TÉCNICO LISBOA

LEIC-T 2024/2025
Aprendizagem - Machine Learning
Homework IV

Group 36
Francisco Uva – 106340
Pedro Pais - 107482

I) Clustering

i)

E-Step

a) Likelihood

p(x1|c1=1) = 0.15915494309

p(x2|c1=1) = 0.00291502446

p(x3|c1=1) = 0.19956264

p(x1|c2=1) = 0.00291502446

p(x2|c2=1) = 0.15915494309

p(x3|c2=1) = 0.015340599734

b) Joint Distribution

p(x1,c1=1) = 0.09549296585

p(x2,c1=1) = 0.00174901467

p(x3,c1=1) = 0.119737584

p(x1,c2=1) = 0.00116600978

p(x2,c2=1) = 0.06366197723

p(x3,c2=1) = 0.06136239893

c) Data

p(x1) = 0.09665897563

p(x2) = 0.0654109919

p(x3) = 0.18109998293

d) Posterior Probability

$\gamma(c_{11}) = p(c1=1|x1) = 0.987936870029120$

$\gamma(c_{21}) = p(c1=1|x2) = 0.026738849661063$

$\gamma(c_{31}) = p(c1=1|x3) = 0.6611683921$

Group 36
Francisco Uva – 106340
Pedro Pais - 107482

$\gamma(c_{12}) = p(c2=1|x1) = 0.01206312911$

$\gamma(c_{22}) = p(c2=1|x2) = 0.97326115047$

$\gamma(c_{32}) = p(c2=1|x3) = 0.3383161078$

M-Step

$N_1 = 1.675844$

$N_2 = 1.324156$

New Means

$$\mu 1=\frac{1}{1.675844}*\begin{pmatrix}0.987936870029120\\0.987936870029120\end{pmatrix}+\begin{pmatrix}-0.026738849661063\\-0.026738849661063\end{pmatrix}+\begin{pmatrix}0.3383161078\\0.36364261406\end{pmatrix}$$

$$\mu 1=\begin{pmatrix}0.77082480904\\0.79055123782\end{pmatrix}$$

$$\mu 2=\frac{1}{1.324156}*\begin{pmatrix}0.01206312911\\0.01206312911\end{pmatrix}+\begin{pmatrix}-0.97326115047\\-0.97326115047\end{pmatrix}+\begin{pmatrix}0.16941580539\\0.18635738592\end{pmatrix}$$

$$\mu 2=\begin{pmatrix}-0.59795241667\\-0.58515816741\end{pmatrix}$$

New Covariance matrices

$$\sum_1=\frac{1}{1.675844}*\begin{pmatrix}0.5090268 & 0.49023896\\0.49023896 & 0.4727479\end{pmatrix}=\begin{pmatrix}0.303744 & 0.2925325\\0.2925325 & 0.282095\end{pmatrix}$$

$$\sum_2=\frac{1}{1.324156}*\begin{pmatrix}0.281943 & 0.290533\\0.290533 & 0.2994324\end{pmatrix}=\begin{pmatrix}0.21292 & 0.21941\\0.21941 & 0.226131\end{pmatrix}$$

Mixing Parameter equal to Nk/N

$\pi_1 = \frac{N1}{3} = 0.55861$

$\pi_2 = \frac{N2}{3} = 0.44138$

Group 36
Francisco Uva – 106340
Pedro Pais - 107482

ii)

