An overview of supervised and unsupervised learning

Examining Data Types and Model Selection

Machine Learning

Types and Models

Types of Machine Learning

supervised learning

unsupervised learning

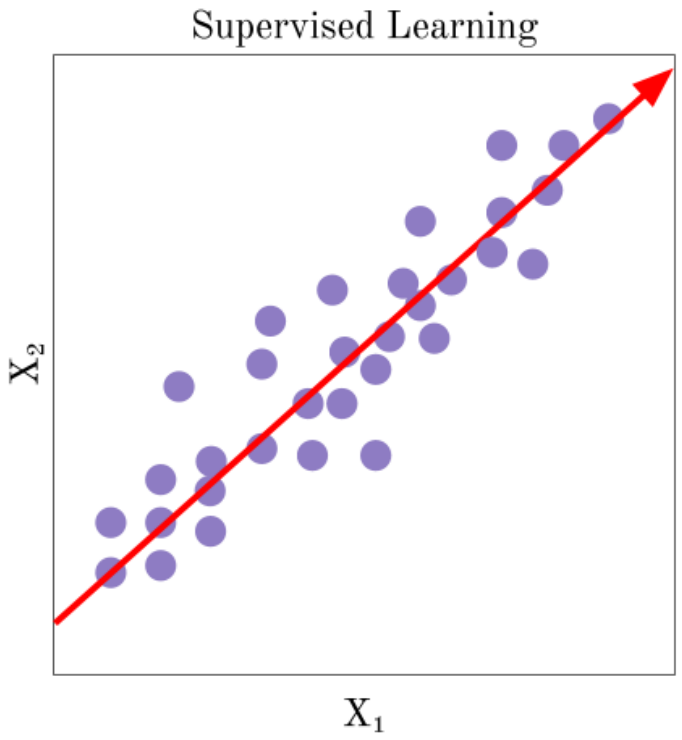
semi-supervised learning

reinforcement learning

Comparison of Supervised and Unsupervised Learning

Supervised learning	Unsupervised learning
Uses Known and Labeled Data as input	Uses Unknown Data as input
The number of Classes is known	The number of Classes is not known
In supervised learning training data is used to infer model	In unsupervised learning training data is not used.
We can test our model.	We can not test our model.
The desired output is given.	The desired output is not given.

Supervised Learning



- Involves training a model using labeled data (input-output pairs). The algorithm learns to map inputs to desired outputs.

- Supervised learning is widely used for tasks where predictions about future or unknown outcomes are necessary, such as spam detection or credit scoring.

Regression Models

- Linear Regression

```
class sklearn.linear_model.LinearRegression(*, fit_intercept=True, copy_X=True,
n_jobs=None, positive=False)
```

- Polynomial Regression

```
class sklearn.preprocessing.PolynomialFeatures(degree=2, *, interaction_only=False,
include_bias=True, order='C')
```

Degree: *int or tuple (min_degree, max_degree), default=2*

Evaluation Metrics
(Regression)

Mean Absolute Error (MAE)

Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)

R-squared (r2 score)

Classification Models

- Logistic Regression

```
class sklearn.linear_model.LogisticRegression(penalty='l2', *, dual=False, tol=0.0001,
C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None,
solver='lbfgs', max_iter=100, multi_class='deprecated', verbose=0, warm_start=False,
n_jobs=None, l1_ratio=None)                                                    [sour
```

Penalty: *{'l1', 'l2', 'elasticnet', None}, default='l2' (regularization)* / C: *float, default=1.0* (Inverse of regularization strength)

solver: *{'lbfgs', 'liblinear', 'newton-cg', 'newton-cholesky', 'sag', 'saga'}, default='lbfgs'*

