# Exam 3

Benny Chen

December 13, 2023

## 1 Question 1

(a) The "Top 10 Data Mining Algorithms" article says about k-means "The greedy-descent nature of k-means on a non-convex cost also implies that the convergence is only to a local optimum, and indeed the algorithm is typically quite sensitive to the initial centroid locations... **The local minima problem can be countered to some extent by running the algorithm multiple times with different initial centroids**."
Please explain why the suggestion in boldface is a potential solution to the local maximum problem.

(b) The soft margin support vector machine solves the following optimization problem:

$$\begin{aligned} \min \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(w'x_i - b) \geq 1 - \xi_i \quad \forall i = 1 \dots n \\ & \xi_i \geq 0 \end{aligned}$$

What is the purpose of the first term in the objective function?
What does $\epsilon_i$ measure?
What role does parameter C play? Explain qualitatively what happens as we decrease C.

### 1.1 Answer 1(a)

K-Means with diffrent centroids would result in different local maxima so having different initial centroids around the local maxima would also result in something different.

### 1.2 Answer 1(b)

The first term in the objective function is $\frac{1}{2}||w||^2$. It is used to maximize the margin between the two classes. By maximizing the margin, it leads to generalization of the model which reduces the chance of overfitting.

$\epsilon_i$ measures the distance between the $i^{th}$ data point and the hyperplane. It is used to measure the error of the model.
Parameter C is a regularization parameter. It is used for maximizing the margin and minimizing the error. As C decreases, the margin increases and the error decreases. This is because the model is more focused on maximizing the margin and less focused on minimizing the error.

# 2 Question 2

Consider the following set of one-dimensional data points: {0.1, 0.2, 0.42, 0.5, 0.6, 0.8, 0.9}. Suppose we apply k-means clustering to obtain three clusters, A, B, and C. If the initial centroids are located at {0, 0.25, 0.6}, respectively.

(a) Show the cluster assignments and locations of the centroids in the first three iterations.

(b) Calculate the overall sum-of-squared errors (SSE) of the clustering after the third iteration.

## 2.1 Answer 2(a)

| Iter | 0.10 | 0.20 | 0.42 | 0.50 | 0.60 | 0.80 | 0.90 | A | B | C |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | — | — | — | — | — | — | — | 0.00 | 0.25 | 0.60 |
| 1 | A | B | B | C | C | C | C | 0.10 | 0.31 | 0.70 |
| 2 | A | A | B | B | C | C | C | 0.15 | 0.46 | 0.77 |
| 3 | A | A | B | B | B | C | C | 0.15 | 0.51 | 0.85 |

## 2.2 Answer 2(b)

Cluster A:

$$SSE_A = (0.10 - 0.15)^2 + (0.20 - 0.15)^2 = 0.005$$

Cluster B:

$$SSE_B = (0.42 - 0.51)^2 + (0.50 - 0.51)^2 + (0.60 - 0.51)^2 = 0.0163$$

Cluster C:

$$SSE_C = (0.80 - 0.85)^2 + (0.90 - 0.85)^2 = 0.005$$

Overall SSE:

$$SSE = SSE_A + SSE_B + SSE_C = 0.005 + 0.0163 + 0.005 = 0.0263$$