

The background of the slide is a complex, abstract composition. It features a dark, muted purple or brownish background. Overlaid on this are several geometric and data-like elements: a network of thin, light-colored lines forming a web-like structure; numerous small, green circular dots scattered across the space; and a series of faint, light-colored plus signs arranged in a grid-like pattern. A prominent white, angular shape, resembling a stylized 'V' or a folded piece of paper, cuts across the middle of the slide, serving as a backdrop for the title. In the bottom left corner, there is a small, rectangular inset image showing a cluster of orange and red dots on a light blue background, with a white line and a small plus sign nearby.

Probabilistic Hierarchical Clustering

Probabilistic Hierarchical Clustering

- ❑ Algorithmic hierarchical clustering
 - ❑ Nontrivial to choose a good distance measure
 - ❑ Hard to handle missing attribute values
 - ❑ Optimization goal not clear: heuristic, local search
- ❑ Probabilistic hierarchical clustering
 - ❑ Use probabilistic models to measure distances between clusters
 - ❑ Generative model: Regard the set of data objects to be clustered as a sample of the underlying data generation mechanism to be analyzed
 - ❑ Easy to understand, same efficiency as algorithmic agglomerative clustering method, can handle partially observed data
- ❑ In practice, assume the generative models adopt common distribution functions, e.g., Gaussian distribution or Bernoulli distribution, governed by parameters

Generative Model

- Given a set of 1-D points $X = \{x_1, \dots, x_n\}$ for clustering analysis & assuming they are generated by a Gaussian distribution:

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

- The probability that a point $x_i \in X$ is generated by the model:

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

- The likelihood that X is generated by the model:

$$L(\mathcal{N}(\mu, \sigma^2) : X) = P(X | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

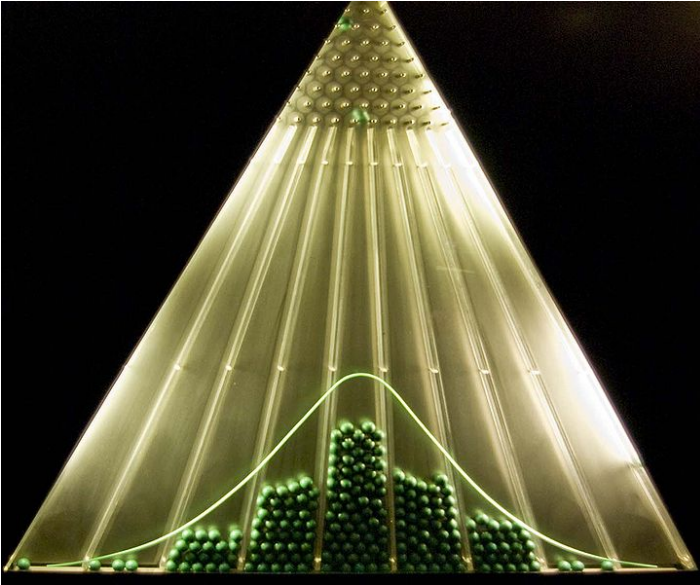
- The task of learning the generative model: find the parameters μ and σ^2 such that

$$\mathcal{N}(\mu_0, \sigma_0^2) = \arg \max \{L(\mathcal{N}(\mu, \sigma^2) : X)\}$$

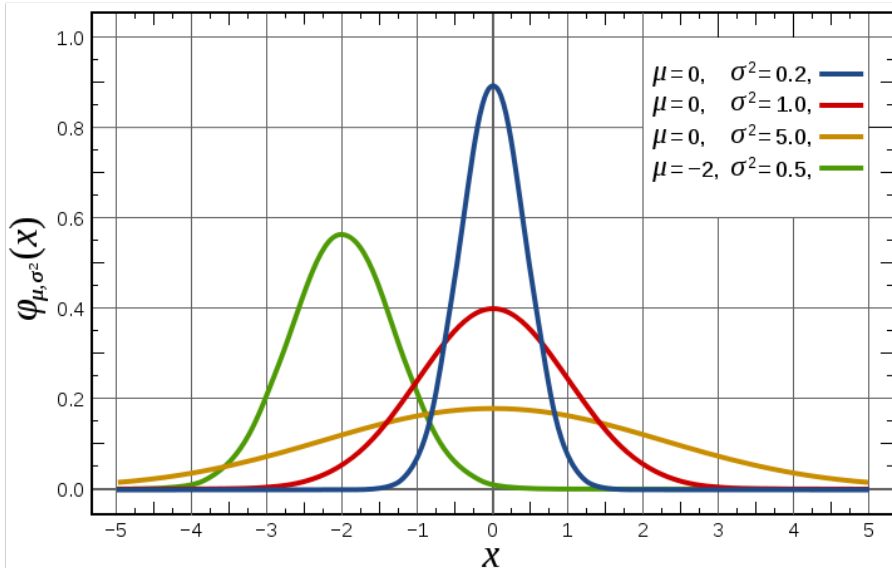
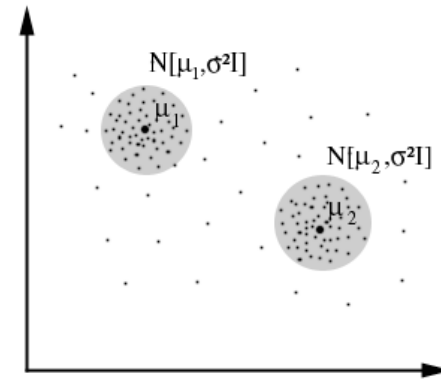
the maximum likelihood



