Pattern Recognition: Coursework I

Mohyeldin Aboualam
Imperial College London
SW7 2AZ
mohyeldin.aboualam15@imperial.ac.uk

Cao An Le
Imperial College London
SW7 2AZ
caoan.le15@imperial.ac.uk

Abstract

ok

1. Introduction

1.1. Dataset Description

The CUHK03 dataset contains 14096 images picturing 1.467 identities. Feature vectors are stored as rows in a 14096×2048 matrix, from which a training set, validation set and test set are produced[1]

1.2. Problem Formulation

In this study, we explore Distance Metrics Learning for the purpose of improving Machine Learning (ML) algorithms. Some supervised and unsupervised learning algorithms, such as k-nearest neighbors and k-means clustering, depend on distance calculations. Generally, finding an adequate metric is a data-dependent problem, where the goal is to learn one that assigns small distances to similar data points (and large distances to dissimilar examples).

We first start by performing an evaluation experiment on a baseline approach using k-nearest neighbors (KNN) algorithm with no modification on the features on the test set.

Then, we propose methods to improve the baseline using distance metric learning algorithms and repeat the testing procedure as means of assessing the performance of our approach.

In order to assess the performance of different approaches, we analyse the Cumulative Matching Characteristics top-K accuracy. In the KNN algorithm this correponds to the percentage of query images that were matched with correct (same label) gallery image within K nearest neighbours. Within the K nearest neighbours, the CMC top-k accuracy is:

$$A_{cck} = \begin{cases} 1, \\ 0 \end{cases}$$

The final CMC curve is computed by averaging the shifted step functions over all the queries.

2. Baseline Approach

In this section, we evaluate the performance of (metrics wise) of untouched feature representations, by conducting the experiment on the gallery set. The calculated distance in the KNN method depends on the metric used (e.g. euclidean, minkowski, mahalanobis). For our baseline approach, the standard euclidean L2 distance is employed (see equation 1 for the Euclidean distance d between vectors **p** and **q** where **p** and $\mathbf{q} \in R^n$).

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{1}$$

3. Alternative Approaches

To improve the metrics values obtained from the baseline approach, several ideas are proposed. It is important to note that now, the

Different metrics of the KNN algorithm other than the Euclidean distance can be analysed and compared. A different metric will place different weights on the neighbouring points which could therefore be different. For instance, we first study the effect of using the mahalanobis metric, which is defined as:

$$d(p,q) = \sqrt{(\mathbf{p} - \mathbf{q})^T \, \mathbf{S}^{-1} \, (\mathbf{p} \textbf{ - } \mathbf{q}(2)}$$

where **S** is the covariance matrix between the two vectors **p** and **q**. Thus, the Mahalanobis metric accounts for how correlated the variables are to one another and avoids the redundant information in the Euclidean distance calculation. In fact, computing the Mahalanobis distance corresponds to computing the Euclidean distance on a PCA-transformed data[2]. Another way to think of this is that the

[1] A standard 60-20-20 split was performed.

[2] where the data is "squished" around the principal components

Mahalanobis distance leads to an elliptic decision boundary in a 2-D case, as opposed to a circular boundary in the Euclidean distance case. Linear Discriminant Analysis can be used to maximise class separability, which should result in In search of reducing the feature dimensions, Principal Component Analysis can be

References