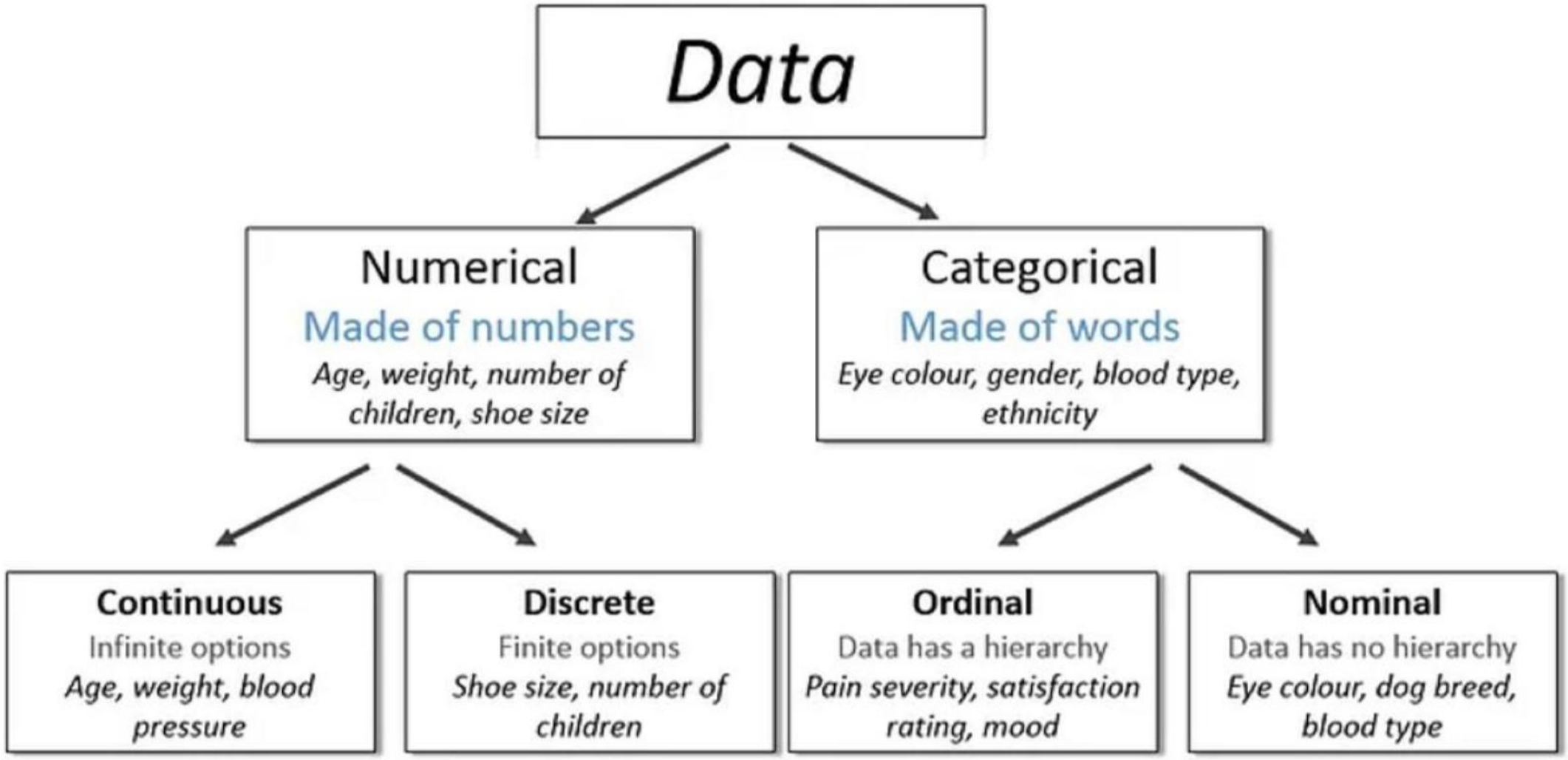
- Decision Trees, Random Forests, SVM, k-NN, etc.

