Problem 1 – Cluster a real data set– Extra credit  For this problem, submit your code through the Assignments page link.

The data files are in ascii, 1 point per line. If your clustering algorithm takes too long, “decimate” the data by taking every 5-th point. If this is still too slow, cluster a random subset of the data.

Choose the number of clusters $K$, or a range for it (suggested range $K = 2, \ldots, 10$).

Describe clearly what you did. What clustering algorithm did you use? If EM, what model? What parameters were estimated? If Mean-Shift, what was $h$, and what was the kernel? How did you initialize (wherever it applies)? Did you do any post-processing?

How did you choose $K$? If you used a plot or a score, include them. If you used BIC or AIC, give the formula and the values you obtained (or graph them).

You will get points for:

- How carefully you analyzed the data
- How completely and precisely you describe what you did. I.e could someone else repeat the experiment from your description?
- How well you motivate what you did
- How many different clustering algorithms or models you used
- Legibility of graphs and figures
- Bonus points: clusterings that “look good” when the results are unveiled
- Negative points: blatantly incorrect results or statements.

You will get full credit of 12 points if:

1. At least one clustering algorithm used is implemented by you
2. Your descriptions of what you did are complete enough to be reproducible (about 1 page, but you can take more if you need to)
3. You describe a reasonable way to choose $K$
4. The results are displayed legibly
5. Give a well thought answer to the question “Should I believe these clusters I obtained are “good” (“true”)?” You can do some investigation to answer this question, like cluster multiple times and compare the results, display the clusters, display or compare characteristics of the clusters, etc?

Upload your results on catalyst in an ascii file, with one row for each point containing

<table>
<thead>
<tr>
<th>point_number</th>
<th>cluster_label</th>
</tr>
</thead>
</table>

where point_number is the number of the point in the data set, starting with 0. If you cluster the whole data set, the point_number will range from 0 to 225749. If you use other data, the point numbers will tell us which data did you use. cluster_label should be a positive integer; you are not required to number the clusters 1, 2, ..., K; any set of K distinct integers > 0 are acceptable.

If you work on windows, enter plain text and not .rtf or other formats. We will test your results and unveil them with the solutions. To summarize:

- If you do everything in a way that shows good understanding of clustering, and describe it clearly, you get full credit no matter what clustering results you have (we do not subtract for “bad” clusterings, only for bad statistics.

- If you use more than one algorithm to cluster, you get bonus points in proportion to the added effort.

- If you enter your clustering results in the correct format (otherwise we can’t test them) and the clustering is reasonably good, you get bonus points. If you don’t enter the clustering results at all, nothing is subtracted.

- If you show some good analytical thinking or working knowledge of stats and probability in the comments you make, you get bonus points. In other words, anything that shows thinking in addition to mechanically using a method (or more) is rewarded.
Problem 2 – Likelihood ratio classifiers in one dimension

The data for this problem is in \texttt{hw7-toy-classif1d.dat} and has \( n = 10 \) points, with \( x \in \mathbb{C} \) on each row.

\textbf{a.} Fit a Likelihood Ratio classifier where \( f_{X|C} = \text{Normal}(\mu_c, \sigma^2_C) \), i.e. where each class has its own variance parameter. Find the values for \( \mu_0, \sigma^2_0, P_C(1) \).

What are the decision regions?

\textbf{b.} Fit a LDA classifier to this data set; in other words, find the Likelihood Ratio classifier for which \( f_{X|C} = \text{Normal}(\mu_C, \sigma^2) \) for \( C = 0, 1 \). Find the values for \( \mu_0, \sigma^2, P_C(1) \) from the data by Maximum Likelihood. Then write the expression of the linear classifier \( \phi(x) \) that results.

\textbf{c.} Plot the probability \( P_{C|X} \) for \( X \in [-10, 10] \) and mark the decision regions.

Plot the function \( \phi(x) \) for \( x \in [-10, 10] \) and mark the decision regions on this graph as well. Comment on what you observe.

Problem 3 – Classification in 2 dimensions

\textbf{a. Warmup} Reproduce Figure 16.3 in the notes.

\textbf{b.} Assume that the class densities are given by \( f_{X|C} = \text{Normal}(\mu_c, I) \) (in other words, we consider \( \sigma^2 \) fixed to the unit matrix and do not estimate it). Estimate from \texttt{hw7-toy-classif2d.dat} the parameters \( \mu_{0,1} \in \mathbb{R}^2 \) and write the expression of the classifier \( \phi(x) \) for this data set. The data is represented below

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{example_plot.png}
\end{figure}

\textbf{c.} Plot the decision regions or decision boundary for this classifier.

\textbf{d.} What are the predicted class values for

\[
x = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ -1 \\ 3 \end{bmatrix}
\]

Which of these points is classified with higher confidence?

\textbf{e.} Draw a decision tree that classifies this data set correctly. What class labels does this decision tree assign to the points in \textbf{d}?

\textbf{f.} Will every decision tree on this data set assign the same labels? Prove or disprove (a counter-example is sufficient to disprove).

\textbf{g.} Draw the decision regions of the 1-nearest neighbor classifier for the same data set. What class labels does this classifier assign to the points in \textbf{d}?

\textbf{h.} We say that data set \( D \) in 2 dimensions is \textit{linearly separable} if there exists a line separating the class 1 examples from the class 0 examples. Usually,
whenever $D$ is linearly separable, there are an infinity of lines that can separate the classes.

Is the data set hw7-toy-classif2d.dat linearly separable?

i. The data set in hw7-toy-classif2d-separable.dat is linearly separable.

In this case, the Maximum Margin linear classifier is the line that stays farthest away from both classes, i.e the line separating the dataset correctly with the property that

$$\min_{x^i \in D} \text{distance}(x^i, \text{line})$$

is maximized over all such lines

Draw the maximum margin linear classifier for this data set.

**Problem 4 – Naive Bayes Model**

Estimate the Naive Bayes classifier from the data set hw7-naive-bayes.dat (each row contains $x_1, x_2, x_3, c$).

Classify the points $x = (0, 0, 0), (1, 1, 1), (0, 1, 0)$. Which point is classified with highest confidence?