



National Technical
University of Ukraine
"Igor Sikorsky
Kyiv Polytechnic Institute"



Institute of
Physics and
Technology

Intellectual Data Analysis

Practice 3: Unsupervised Learning

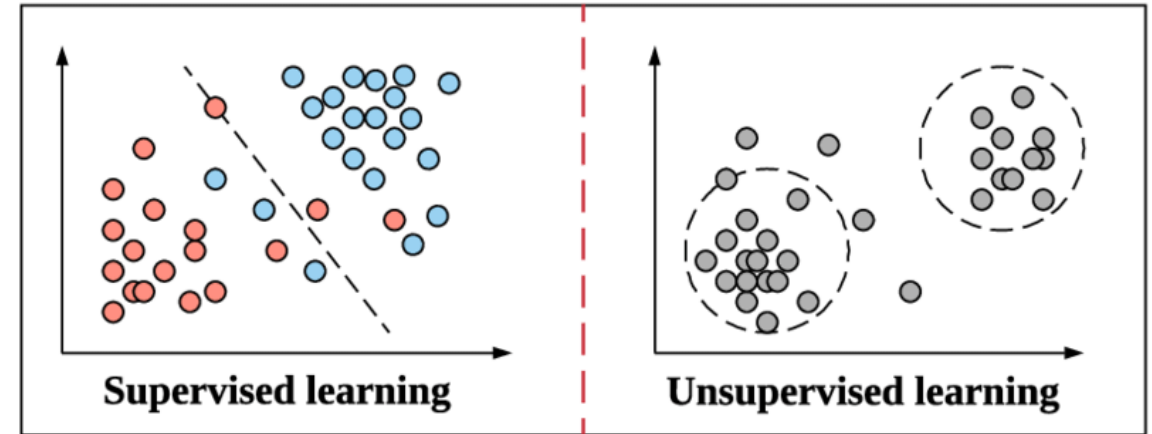
Dr. Nataliya K. Sakhnenko

Please review Lecture 4 before this practical

Recap: Unsupervised Learning

Supervised learning: discover patterns in the data that relate data attributes with a target (class) attribute. These patterns are then utilized to predict the values of the target attribute in future data instances.

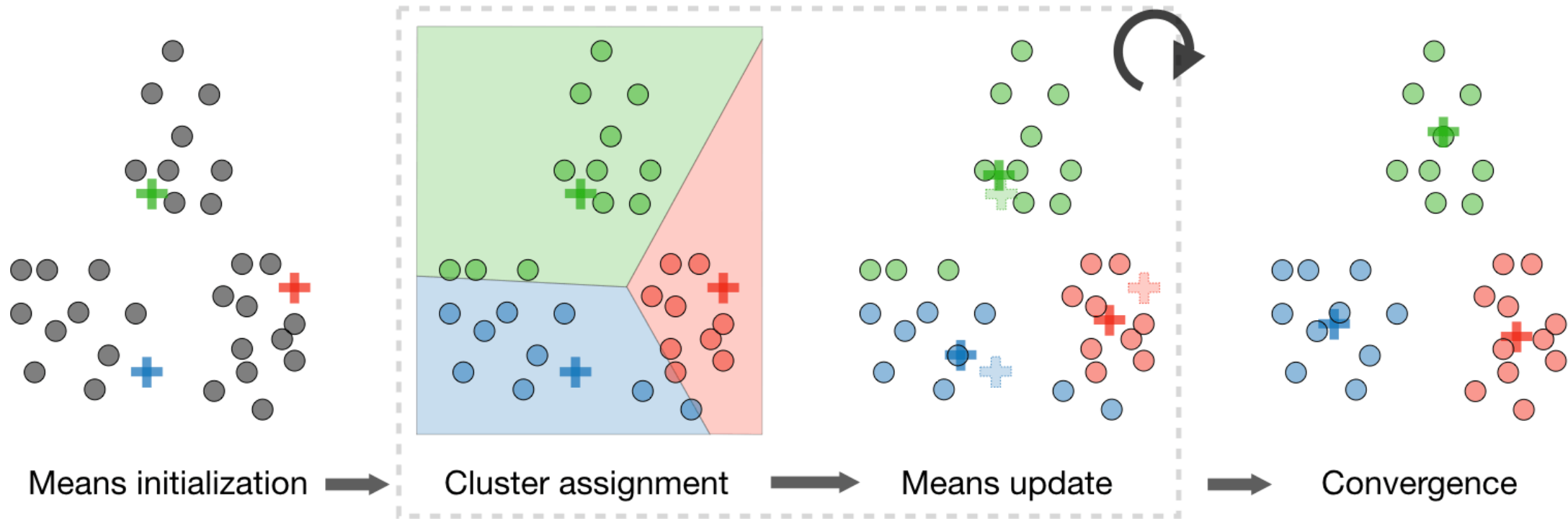
Unsupervised learning: The data have no target attribute. We want to explore the data to find some intrinsic structures in them. The goal of unsupervised learning is to find hidden patterns in unlabeled data .