Evaluation Metrics
(Classification)

- Accuracy
- Precision and Recall
- F1 Score
- ROC-AUC
- Confusion Matrix

Data types in machine learning

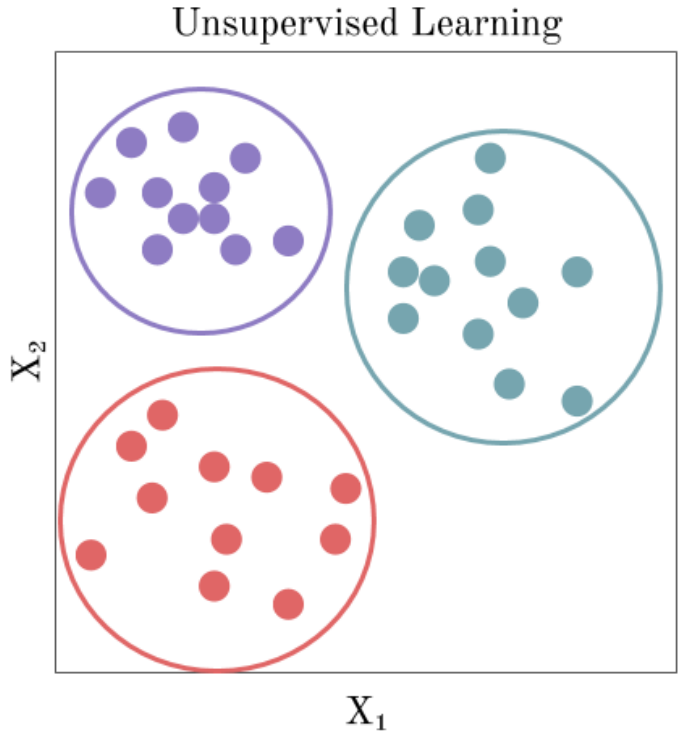
How to choose the right model?


Data
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type, ethnicity
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

Understand Your Data

Define the Problem

Unsupervised Learning



- Involves training a model on data without labeled responses. The goal is to find hidden patterns or intrinsic structures within the input data.

- Commonly used in customer segmentation, anomaly detection, and association mining.

Clustering Models

○ Hierarchical Clustering

- Builds a hierarchy of clusters either in a top-down (divisive) or bottom-up (agglomerative) manner.

```
class sklearn.cluster.AgglomerativeClustering(n_clusters=2, *, metric='euclidean',
memory=None, connectivity=None, compute_full_tree='auto', linkage='ward',
distance_threshold=None, compute_distances=False)
```

n_clusters: *int or None, default=2* (The number of clusters to find. It must be None if distance_threshold is not None.)

distance_threshold: *float, default=None* (The linkage distance threshold at or above which clusters will not be merged.)

K-Means Clustering

- Partitions data into K clusters by minimizing the variance within each cluster.

```
class sklearn.cluster.KMeans(n_clusters=8, *, init='k-means++', n_init='auto', max_iter=300,
tol=0.0001, verbose=0, random_state=None, copy_x=True, algorithm='lloyd')          [source]
```

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise.

```
class sklearn.cluster.DBSCAN(eps=0.5, *, min_samples=5, metric='euclidean',
metric_params=None, algorithm='auto', leaf_size=30, p=None, n_jobs=None)
```

Eps: *float, default=0.5 (The maximum distance between two samples)*

min_samples: *int, default=5 (*The number of samples (or total weight) in a neighborhood)

Association Rule Learning

- **Apriori Algorithm**

Is a foundational method in data mining used for discovering frequent itemsets and generating association rules.

- Eclat Algorithm

It is a more efficient and scalable version of the Apriori algorithm.(works in a vertical manner)

Dimensionality Reduction Techniques

- is the process of reducing the number of features (or dimensions) in a dataset while retaining as much information as possible.
  - **PCA, SNE, SVD, LDA, etc.**

Challenges

Overfitting and Underfitting

when a model learns the training data too well, capturing noise, whereas underfitting occurs when it fails to capture the underlying trend.