$\gamma(c_{11}) = p(c1=1|x1) = 0.987936870029120$ x1 ∈ C1

$\gamma(c_{21}) = p(c1=1|x2) = 0.026738849661063$ x2 ∈ C2

$\gamma(c_{31}) = p(c1=1|x3) = 0.6611683921$ x3 ∈ C3

$a(x1) = |1 - 0.5| + |1 - 0.55| = 0.5 + 0.45 = 0.95$

$b(x1) = |1 - (-1)| + |1 - (-1)| = 2 + 2 = 4$

$s(x1) = 1 - a(x1) / b(x1) = 1 - 0.95 / 4 = 0.7625$

$a(x3) = 0.95$

$b(x3) = |0.5 - (-1)| + |0.55 - (-1)| = 1.5 + 1.55 = 3.05$

$s(x3) = 1 - a(x3) / b(x3) = 1 - 0.95 / 3.05 = 0.6885$

$s(C1) = (s(x1) + s(x3)) / 2 = 0.7255$

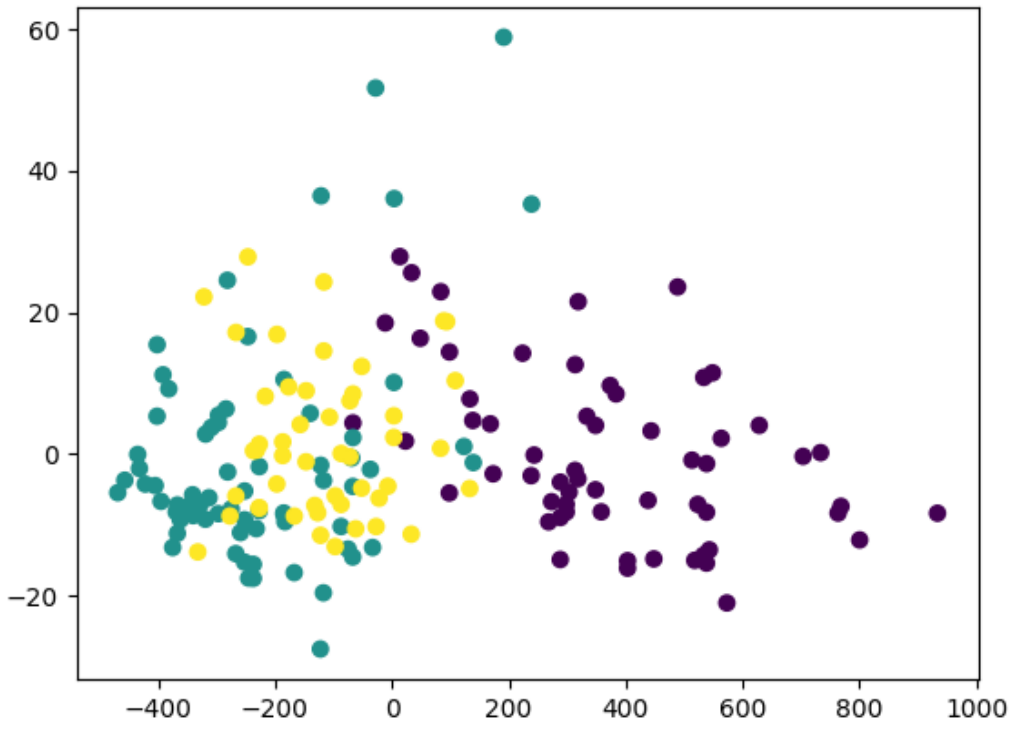
II) Software Experiments

a) Based on the silhouette scores obtained for each clustering method, it is evident that K-Means outperforms EM Clustering on the wine dataset. With a silhouette score of 0.5711, K-Means demonstrates a stronger clustering structure, where data points are more cohesive within clusters and more distinct from points in other clusters. In contrast, the EM Clustering method yields a silhouette score of 0.2833, indicating less

Group 36
Francisco Uva – 106340
Pedro Pais - 107482

defined clusters with increased overlap between clusters. Therefore, K-Means is the preferred method for clustering this dataset, as it provides clearer and more well-separated groupings, making it a better choice for this analysis.

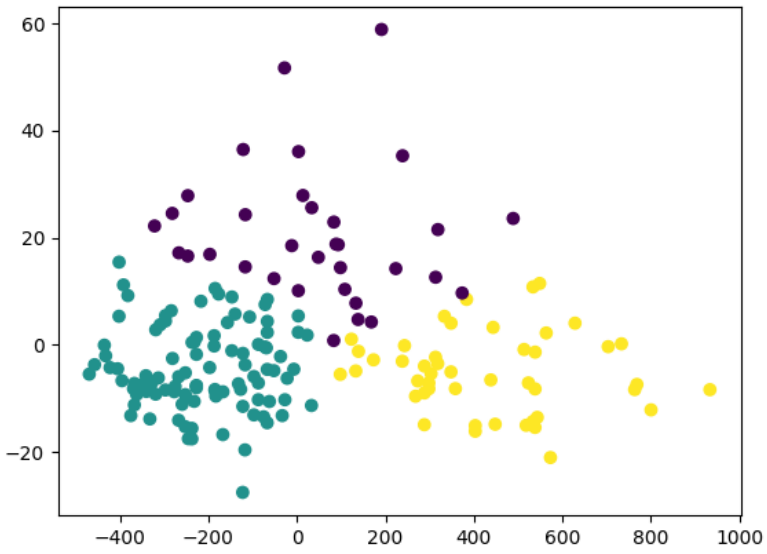
b) The scatter plot of the wine dataset, reduced to two principal components, shows a degree of separation among the three classes, though there is significant overlap between clusters, suggesting that the classes cannot be fully separated with just two principal components.



c) For the PCA-reduced wine dataset, **K-Means** clustering achieves a silhouette score of **0.5723**, while **EM Clustering** has a lower silhouette score of **0.2623**. This indicates that K-Means forms more well-defined clusters in the reduced data space. The silhouette scores differ from those in the original dataset (question a) because PCA reduces dimensionality, potentially losing some information and altering the

Group 36
Francisco Uva – 106340
Pedro Pais - 107482

data's structure, which affects the clustering outcome and compactness.

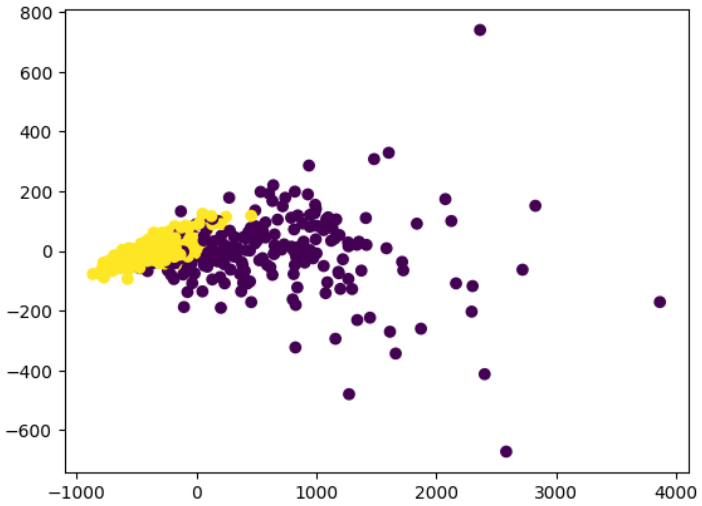


d) K-Means clustering on the original breast cancer dataset has a silhouette score of **0.6972**, outperforming EM Clustering, which has a score of **0.4963**. After PCA, K-Means achieves a slightly higher silhouette score of **0.6984**, while EM Clustering improves to **0.5262**. This indicates that PCA enhances the clustering quality for both methods, though **K-Means (0.6984)** still performs better than **EM Clustering (0.5262)** on the reduced data.

(Original breast cancer dataset)

LEIC-T 2024/2025
Aprendizagem - Machine Learning
Homework IV

Group 36
Francisco Uva – 106340
Pedro Pais - 107482



(after performing PCA)