1 Problem setup

Suppose we have a labeled training set with N independent and identically distributed (iid) samples:

$$\mathbf{X} = \{x_i\}_{i=1\cdots N}.$$

We assume that each x_i is an instance drawn from a known probability density family, $p(x|\theta)$, defined by parameters, θ.

In Problem 4, our goal is to train three classifiers using the labeled dataset. The class index is indicated by j in C_j:

$$x \in C_0 \sim p(x|\theta_0)$$

$$x \in C_1 \sim p(x|\theta_1)$$

$$x \in C_2 \sim p(x|\theta_2).$$

For each class, the data samples x_i are assumed to be normally distributed.

2 Posterior distribution

We need to determine the parameter θ_j that maximizes the posterior distribution $p(C_j|x)$. This will reveal the distribution of class predictions for a given input x. Since we assume the samples x are normally distributed, we can use a parametric approach.

Using Bayes' rule, we can express the posterior distribution as:

$$p(C_j|x) = \frac{p(x|C_j)p(C_j)}{p(x)},$$

where $p(x|C_j)$, $p(C_j)$, and $p(x)$ are the likelihood, prior, and evidence, respectively. As $p(x)$ is fixed and does not affect the maximization of the posterior $p(C_j|x)$, we can ignore it.

3 MLE and MAP

We can rewrite Bayes' rule without the evidence term:

$$p(C_j|x) = p(x|C_j)p(C_j).$$

This formula highlights the difference between MLE and MAP. MLE maximizes the likelihood $p(x|C_j)$, assuming a uniformly distributed prior $p(C_j)$. In contrast, MAP directly maximizes the posterior by optimizing both the likelihood and the prior.

For the likelihood $p(x|C_j)$, we can estimate the parameter θ_j using MLE. For MAP, we follow a similar procedure if the prior is known. If the prior is unknown, we can estimate the distribution by counting the number of samples per class:

$$p(C_j) = \frac{N_j}{N_{total}}.$$

Combining this, the posterior distribution for each class is:

$$p(C_j|x) = p(x|C_j)\frac{N_j}{N_{total}}.$$

Applying a log transform to the posterior, we have:

$$\log p(C_j|x) = \log p(x|C_j) + \log \frac{N_j}{N_{total}}.$$

In this formula, the prior is constant, which indicates that MAP and MLE have the same solution based on the first-order derivative.

Given that $p(x|\theta_j)$ is normally distributed, the discriminant function for class j is:

$$g_j(x) = -\frac{1}{2}\log 2\pi - \log \sigma_j - \frac{(x-\mu_j)^2}{2\sigma_j^2} + \log \frac{N_j}{N_{total}}$$

4 Further read

https://courses.cs.washington.edu/courses/cse312/20su/files/student_drive/7.5.pdf