Hawaii Machine Learning

Review Session Meetup

Part I – Data Preprocessing

Data Preprocessing Steps

- Data Consolidation – collect/select/integrate
- Data Cleaning – imputation/outlier removal
- Data Transformation – scaling/normalization
- Data Reduction – minimize features/dimensions

Machine Learning Terminology

- Features: inputs, independent variables, column headings, e.g. age, salary
- Prediction: outputs, dependent variables, results
- Fitting: training, extracting rules from data
- Model: algorithm applied to data

What is Machine Learning?

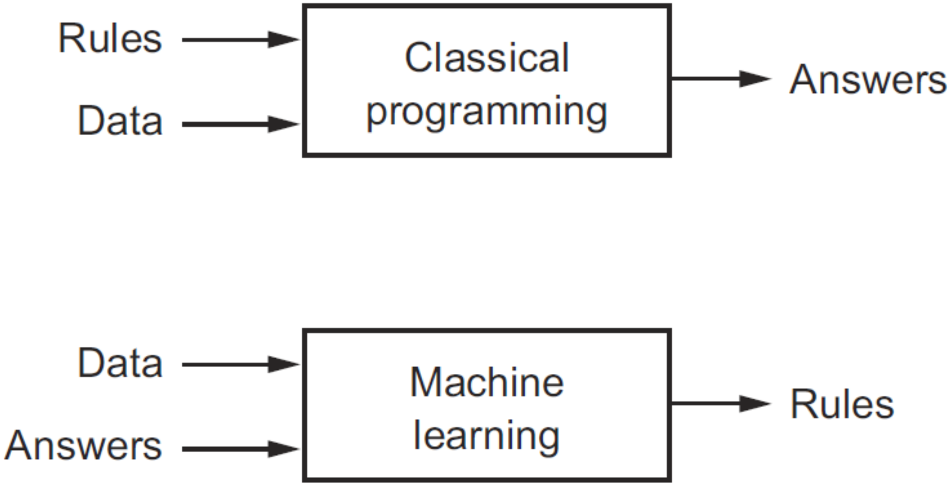


Figure 1.2 Machine learning: a new programming paradigm

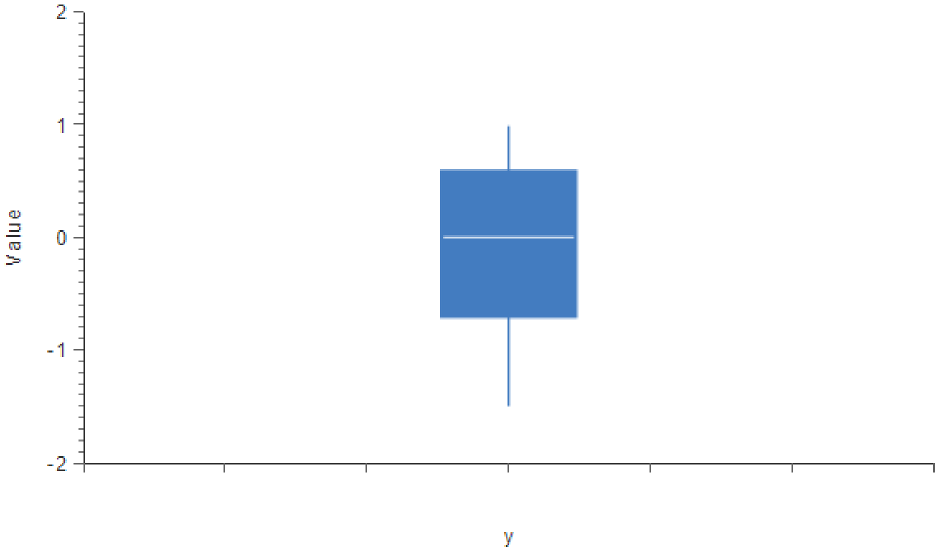
Excerpt from Deep Learning with Python, by Francois Chollet
©2018 Manning Publications Co.

