**Unsupervised Learning and Dimensionality Reduction**

# Tichakunda Mangono, tmangono3

# March 24th, 2019

**Introduction**

This paper will apply several unsupervised learning and dimensionality reduction algorithms to two separate datasets – one for **handwritten digit recognition** and another for **breast cancer diagnosis**. For unsupervised learning, k-means clustering (K-means) and the Expectation-Maximization (EM) algorithms will be explored while Principal Components Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections (RP), Forward Stepwise and Backward Elimination will be investigated. The main focus is to comparatively analyze. For each algorithm the report explores key algorithm performance in different settings given hyperparameters.

This analysis relies heavily on the Python libraries machine learning algorithms while using Pandas, Numpy, Matplotlib and Seaborn for data manipulation and visualization.

**Datasets**

The two datasets selected for this analysis are the following classification problems: i) **handwritten digit recognition** (referred to as “*digits*” for the rest of this report) – a labelled dataset of 1797 (8x8 pixel) images of handwritten digits with 10 classes (from 0 to 9), and ii) **breast cancer diagnosis** (referred to as “*diagnosis*” for the rest of this report) – a labelled dataset of 569 instances with 30 features computed from a digitized image of a fine needle aspirate of a breast mass, describing the characteristics of cell nuclei present in the image.

While the digits and diagnosis datasets are individually interesting for representing the major topics of optical character recognition and predictive disease diagnosis, they also represent two contrasting approaches of classifying data from images. The digits dataset represents raw, primary pixel-level information, while the diagnosis dataset is a set of carefully computed secondary features of the area of interest (cell nucleus) in an image. As a result, the feature-space of the two datasets represent two extremes - the digits data have a homogenous while diagnosis data has heterogenous feature-space. As such, we should observe interesting similarities and contrasts in the performance of different supervised learning algorithms on these datasets.

For analysis, the datasets are split into train, validation and test sets comprising 60%, 20%, and 20% of the full dataset respectively. The test set is extracted first, and then the algorithms learn on the train set and are cross-validated on the valid set several times before finally analyzing performance on the test set as a proxy for real-world, generalizable performance. Both datasets are available as native datasets within the Scikit Learn library.