Unified Algorithm and Datatype Taxonomy

Darkar Dengeno

March 21, 2016

1 Basic Datatype Spaces

A datapoint, d, is a pair of vectors $\{z, r\}$ such that $z \in \mathbb{Z}^\alpha, r \in \mathbb{R}^\beta$ For convenience, a datapoint may also include a map $l = \{z^* \in \mathbb{S}^{\alpha^*}, r^* \in \mathbb{S}^{\beta^*}\}$ where $\mathbb{S}$ is the set of strings and $\alpha^* \leq \alpha, \beta^* \leq \beta$ and surjection $F : d \to l$. A datatype is defined with α, β, l, and F.

2 Interpretation of Datatypes

Essentially, all categorical data is defined by the z vector and all continuous data is defined by the r vector. The map and surjection provide labels to elements of both vectors.

3 Taxonomy and Behavior of Machine Learning Algorithms

3.1 Classification Algorithms

A classification algorithm maps from $\{z, r\} \to \mathbb{Z}$

3.2 Clustering Algorithms

A clustering algorithm maps from $\{0, r\} \to \mathbb{Z}$. Notice that it is a subset of Classification.

3.3 Dimensionality Reduction

A dimensionality reduction algorithm maps from $\{z_0 \in \mathbb{Z}^{n_0}, r_0 \in \mathbb{R}^{m_0}\} \to \{z_1 \in \mathbb{Z}^{n_1}, r_1 \in \mathbb{R}^{m_1}\}$ such that $n_0 \gg n_1$ and $m_0 \gg m_1$.
Families: Input/Output, Classifier, Clustering, Extraction, Operation, Misc
Supervised learning Clustering Dimensionality reduction Structured prediction Anomaly detection Neural nets
Operation: 0 These are algorithms that are stateless - they cannot be trained or saved
Input/Output: 1
Classifier: 2
Clustering: 3
Extraction: 4
Structure: 5
Outlier: 6
NeuralNet: 7
Misc: 8