Imputation = handling missing values

- Remove record (usually not desired)
- Replace with mean of other records
- Replace with median of other records
- Replace with mode of other records (for categorical data)
- Use regression to predict missing values

Handling Outliers

- Univariate method – box plot/median/quartile



- Multivariate method – based on multiple features
- Minkowski error – minimize loss

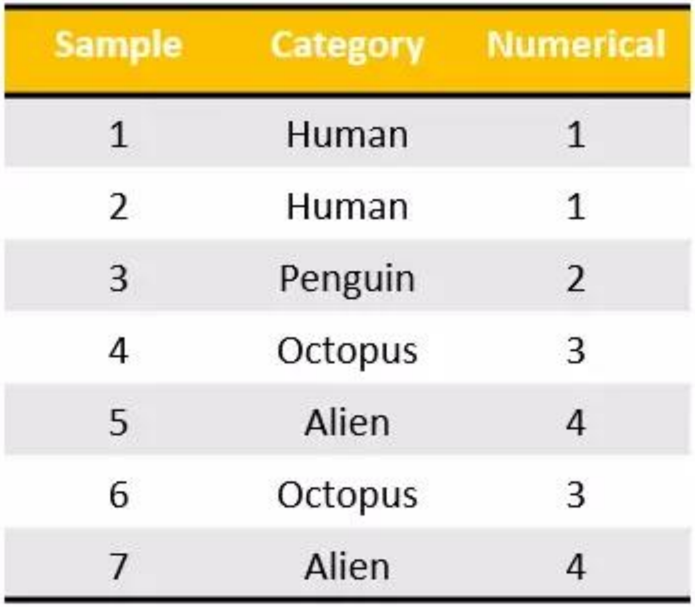
Data Types

- Numeric – float, int
- Categorical – string description of category without rank, e.g. France, Spain, Germany
- Ordinal – category with rank, e.g. good, better, best

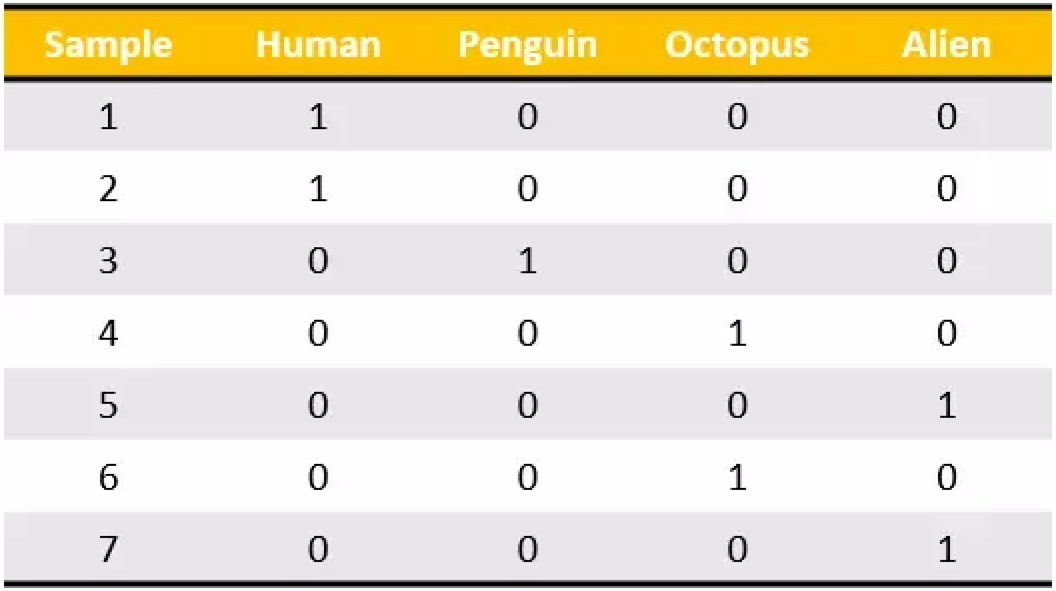
Everything must be converted to numeric

Categorical Encoding

- AKA One-Hot encoding, dummy encoding
- Converts feature set to vector of zeroes with a one indicating feature by position



Sample	Category	Numerical
1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4



Sample	Human	Penguin	Octopus	Alien
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	0	1	0
7	0	0	0	1

Data Transformation

- Scaling – applying scalar transformation
- Normalization – transform values to fit a numeric range, commonly 0.0 – 1.0
- Standardization – remove mean and scale to variance