JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

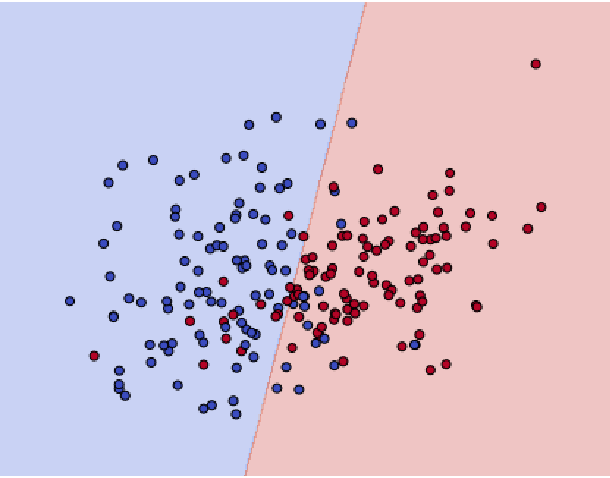
685.621 Algorithms for Data Science

Supervised Learning: Classification Algorithms

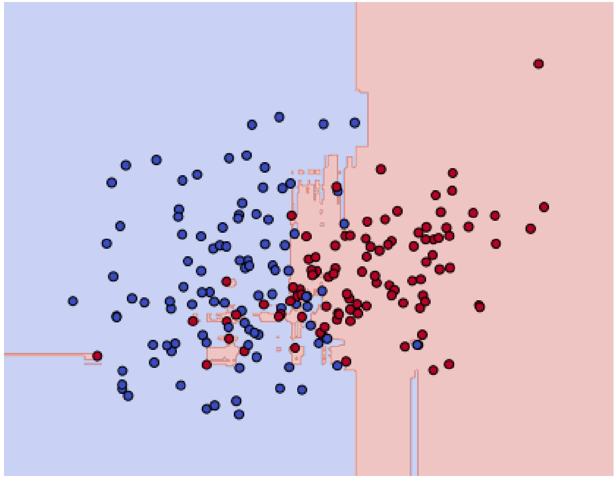
How Classifiers Separate Data

Linear Boundaries	Non-Linear Boundaries
• Logistic Regression • SVM	• KNN • Decision Trees, Random Forest

Logistic Regression Decision Boundary



Random Forest Decision Boundary



How Classification Models Learn

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

Log Loss

$$L = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij})$$

Softmax Cross-Entropy Loss

Handling More Than Two Classes

Problem: Many classification algorithms are naturally binary

Solution:

- **One-vs-Rest (OvR):**
  - **Train one classifier per class.**
  - **The class with the highest confidence score wins**
- **One-vs-One (OvO):**
  - **Train classifiers for every pair of classes.**
  - **Use majority voting**

Choosing the Right Model

Algorithm	Type	Key Characteristics
Logistic Regression	Linear Model	Simple, interpretable, probabilistic
K-Nearest Neighbors (KNN)	Instance-based	No training phase, works well for small datasets
Decision Trees	Rule-based	Intuitive, interpretable, prone to overfitting
Random Forest	Ensemble	Reduces overfitting, handles large feature sets
Support Vector Machines	Hyperplane-based	Effective in high-dimensional spaces, kernel trick
Neural Networks	Deep Learning	Complex, data-hungry, highly accurate

The Simplest Classifier

$$\hat{y} = \frac{1}{1 + e^{-(\omega^T x + b)}}$$

Advantages

- **Interpretable**
- **Probabilistic** output
- **Efficient** & scalable

Limitations

- Only finds **linear** decision boundaries
- **Assumes** feature independence
- Struggles with **class imbalance**

Instance-Based Learning

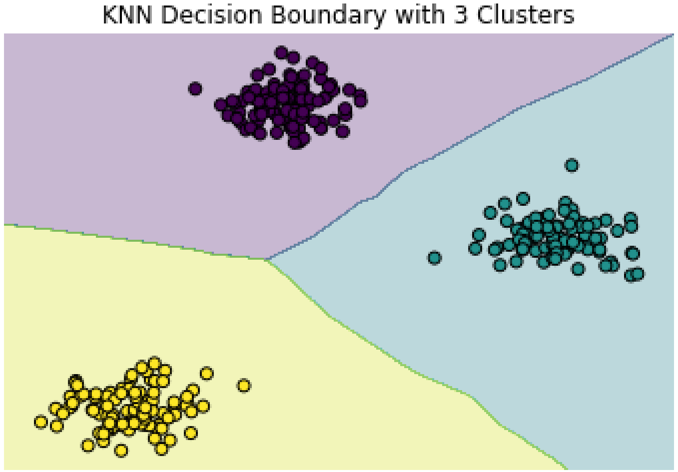
$$d\left(x, x'\right) = \sqrt{\sum_{i=1}^{n} \left(x_i - x'_i\right)^2}$$

