**I. Pen-and-paper**

1. To compute the solution given by the EM Clustering algorithm we need three major steps. After going through all the steps, we will have computed the clustering solutions given by one iteration of the algorithm.
2. **Initialization**

Priors:

Multivariate normal distributions likelihoods: 



1. **E-Step:** Assign each point to the cluster with the higher normalized posterior

Notes:

Bayes rule:

Likelihood:

Joint Probability:

Normalized Posterior:

|  | **Cluster 1** | **Cluster 2** |
| --- | --- | --- |
| **Likelihood** | 0.15915494 | 4.71938976 |
| **Joint Probability** | 0.15915494 | 1.41581693 |
| **Normalized Posterior:** | **1** | 1.2708343 |

is assigned to **cluster 1**

|  | **Cluster 1** | **Cluster 2** |
| --- | --- | --- |
| **Likelihood** | 2.23908996 | 1.56736297 |
| **Joint Probability** | 0.03978874 | 0.01193662 |
| **Normalized Posterior:** | 1.31307093 | **1** |

is assigned to **cluster 2**

|  | **Cluster 1** | **Cluster 2** |
| --- | --- | --- |
| **Likelihood** | 0.00023928 | 4.91032009 |
| **Joint Probability** | 0.0001675 | 1.47309603 |
| **Normalized Posterior:** | **0.99128185** | 0.00871815 |

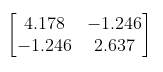
is assigned to **cluster 1**

|  | **Cluster 1** | **Cluster 2** |
| --- | --- | --- |
| **Likelihood** | 7.22562324 | 1.40683026 |
| **Joint Probability** | 5.05793627 | 4.22049078 |
| **Normalized Posterior:** | **0.92298354** | 0.07701646 |

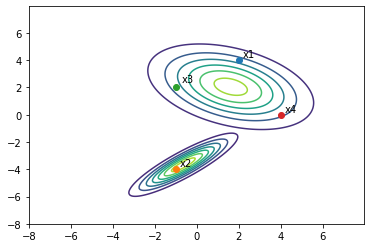
is assigned to **cluster 1**

1. **M-Step:** Update the cluster parameters and the priors

Notes:

* Cluster 1
* Cluster 2
* Priors

Sketch of the clustering solution:



1. **Silhouette of an observation**

**Silhouette of a cluster**  **Silhouette of the solution**

Using the Euclidean distance:

The silhouette values are between -1 and 1 and the higher the value the better, since smaller the distances between points from the same cluster - - and the greater the distances between points in different clusters - . This silhouette level is considered high, so we can conclude that this is a good model with cohesive and well-separated clusters.

1. The VC dimension measures the degrees of freedom of a two-class/binary classifier. A good approximation of that value is the number of parameters of the classifier. This will be used in i) and iii).

i) Between the input layer and the first hidden layer we have a weights matrix and a bias vector , so we have 30 parameters.

The same happens between the first and second hidden layers and between the second and the third ones. Knowing this, in total we now have parameters.

Between the third hidden layer and the output layer we have a weights matrix and abias vector , so we have 12 parameters.

There are no more parameters, so the total amount is parameters.

Given this, we can guess that the VC dimension for this classifier is 102.

ii) The maximum amount of different inputs we can have in this scenario is , where d is the data dimensionality. Given this fact, we can immediately say that for this classifier. Using a split of 3 possible outcomes for every feature, we can create a decision tree with that has a leaf for all possible points. All points will be classified correctly independently from the dichotomy chosen. This proves that . Knowing this, we can conclude that for this decision tree.

iii) For this Bayesian classifier, the parameters we need to estimate are the priors and likelihoods.

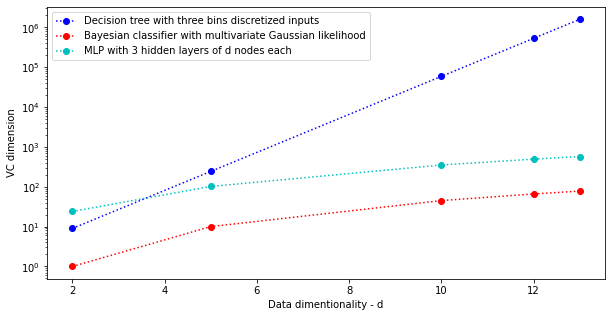
We have one parameter that is the prior of one class and, knowing this one, we can infer the other prior . Thus, we have 1 parameter for the prior.

For the likelihood , which we know is a multivariate normal, we need to know the mean vector and the covariance matrix. The mean vector is a vector, so here we have 5 parameters. The covariance matrix is a matrix but it is symmetric, so we only count the diagonal and upper diagonal part of the matrix, which gives us 15 parameters to estimate.

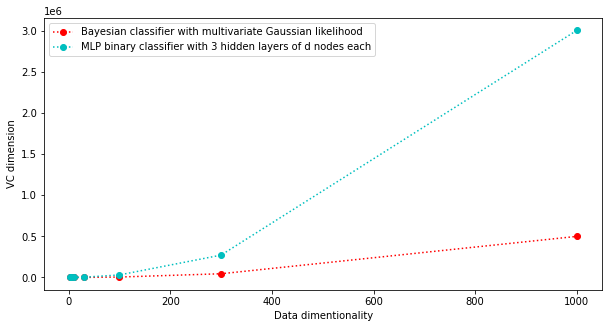
For the other likelihood , we need the same amount of parameters.

Hereupon, in total we have parameters and we can guess that the VC dimension for this classifier is 61.

1. The VC dimension grows exponentially with d for the Decision Tree, whilst it grows quadratically for the Bayesian classifier and MLP.

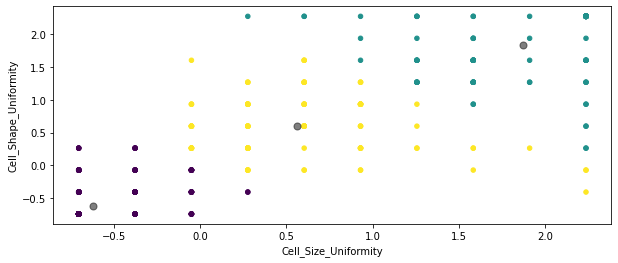


1. The VC dimension of the MLP grows higher and faster than the VC dimension of the Bayesian classifier when increasing the data dimensionality.



**II. Programming and critical analysis**

1. For the k-means clustering with k = 2 we have an error classification rate of 14.5, whereas, for the solution with k = 3 the ECR is 7.3. We can conclude that with 3 clusters we have more points from the same class assigned to the same cluster.
2. The k-means clustering with k = 2 has a silhouette coefficient of 0.5732, whereas, the solution with k = 3 has a silhouette of 0.5578. The higher the silhouette, the more cohesive and well-separated the clusters are, however, for these two models, the difference is not significant. The silhouettes of the models are both moderately high, so we may say that they are both good models in terms of internal measures.



1. The silhouette of this model is high (0.7064), which means that the clusters are compact and well-separated. We can observe that in the plot above: the clusters are quite distinguishable and the points assigned to the same cluster are relatively close to each other.

Also, when using feature selection, the results are a lot better than when using all features to perform the k-means. We can see that by comparing the silhouette coefficients of the k-means model with k = 3 from exercise 4 (0.5578) and the silhouette of this new solution.

In conclusion, this is a good model for clustering this dataset.

**III. APPENDIX**

Paste your programming code here using Consolas 9pt or 10pt.

Use **highlighting** or colored text to facilitate the analysis by your faculty hosts.

**END**