Report

Homework 1

Goal: Develop models for 10-class classification problems with medium and large input space

Andrea Massignan

November 20, 2023

Project Overview

The goal is to provide two different solution for a 10-class classification problem.

0.1 Background

Multi-class classification involves assigning instances to one of several predefined classes. This task is fundamental to a multitude of domains, such as healthcare, finance, and telecommunications, where decision-making relies on discerning among multiple potential outcomes.

0.2 Motivation

The motivation behind this project lies in the exploration of multi-class classification across datasets with disparate input space dimensions. By examining two datasets—Dataset 1 with an input dimension of 100 and Dataset 2 with an input dimension of 1000—we aim to understand how models generalize and perform in scenarios with varied feature vector sizes.

0.3 Problem Statement

The challenge posed by datasets with different input space dimensions introduces complexities in model development. Strategies that effectively handle a smaller input space may not seamlessly scale to larger dimensions. Consequently, this project seeks to address the nuances and intricacies associated with multi-class classification when confronted with varying input space sizes.

0.4 Objectives

The primary objectives of this project are:

- Develop robust multi-class classification models for datasets with input space dimensions of 100 and 1000.
- Evaluate model performance on both training datasets and blind test datasets to assess generalization capabilities.
- Compare and analyze the results obtained from models trained on datasets with different input space dimensions.

0.5 Significance

The insights gained from this project extend beyond the realm of multi-class classification and contribute to the broader understanding of model adaptability to varying input dimensions. Such knowledge is instrumental in guiding the development of intelligent systems capable of handling diverse and dynamic data representations.

This project embarks on a journey to uncover the intricacies of multi-class classification within the context of datasets characterized by different input space dimensions. The subsequent sections will delve into the datasets, methodologies employed, experimental setups, and the anticipated outcomes, providing a comprehensive overview of the undertaken exploration.

1 Dataset

The datasets used in this project consists in training sets and blind test sets. The training sets are composed by 50000 samples, each one with 100 features for Dataset 1 and 1000 features for Dataset 2. The blind test sets are composed by 10000 samples, each one with 100 features for Dataset 1 and 1000 features for Dataset 2. The labels are 10, one for each class.

- `X_train`: 50000 samples for each dataset.
- `n_features`: 100 for Dataset 1 and 1000 for Dataset 2.

2 Data Exploration

The datasets, that are in a csv file, are loaded with an helper function called `load_data` that returns the training set in the X matrix and the labels in the Y vector.

Then the training set is split into `X_train` and `Y_train` with a test size of 0.2 and different random states.