Data-Driven Metric Learning for Healthcare Analytics

Project Description. Selecting appropriate similarity measure (Euclidean distance is one of many possible choices) is fundamental to many learning algorithms such as k-means and nearest neighbor searches. However, choosing such a similarity is highly problem-specific, and ultimately dictates the success or failure of the learning algorithm.

For instance, suppose a clinician wants to analyze the characteristics of over thousands patients with some diabetic or obesity condition. She first needs to group similar patients together. The patients are described, e.g., by their physical activity level throughout several days, and the clinician is able to tell for pairs of patients if they are similar or not. However, there are over thousands of patient to group, and it is not easy to tell exactly which factors are at play in her judgment, thus making it difficult to automatize the grouping with standard clustering techniques, like k-means.

We propose to use a metric learning approach. We show the clinician few pairs of patients using some graphical representation like bar charts, and we ask her to decide whether she perceives them as similar or not. The metric learning technique will learn the similarity measure (from the data) to apply for all patients based on these few pairwise judgments. This learned similarity measure will then be used to group the 1000s patients automatically in a way matching the clinician judgment.

In this project, we will study several approaches to learn the similarity functions by exploiting the distance information that is intrinsically available in many learning settings, i.e.,

1. In semi-supervised clustering, points are constrained to be either similar or dissimilar.
2. In fully supervised settings, constraints can be inferred so that points in the same class have smaller distances to each other than to points in different classes.

We will explore some metric learning algorithms which can be sufficiently flexible to support the variety of constraints realized across different learning paradigms. These algorithms should be able to learn a similarity function that generalizes well to unseen test data.

eHealth domain. We will apply these techniques in the eHealth domain, to support clinician in interactively grouping patients with certain behaviors based on their physical activity data.

Disclaimer and Learning Opportunities: The interns need not have prior knowledge about the concepts mentioned above. They will enhance their programming skills in Python and acquire new knowledge in metric learning; supervised and unsupervised learning; and Health analytics.

Duties/Activities: The intern will run and test different instances of the machine learning code on real data. The code in python will be provided.

Required Skills: Python

Preferred Intern Academic Level: undergrad / B.Sc.

Expected Team Size: 2 students

Main Mentor: Dr. Abdelkader Baggag <abaggag@hbku.edu.qa>

Co-Mentor: Dr. Michael Aupetit <maupetit@hbku.edu.qa>