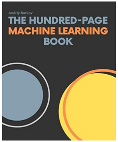
MSML610: Advanced Machine Learning

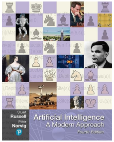
Lesson 02.3: ML Techniques - Input Processing

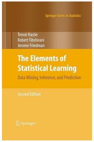
Instructor: Dr. GP Saggese - `gsaggese@umd.edu`

References:

- Burkov: *“The Hundred-Page Machine Learning Book”* (2019)
- Russell et al.: *“Artificial Intelligence: A Modern Approach”* (4th ed, 2020)
- Hastie et al.: *“The Elements of Statistical Learning”* (2nd ed, 2009)







- ***Input Processing***

Data Processing Transformations

- **Purpose of data processing**
  - Prepare raw data for effective machine learning
  - Improve model performance and generalization
- **Data cleanup**
  - Apply filters or smoothing to remove irrelevant variations
- **Handling missing data**
- **Types of transformations**
  - Normalization and standardization
  - Encoding categorical data
  - Feature construction
  - Dimensionality reduction
  - Discretization
- **Data augmentation**
  - Increase dataset size using transformations (common in vision)

Data Cleaning

- **Purpose of data cleaning**
  - Ensure data quality for accurate model training
  - Detect and correct errors or inconsistencies in the dataset
- **Typical steps in data cleaning**
  - *Remove duplicates*: Identical records are eliminated
  - *Correct data entry errors*: Fix misspellings or misformatted entries
  - *Standardize data*:
    - Convert dates formatted as both `MM/DD/YYYY` and `DD-MM-YYYY` into consistent format
    - String normalization (e.g., lowercase conversion)
    - Type conversion (e.g., strings to integers)
    - Dealing with unexpected characters or encodings
- **Relevance to ML models**
  - Poor data quality leads to biased or incorrect predictions
  - Clean data reduces variance and improves generalization

Handling Outliers and Missing Data

- **Outliers**
  - *Definition*: Data points significantly different from others
  - *Causes*: Measurement errors, variability in the data
  - *Detection*: Box plots, Z-scores, or interquartile method
  - *Treatment*: Removal, capping, or transformation (e.g., log scale)
- **Missing data**
  - *Detection*: Count of null values or incomplete entries
- **Remediation**
  - *Deletion*: Remove rows or columns with too many missing values
  - *Imputation*:
    - Mean/median/mode substitution
    - K-nearest neighbors (KNN)
    - Regression or model-based approaches
    - E.g., for a missing temperature reading, impute using the mean of the day's surrounding values

Normalization and Standardization

- **Goal**: Adjust feature scales for better convergence and learning
  - *Normalize*: Rescale to $[0, 1]$
  - *Standardize*: Zero mean, unit variance
- **Why it helps**
  - Equal feature contribution in distance-based models
  - Faster convergence in gradient-based algorithms
  - Enables regularization
  - Easier feature interpretation
- **Common methods**
  - Min-Max normalization: $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$
  - Z-score standardization: $x' = \frac{x - \mu}{\sigma}$

Encoding Categorical Data

- **Goal**: Convert non-numeric categories into numeric representations
- **Label encoding**
  - Assigns an integer to each category
    - E.g., `red`, `green`, `blue` $\rightarrow$ 1, 2, 3
  - Can mislead models if order is not meaningful
- **One-hot encoding**
  - Creates binary vector per category
    - E.g., `red`, `green`, `blue` $\rightarrow$ `[1,0,0]`, `[0,1,0]`, `[0,0,1]`
  - Avoids ordinal assumption
  - Increases dimensionality

Feature Construction

- **Goal**: Derive more informative features from raw inputs
- **Methods**
  - Combining variables (e.g., area $=$ height $\times$ width)
  - Extracting parts (e.g., year from a date)
  - Logical features
    - E.g., transform `2023-04-15` $\rightarrow$ (`Saturday`, `is_weekend = True`)
- **Why it helps**
  - Encodes domain knowledge
  - Improves model expressiveness and performance

Dimensionality Reduction

- **Goal**: Reduce number of features while preserving key information
- **Why it helps**
  - Reduce overfitting
  - Reduce data redundancy
  - Improve model speed
  - Allow visualization
- **Common techniques**
  - PCA: linear combinations that maximize variance
  - LDA: projects data to maximize class separability
- **Example**
  - Reduce 1024x640 image pixels to 10 principal components
  - Quantize color images into gray scale

Discretization

- **Goal**: Convert continuous values into categorical bins
- **Why it helps**
  - Simplifies models or enables categorical algorithms
  - Helps detect threshold effects in data
- **Techniques**
  - Equal-width binning
  - Quantile binning
- **Example**
  - Discretize age
    - `Child`: $[0, 13)$
    - `Teen`: $[13, 20)$
    - `Adult`: $[20, 65)$
    - `Senior`: $[65, \infty)$
  - Age 32 $\rightarrow$ `Adult`

Noise Removal

- **Goal**: Remove irrelevant or corrupt data variations
- **Why it helps**
  - Improves signal clarity and model robustness
    - E.g., clean noisy speech by removing high-frequency noise
- **Methods**
  - Smoothing (e.g., moving average)
  - Filtering (e.g., low-pass filter in audio)