Supervise Learning		Unsupervised Learning	
Regression	Classification	Clustering	Dimensionality reduction

K-means Clustering

K clusters

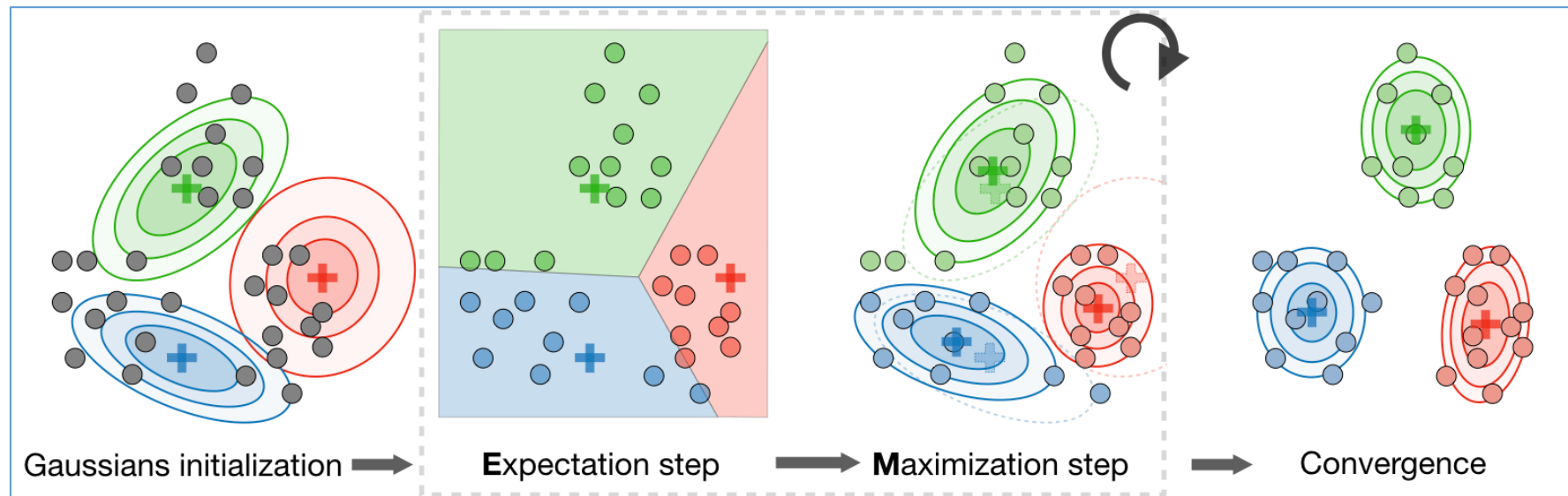


The K-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion:

$$J(c, \mu) = \frac{1}{M} \sum_{i=1}^M \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

centroids

EM for Gaussian Mixtures (2D)



Initialization:

- Initialize K clusters: C_1, \dots, C_K
- Parameters for each cluster j : (μ_j, Σ_j) , prior $P(C_j)$

Iteration steps:

1. Expectation step (E-step):

Estimate the probability that data point x_i belongs to cluster j :

$$p(C_j \mid x_i)$$

2. Maximization step (M-step):

Re-estimate cluster parameters for each j :

$$(\mu_j, \Sigma_j), \quad P(C_j)$$

Input:

$$\{x_i\}_{i=1}^M$$

Output:

