

Exercises

§8.5: Clustering and k-Means

Python: There is an optional starter script that supports Python-based problem(s) on this sheet. Use it as scaffolding: read the comments, complete the lines marked **TODO**, and compare the output with the mathematics requested in the problem(s).

Problem 1 For one-dimensional observations $x = (0, 1, 4, 5)$ and fixed centroids $\mu_1 = 0$ and $\mu_2 = 5$, (a) assign each observation to its nearest centroid, (b) compute the current k -means objective, (c) update the two centroids, (d) compute the objective after the centroid update.

Problem 2 Let a fixed cluster contain vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$. Starting from $J(\boldsymbol{\mu}) = \sum_{p=1}^n \|\mathbf{x}_p - \boldsymbol{\mu}\|^2$, derive $\boldsymbol{\mu} = \frac{1}{n} \sum_{p=1}^n \mathbf{x}_p$. Why is this a minimum?

Problem 3 Let two centroids be $\boldsymbol{\mu}_1 = (0, 0)^T$ and $\boldsymbol{\mu}_2 = (4, 2)^T$. Find the equation of the boundary containing points equally distant from both centroids. Verify that: (a) the boundary passes through the midpoint, (b) its normal vector is parallel to $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

Problem 4 Show that nearest-centroid assignment is equivalent to $\hat{c} = \arg \max_k [2\boldsymbol{\mu}_k^T \mathbf{x} - \|\boldsymbol{\mu}_k\|^2]$. Why does this imply that the boundaries between fixed centroids are linear?

Problem 5 Suppose an assignment step is performed with fixed centroids. Explain carefully why the k -means objective cannot increase. Then explain why the centroid-update step also cannot increase the objective. Why do these facts still not prove that k -means finds the global minimum?

Problem 6 Suppose the optimal k -means objective values are $J_1^*, J_2^*, \dots$. Prove that $J_{K+1}^* \leq J_K^*$. Why does this mean we cannot choose the number of clusters merely by minimizing the training objective over K ?

Problem 7 Consider

```
x = np.array([[1.],
              [2.],
              [3.],
              [8.],
              [9.],
              [10.]])
```

with initial centroids

```
centroids = np.array([[2.],
                      [8.]])
```

Use Python to perform: (a) one assignment step, (b) one centroid-update step, (c) compute the objective before and after the centroid update, (d) verify that the next assignment step makes no changes.

Problem 8 A dataset has two features. Feature 1 ranges approximately from 0 to 100,000. Feature 2 ranges approximately from 0 to 1. Explain why unscaled Euclidean k -means may be dominated by feature 1. What preprocessing could be considered? Why might that preprocessing nevertheless be inappropriate in some applications?