MSc in Computer Science
at University of Milan

Statistical Methods for Machine Learning
Kernelized Linear Predictor

course held by **Nicoló Cesa-Bianchi**

Email:
davide.cologni@studenti.unimi.it

Created by:
Davide Cologni
mat. 09732A

Academic year of 2023/2024

Contents

1 introduction

2 Dataset analysis and preprocessing

2.1 Dataset description

The provided dataset contains 10000 points with 10 features named from $x1$ to $x10$ and a label coloumn named y.
All the feature are floating point values and the dataset is well formed (in the sense that there are no missing values).
The label coloumn contains values that are either -1 or $+1$.
I collect the major statistics from each feature in the dataset: mean std min max It's clear that the different features are not normalized and follow different probability distribution.

2.2 Normalization

(advantage from the theorical prospective) (avoid data leakage)

2.3 Standardization

(advantage from the theorical prospective) (avoid data leakage)

2.4 Outliers removal

another approach I tried is removing the outliers from the dataset using the Z-score methods removing 265 outliers (recall that the dataset has size 10_000).
$I calculate the score$ Z = (x - μ)/σ for each value where μ is the mean and σ is the variance of the data on the feature then we remove all the points with a Z-score greater or equals than 3 in absolute value.
I tried training non-kernelized Perceptron and Pegasos over the modified dataset but the results shows that the dataset is already sufficently cleaned: in fact it affects the performance of the models in a minimum way with no significative changes, and even in some cases it is (even if only slightly) worsening

comparison data table

2.5 Feature correlation

I tried plotting the feature on the training set on both axis to spot correlation and I observed that the feature 2 and 5 have a linear correlation (with a negative coefficent) as the feature 5 and 9 (with positive coefficent).

plot images

One possibility in this case during the preprocessing of the data is to remove the correlated features and leave only one of them to avoid redundancy of the data.
I don't follow this approach because there is a sensibile noise in the correlation and removing some features can lead to also removing this noise that can encode important information on the model.

2.6 Feature expansion

To being able to express non-homogeneous linear separators (hyperplane that don't pass through the origin) we add a constant feature of value 1 to each point in the dataset.
Let $\vec(x)$ be any point in the dataset and $\vec(w)$ be the linear separator, if we define $x' = (\vec{x}, 1)$ we can define $w' = (\vec(w), c)$, in that way:

$$w'^T x' = (\vec(w)^T \vec(x) + c)$$