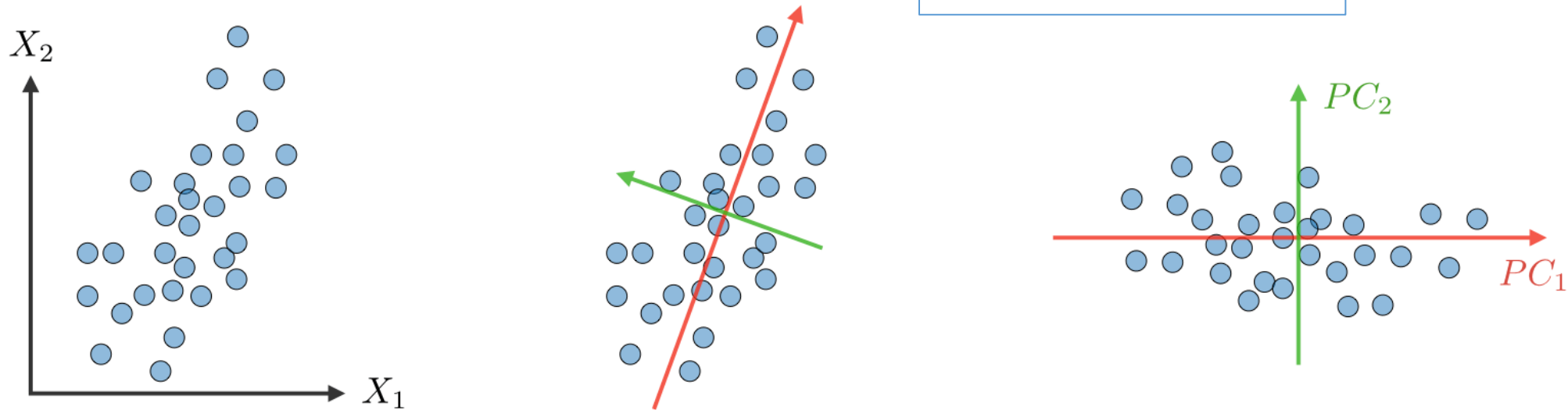
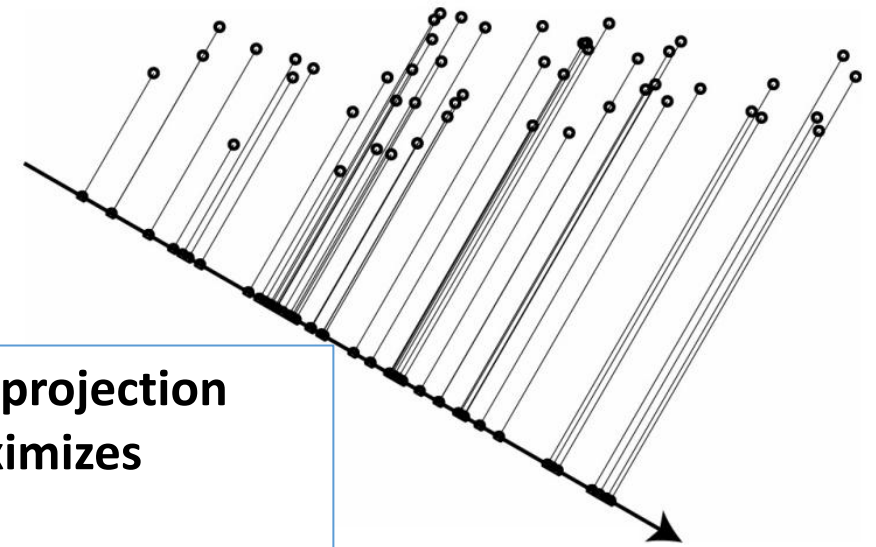
$(\mu_j, \Sigma_j), P(C_j)$ for each cluster j

$$p(C_j \mid x_i)$$

Principal Component Analysis (PCA): 2D case, intuition

PCA is the most popular dimensionality reduction algorithm. First it identifies the hyperplane that lies closest to the data, and then it projects the data onto it.


It is essential rotating the coordinate axes so higher-variance come first



Data in feature space → Find principal components → Data in **principal components** space

Covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \dots & \sigma_{1n} \\ \dots & \dots & \dots \\ \sigma_{1n} & \dots & \sigma_n^2 \end{pmatrix}$$

$$\sigma_{jk} = \frac{1}{m-1} \sum_{i=1}^m (x_j^{(i)} - \mu_j)(x_k^{(i)} - \mu_k)$$


For standardized data:

$$\Sigma = \frac{1}{m-1} X^T X$$

Principal axes are eigenvectors of matrix Σ

$$\Sigma \vec{v} = \lambda \vec{v}, \quad \det(\Sigma - \lambda E) = 0$$

λ are eigenvalues

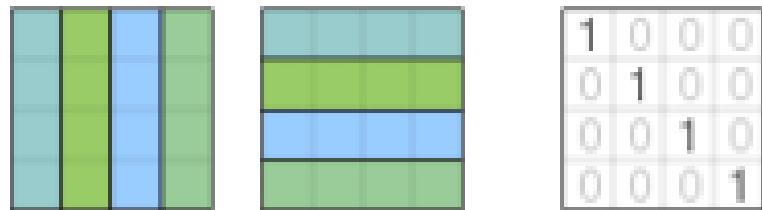
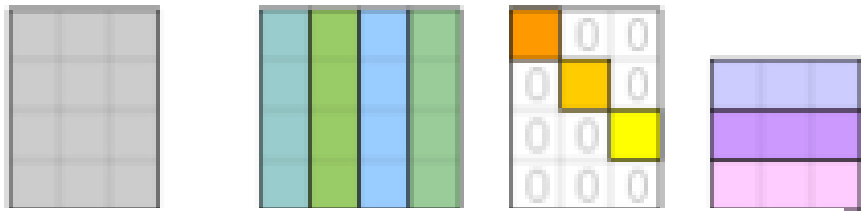
\vec{v} are eigenvectors