Advantages

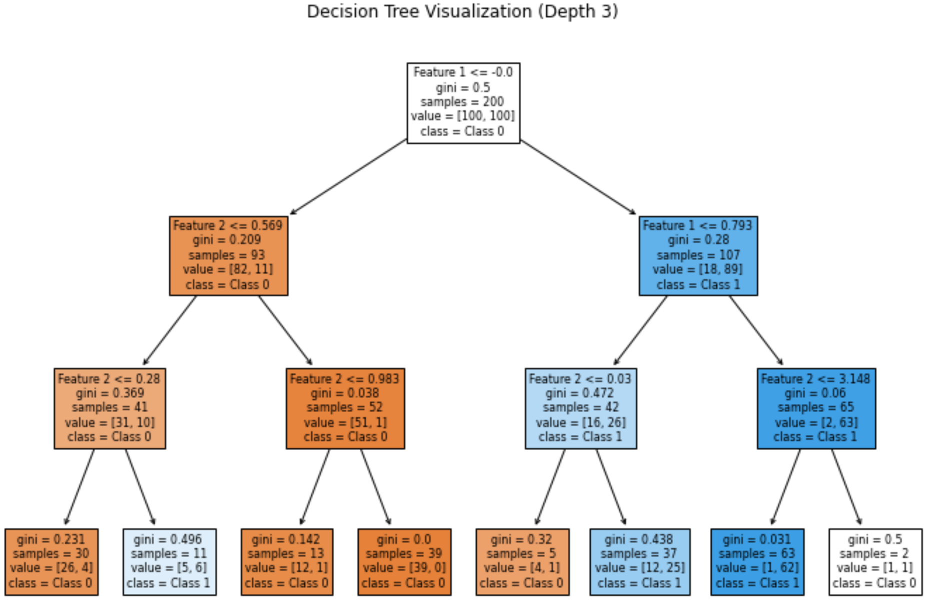
- **Simple & Intuitive**
- Works well with **non-linear** relationships
- **Adapts** to new data quickly

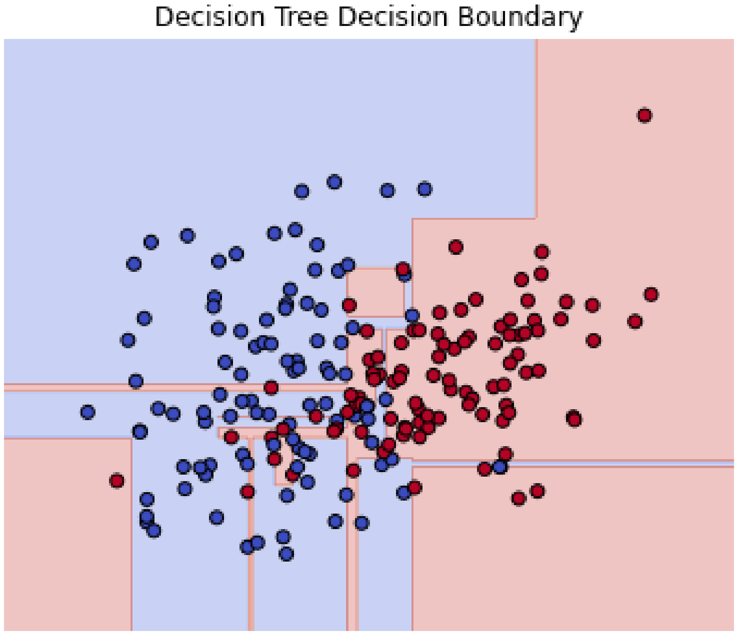
Limitations

- **Slow** on large datasets
- **Sensitive** to irrelevant features
- Choice of **k matters**