Data Quality and Preprocessing

The performance of machine learning models heavily relies on the quality of input data, necessitating data cleaning and preprocessing

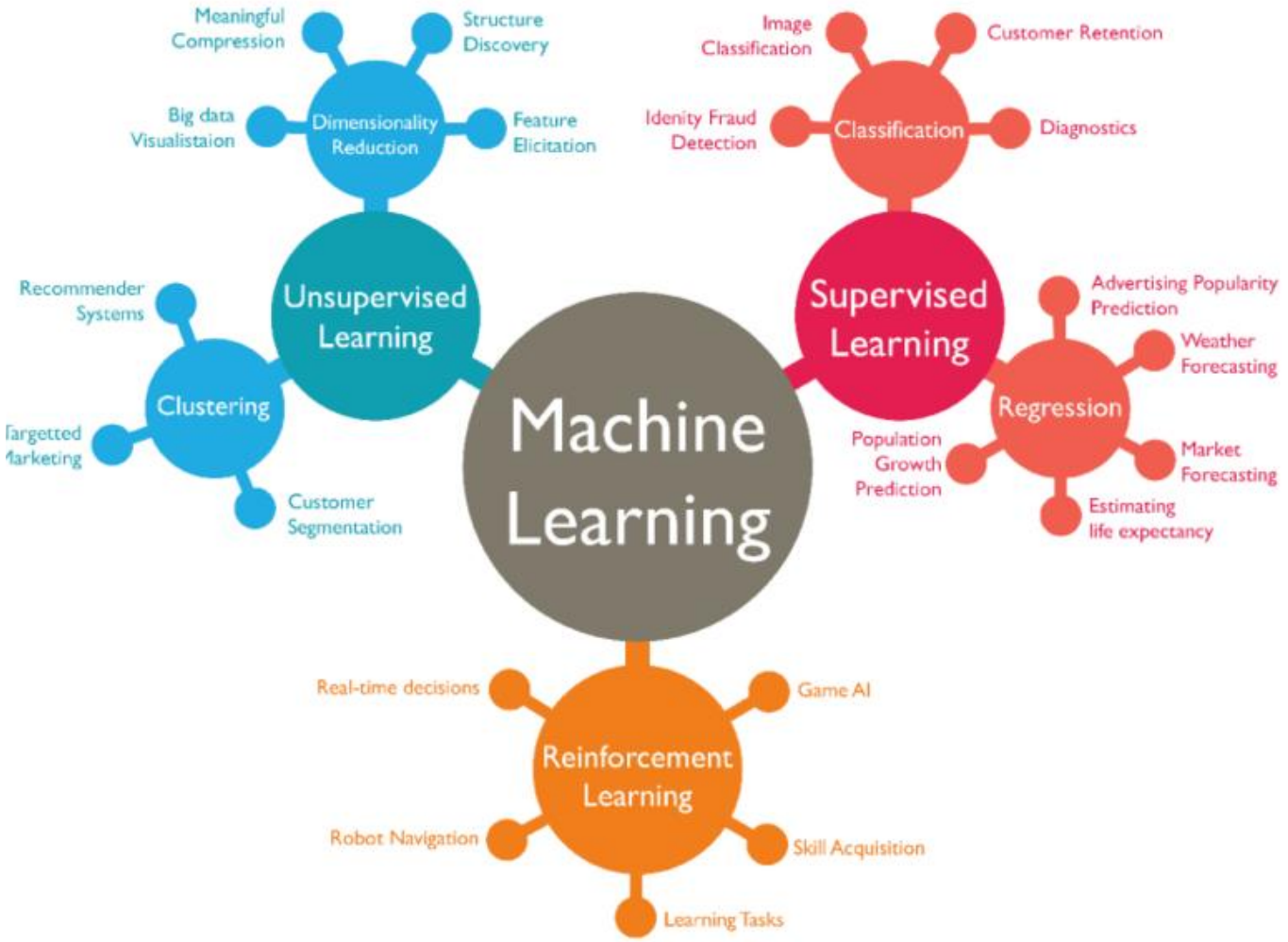
Scalability Issues

As data volume grows, some algorithms may struggle with performance and computational efficiency.

Codes

Colab Link

Statistic Project



Thank you

Nour Sajadi