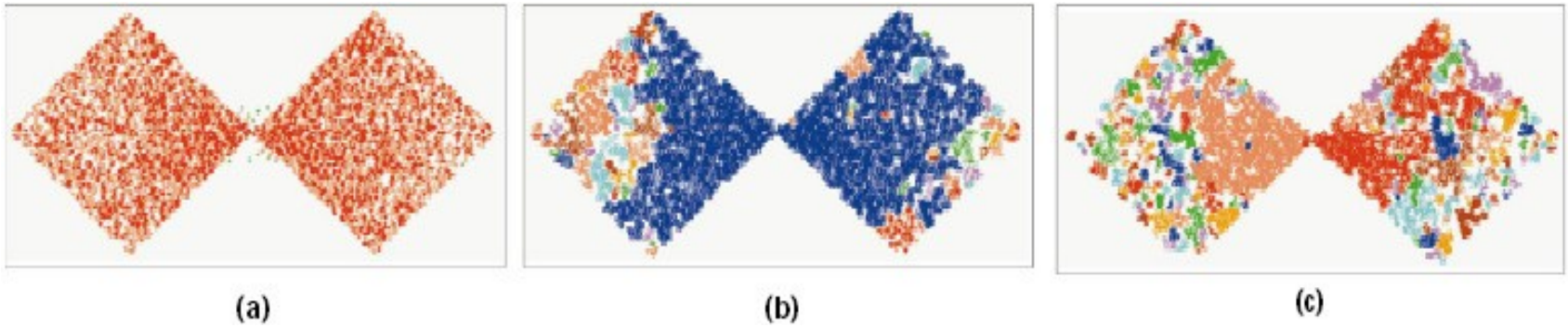
MinPts = 4

Data Set 1

Eps at (a) 0.5, (b) 0.4



Data Set 2



Eps at (a) 5.0, (b) 3.5, and (c) 3.0

G. Karypis, Han, Kumar, Computer, 32(8), 1999