Rule-Based Classification





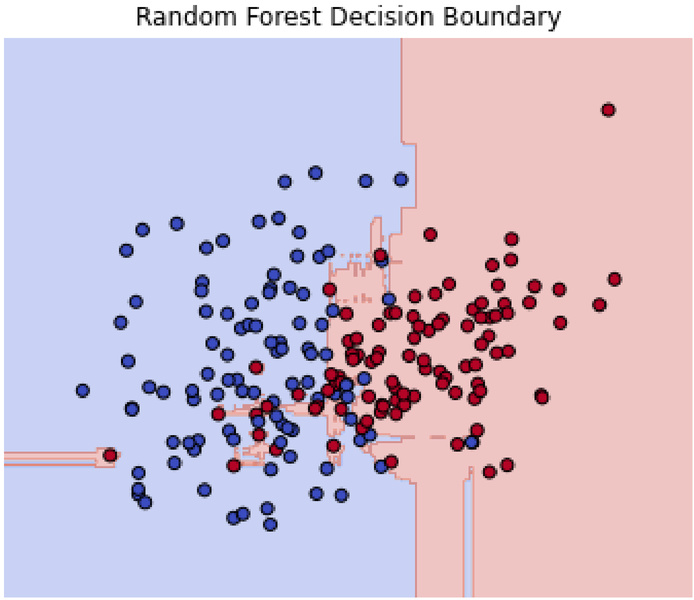
The Power of Ensembles

Advantages

- **Reduces overfitting – ** More stable than a single Decision Tree.
- **Handles high-dimensional data well – ** Works even when many features exist.
- **Works with missing data – ** Can still make predictions even if some values are missing.

Limitations

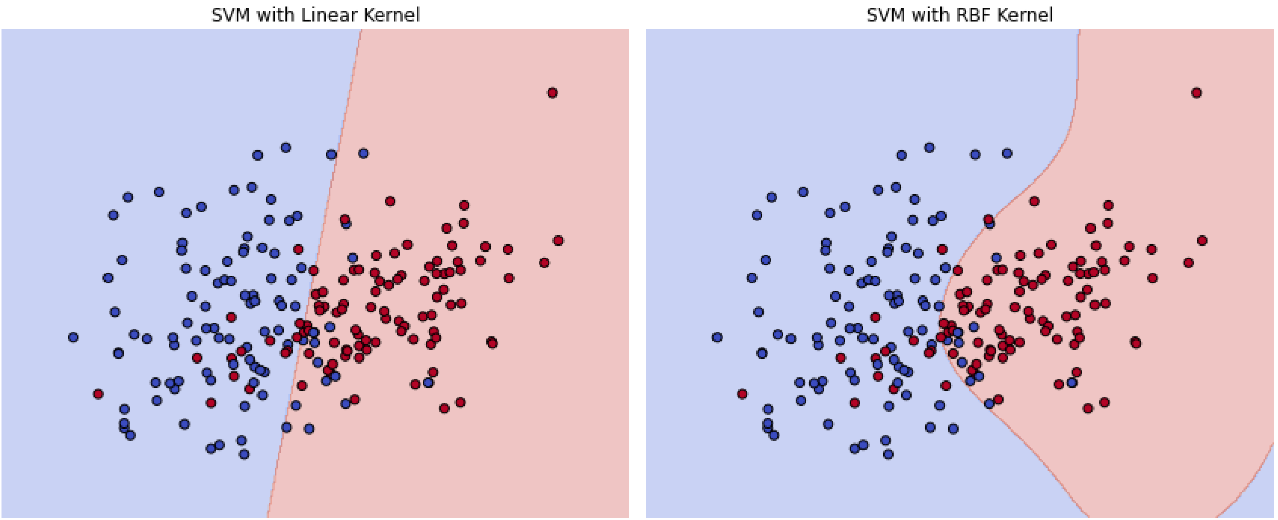
- **Less interpretable – ** Unlike a single Decision Tree, it's hard to visualize.
- **Computationally expensive – ** Training multiple trees takes more time than a single model.
- **May not work well for small datasets – ** Too many trees can lead to unnecessary complexity.



Finding the Optimal Decision Boundary

Hard Margin $\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$

Soft Margin $\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i,$



Support Vector Machine Optimization

Primal Form

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i,$$

Dual Form

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j),$$

Subject to: $0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0,$

JOHNS HOPKINS

WHITING SCHOOL *of* ENGINEERING