KAZAKH-BRITISH TECHNICAL UNIVERSITY

Introduction to Machine Learning
Week 2

Olivier JAYLET

School of Information Technology and Engineering

Quantitative Vs. Qualitative data

- **Quantitative** data are numbers-based, countable, or measurable.
- **Qualitative** data is interpretation-based, descriptive, and relating to language.

Give three examples for both types.

Numerical data

Numerical data: Refers to data that is represented in numbers.

- All of the quantitative data are numerical.
- A numerical data can also be ~~quantitative~~ qualitative.

Numerical data can further be divided into **continuous** and **discrete**.

Continuous variable

Continuous variables can take any value within a given range, including fractions or decimals. They are typically **associated with measurements** that can be infinitely precise, depending on the level of measurement detail.
Examples:

- height of a person
- execution time of a program
- distance traveled
- speed of a vehicle

Discrete variable

Discrete variables can only take specific, distinct values, often whole numbers. These values are countable, and there are no intermediate values between them. Discrete variables are usually related to counting:

- number of employees in a company
- number of products sold
- goals scored in a soccer match

Categorical variable

Categorical data represent qualitative attributes and are often divided into distinct categories, **nominal** & **ordinal** :

- Categorical **nominal** variable consists of categories with no inherent order or ranking.
- Categorical **ordinal** variable includes categories with a meaningful order or ranking.

What is a dataset

A dataset is a structured collection of data organized and stored together for analysis or processing.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
59	23.0	4	97.0	54	2254	23.5	72	2	volkswagen type 3
203	29.5	4	97.0	71	1825	12.2	76	2	volkswagen rabbit
191	22.0	6	225.0	100	3233	15.4	76	1	plymouth valiant
220	33.5	4	85.0	70	1945	16.8	77	3	datsun f-10 hatchback
326	43.4	4	90.0	48	2335	23.7	80	2	vw dasher (diesel)
324	40.8	4	85.0	65	2110	19.2	80	3	datsun 210
369	34.0	4	112.0	88	2395	18.0	82	1	chevrolet cavalier 2-door

Data processing

Data processing is the collection and manipulation of digital data to produce meaningful information.

1. Data collection
2. Data cleaning
3. Data exploration and visualization
4. feature engineering

Feature engineering

Feature engineering preprocesses raw data into a machine-readable format. It optimizes ML model performance by transforming and selecting relevant features.

1. Feature creation
2. Feature transformation
3. Feature selection
4. Outliers policy

Feature matrix

A tabular numerical dataset can be represented as a feature matrix X of shape $n \times d$ where :

- $n =$ number of sample (rows)
- $d =$ number of features (columns)

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

How can we do math with categorical features ?

The case of boolean features[1]

A Boolean variable is a type of variable that can hold only two possible values: **true** or **false**
(can also be any other couple of words like yes/no, male/female, win/defeat et cetera.).

$$\mathbf{X} = \begin{bmatrix} \textbf{Age} & \textbf{Height} & \textbf{Is\_Student} \\ 25 & 170 & \textit{False} \\ 30 & 160 & \textit{False} \\ 18 & 180 & \textit{True} \\ 50 & 165 & \textit{False} \end{bmatrix}$$

[1] Binary encoding

The case of categorical nominal variables

$$\mathbf{X} = \begin{bmatrix} \textbf{Age} & \textbf{Height} & \textbf{Weight} & \textbf{Category} \\ 25 & 170 & 65 & \text{Adult} \\ 30 & 160 & 70 & \text{Adult} \\ 18 & 180 & 75 & \text{Youth} \\ 50 & 165 & 80 & \text{Senior} \end{bmatrix}$$

One-hot encoding in action

Using One-hot encoding, a Binary Column is created for each Unique Category in the variable.

$$\mathbf{X} = \begin{bmatrix} \textbf{Age} & \textbf{Height} & \textbf{Weight} & \textbf{Adult} & \textbf{Youth} & \textbf{Senior} \\ 25 & 170 & 65 & 1 & 0 & 0 \\ 30 & 160 & 70 & 1 & 0 & 0 \\ 18 & 180 & 75 & 0 & 1 & 0 \\ 50 & 165 & 80 & 0 & 0 & 1 \end{bmatrix}$$

(Don’t forget to drop ex category column)

One-hot encoding

If a categorical feature belongs to the finite set {1,..., K}, it is encoded by a binary vector :

$$(\delta_1, \ldots, \delta_K) \in \{0, 1\}^K, \quad \sum_{k=1}^{K} \delta_k = 1.$$

where $(\delta_1, \ldots, \delta_K)$ is a vector of K features.

One-hot encoding

If a categorical feature belongs to the finite set {1,..., K}, it is encoded by a binary vector :

$$(\delta_1, \ldots, \delta_K) \in \{0,1\}^K, \quad \sum_{k=1}^{K} \delta_k = 1.$$

where $(\delta_1, \ldots, \delta_K)$ is a vector of K features.

Properties :

1. Each δ_k is binary (0 or 1)
2. Only one δ_k can be 1 at a time.

One-hot encoding issue

Some models like **linear regression** rely on **matrix inversions** and assume that independant variables exhibit **full rank**.

It can also increase the dimensions of the matrix.

Dummy encoding

Similar to one-hot encoding, but seeks to avoid linear dependence.

While one hot encoding uses K binary variables for K categories in a variable. Dummy encoding uses K-1 features to represent K categories (labels).

Ordinal encoding

Each unique category is assigned a unique integer value.

$$X = \begin{bmatrix} \textbf{Name} & \textbf{Education Level} & \textbf{Education Level} \\ \text{Alice} & \text{High School} & 0 \\ \text{Bob} & \text{Bachelor's} & 1 \\ \text{Charlie} & \text{Master's} & 2 \\ \text{Diana} & \text{PhD} & 3 \\ \text{Eve} & \text{High School} & 0 \end{bmatrix}$$

Why Should We Scale Our Data?

- Some ML algorithms are sensitive to feature magnitude.
- Features with large ranges can dominate those with smaller ranges.
- Scaling helps to:
  - Improve model convergence speed (e.g., in gradient descent).
  - Enhance model performance for some algorithms.

Do All Models Need Scaled Data?

- Require scaling:
  - K-Nearest Neighbors (KNN)
  - Support Vector Machines (SVM)
  - Principal Component Analysis (PCA)
  - K-Means Clustering
  - Neural Networks
- Do not require scaling:
  - Decision Trees
  - Random Forests
  - Gradient Boosted Trees (e.g., XGBoost, LightGBM)
- It depends:
  - Linear Regression
  - Logistic Regression

Common Data Scaling Techniques

- Min-Max Scaling (Normalization):

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$$

- Standardization (Z-score Scaling):

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation.

- Robust Scaling (for data with outliers):

$$x_{\text{scaled}} = \frac{x - \text{median}}{\text{IQR}}$$

where IQR is the interquartile range.

Thank you for your attention !