Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

[redacted]

QUESTION [redacted]

Given loss function:

For class c:

$$L(w_c, M_c) = \frac{1}{N_c} \sum_{x_n : y_n = c} (x_n - w_c)^T M_c (x_n - w_c) - \log |M_c|$$

Optimizing with respect to w_c:

The derivative with respect to w_c should be set to zero:

$$\frac{\partial L}{\partial w_c} = 0$$

$$\implies \frac{-1}{N_c} \sum_{x_n : y_n = c} M_c (x_n - w_c) = 0$$

This results in:

$$w_c = \frac{1}{N_c} \sum_{x_n : y_n = c} x_n$$

This is the mean of the data points from class c.

Optimizing with respect to M_c:

Optimizing with respect to the positive definite matrix M_c requires considering the term involving the matrix and the determinant term. The matrix derivative gives:

$$\frac{\partial L(w_c, M_c)}{\partial M_c} = \frac{1}{N_c} \sum_{x_n : y_n = c} (x_n - w_c)(x_n - w_c)^T - M_c^{-1}$$

Now for finding the value of M_c for optimization:

$$\frac{1}{N_c} \sum_{x_n : y_n = c} (x_n - w_c)(x_n - w_c)^T - M_c^{-1} = 0$$

$$\implies M_c^{-1} = \frac{1}{N_c} \sum_{x_n : y_n = c} (x_n - w_c)(x_n - w_c)^T$$

This is the covariance of the data points from class c when w_c is the mean.

Special Case: When M_c is an identity matrix.

For this case, the loss function for class c is:

$$L(w_c) = \frac{1}{N_c} \sum_{x_n : y_n = c} ||x_n - w_c||^2$$

This is a measure of the average squared Euclidean distance between the training examples of class c and w_c. In this scenario, the optimal w_c is simply the mean of the training examples from class c, as previously derived.

Given that the loss function measures the squared Euclidean distance to some point w_c (which is the centroid or mean of the class), the classifier's decision for a new data point x can be thought of as follows:

1. Compute the squared Euclidean distance between x and the mean of each class.
2. Assign x to the class whose mean is closest in terms of this squared distance.

This is a **Learning with Prototypes** (or Nearest Prototype Classifier). For each class, the centroid (mean of training samples) is computed and used as a representative point. To classify a new data point, the distance between the point and each class centroid is calculated. The point is then assigned to the class whose centroid is the closest.