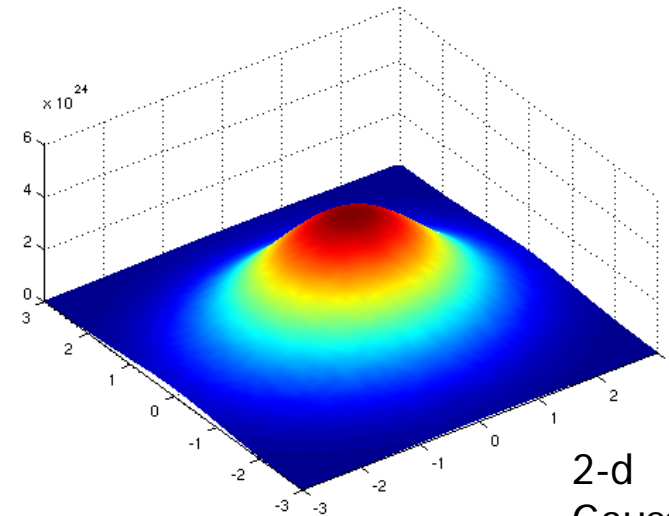
Gaussian Distribution



Bean
machine:
drop ball
with pins



1-d
Gaussian



2-d
Gaussian

From wikipedia and <http://home.dei.polimi.it>

A Probabilistic Hierarchical Clustering Algorithm

- For a set of objects partitioned into m clusters C_1, \dots, C_m , the quality can be measured by,

$$Q(\{C_1, \dots, C_m\}) = \prod_{i=1}^m P(C_i)$$

where $P()$ is the maximum likelihood

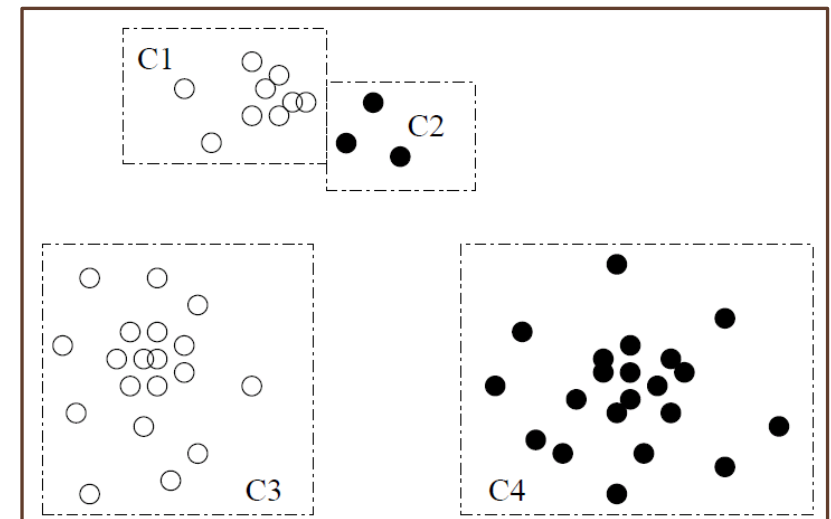
- If we merge two clusters C_{j_1} and C_{j_2} into a cluster $C_{j_1} \cup C_{j_2}$, the change in quality of the overall clustering is

$$\begin{aligned} & Q(\{C_1, \dots, C_m\} - \{C_{j_1}, C_{j_2}\} \cup \{C_{j_1} \cup C_{j_2}\}) - Q(\{C_1, \dots, C_m\}) \\ = & \frac{\prod_{i=1}^m P(C_i) \cdot P(C_{j_1} \cup C_{j_2})}{P(C_{j_1})P(C_{j_2})} - \prod_{i=1}^m P(C_i) \\ = & \prod_{i=1}^m P(C_i) \left(\frac{P(C_{j_1} \cup C_{j_2})}{P(C_{j_1})P(C_{j_2})} - 1 \right) \end{aligned}$$

- Distance between clusters C_1 and C_2 :

$$\text{dist}(C_i, C_j) = -\log \frac{P(C_1 \cup C_2)}{P(C_1)P(C_2)}$$

- If $\text{dist}(C_i, C_j) < 0$, merge C_i and C_j



Recommended Readings

- ❑ A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice Hall, 1988
- ❑ L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, 1990
- ❑ T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An Efficient Data Clustering Method for Very Large Databases. SIGMOD'96
- ❑ S. Guha, R. Rastogi, and K. Shim. Cure: An Efficient Clustering Algorithm for Large Databases. SIGMOD'98
- ❑ G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. *COMPUTER*, 32(8): 68-75, 1999.
- ❑ Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011 (Chap. 10)
- ❑ C. K. Reddy and B. Vinzamuri. A Survey of Partitional and Hierarchical Clustering Algorithms, in (Chap. 4) Aggarwal and Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press, 2014
- ❑ M. J. Zaki and W. Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge Univ. Press, 2014