Proof: https://en.wikipedia.org/wiki/Rayleigh_quotient

PCA and SVD decomposition

Singular value decomposition (SVD) is standard matrix factorization technique.

$$X_{m \times n} = U_{m \times n} \cdot D_{n \times n} \cdot V_{n \times m}^T$$



$$U U^* = I_m$$

$$V V^* = I_n$$

Columns of U matrix are eigenvectors of XX^T

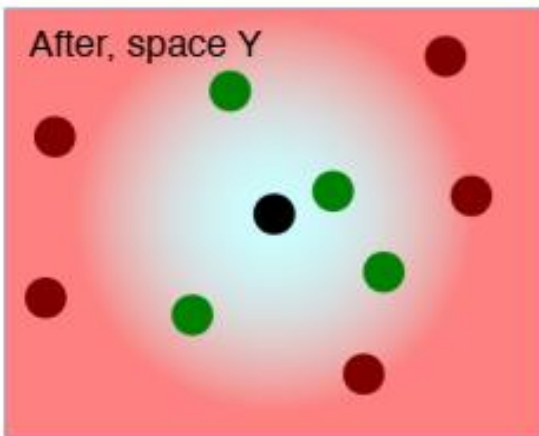
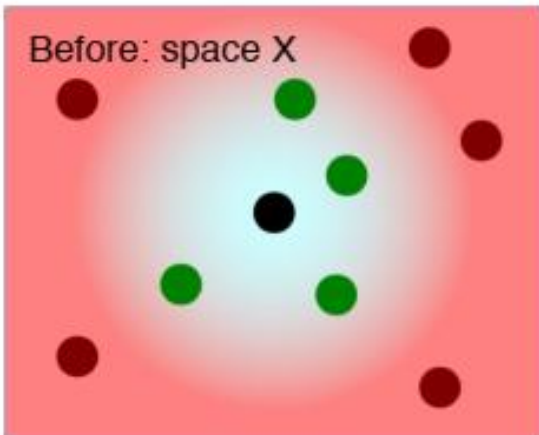
Columns of V matrix are eigenvectors of $X^T X$

Diagonal elements of D matrix are known as singular values of X

$$D = \begin{pmatrix} s_1 & 0 & & \\ 0 & s_2 & & \\ & & \ddots & \\ & & & \dots \end{pmatrix}, s_1 \geq s_2 \dots$$

Stochastic neighbor embedding (SNE)

It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.



$$p(j|i) = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma^2)}$$

Probabilistic input neighborhood

σ is hyperparameter

$$q(j|i) = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}$$

Probabilistic output neighborhood

t-distributed SNE (t-SNE)

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2 / 2\sigma_i^2)}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

Probabilistic **input** neighborhood:

Probability to be picked as a neighbor in space X (input coordinates)

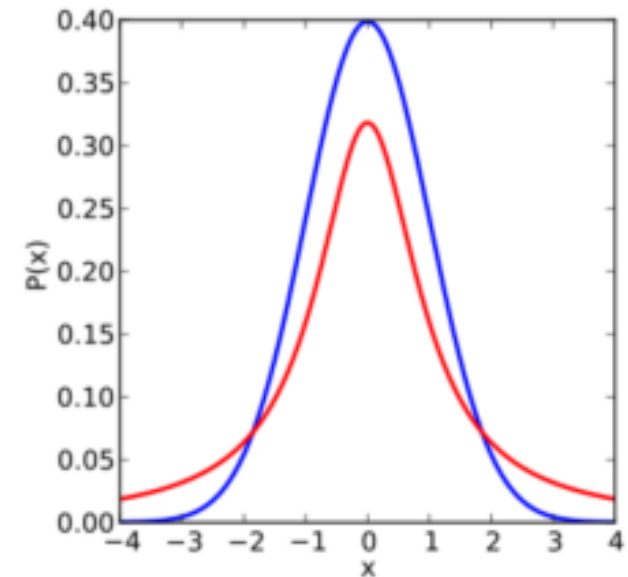
$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l} (1 + ||y_k - y_l||^2)^{-1}}$$

Probabilistic **output** neighborhood:

Probability to be picked as a neighbor in space Y (display coordinates)

$$\text{Cost} = KL(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Visualizing Data using t-SNE,
2008, L.Maaten&G.Hinton



Blue: normal distribution
Red: t-distribution (heavy-tailed)