DATA SCIENCE BRAIN
@datasciencebrain

MACHINE LEARNING

Model Selection Cheat Sheet

Save for later reference ·················›

TYPE OF PROBLEM

Regression:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Decision Trees
- Random Forest
- Gradient Boosting
-

Classification:

- Logistic Regression
- Decision Trees
- Random Forest
- Gradient Boosting
- Support Vector Machines
- k-Nearest Neighbors

Clustering:

- K-Means
- Hierarchical Clustering
- DBSCAN

Time Series Forecasting:

- ARIMA
- SARIMA
- Prophet (for seasonality)

SIZE & LINEARITY OF DATASET

Size of Dataset

- **Large Dataset:**
  - Gradient Boosting
  - Random Forest
  - Deep Learning (Neural Networks)
- **Small/Medium Dataset:**
  - Linear Regression
  - Support Vector Machines
  - k-Nearest Neighbors
  - Naive Bayes

Linearity of Data:

- **Linear Relationship:**
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
  - Support Vector Machines (linear kernel)
- **Non-linear Relationship:**
  - Decision Trees
  - Random Forest
  - Gradient Boosting
  - Support Vector Machines (non-linear kernel)
  - Neural Networks

INTERPRETABILITY & HANDLING

Interpretability:

- **High Interpretability:**
  - Linear Regression
  - Logistic Regression
  - Decision Trees
- **Medium Interpretability:**
  - Random Forest
  - Support Vector Machines
- **Low Interpretability:**
  - Neural Networks
  - Gradient Boosting

Handling High-Dimensional Data:

- **Feature Importance is Crucial:**
  - Random Forest
  - Gradient Boosting
  - Lasso Regression (for feature selection)
- **Many Features, Non-linearity:**
  - Support Vector Machines
  - Neural Networks

HANDLING

Handling Categorical Variables:

- **Categorical Features:**
  - Decision Trees
  - Random Forest
  - Gradient Boosting
  - CatBoost (handles categorical features well)

Handling Imbalanced Classes:

- **Imbalanced Classes:**
  - Random Forest
  - Gradient Boosting
  - Resampling Techniques (oversampling, undersampling)

EFFICIENCY

Computational Efficiency:

- **Fast Training/Prediction:**
  - Linear Regression
  - Naive Bayes
  - k-Nearest Neighbors
- **Slower Training, High Accuracy:**
  - Random Forest
  - Gradient Boosting
  - Neural Networks (with GPU)

Ensemble Methods:

- **High Accuracy, Robustness:**
  - Random Forest
  - Gradient Boosting
  - XGBoost, LightGBM

TIME COMPLEXITY

Fast Training/Prediction:

- Linear Regression
- Naive Bayes

Moderate Time Complexity:

- Decision Trees
- Random Forest
- Gradient Boosting

High Time Complexity:

- Support Vector Machines
- Neural Networks

ONLINE LEARNING

Continuous Learning:

- Stochastic Gradient Descent
- Online Random Forest (if available)

HANDLING NON-NUMERIC DATA

Non-Numeric Features:

- Decision Trees
- Random Forest
- Gradient Boosting
- Naive Bayes

MODEL DEPLOYMENT

Ease of Deployment:

- Simpler Models (Linear Regression, Decision Trees)
- Frameworks with Low Latency Requirements (XGBoost, LightGBM)

Challenging Deployment:

- Complex Models (Neural Networks, Gradient Boosting)