Supervised Learning Algorithms

Mr. Mehul A. Jadhav
Department of Computer Studies (MCA)
Vivekanand College, Kolhapur

12 July 2025

Agenda

Classification Tasks

Definition: Predicting discrete labels (categories).

Examples:

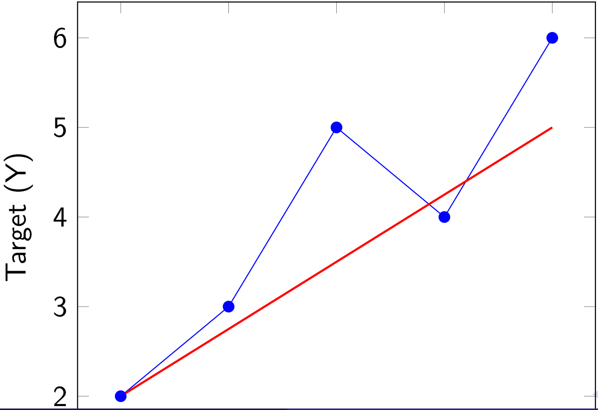
- Spam vs. non-spam email detection
- Image classification (e.g., cat vs. dog)

Regression Tasks

Definition: Predicting continuous values.

Examples:

- House price prediction
- Temperature forecasting



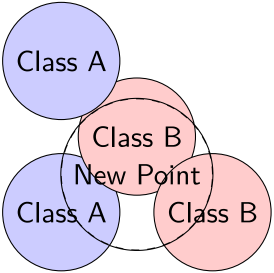
k-Nearest Neighbours (k-NN)

Concept: Classifies data points based on the majority class of their k-nearest neighbors using distance metrics (e.g., Euclidean distance).

Applications:

- Handwritten digit recognition
- Recommendation systems

k-NN Working Mechanism



$k = 3$

The new point is classified as Class A (2 blue, 1 red within radius).

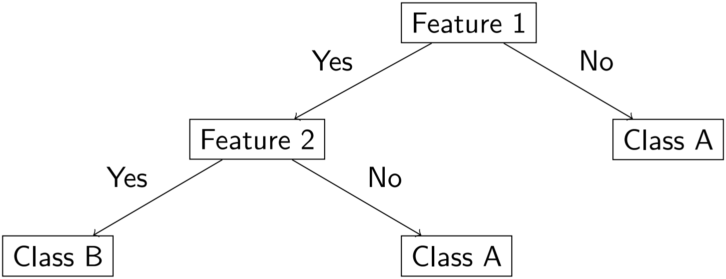
Introduction to Decision Trees

Concept: A tree-like model where decisions are made based on feature conditions.

Applications:

- Credit risk assessment
- Medical diagnosis

Decision Tree Structure



ID3 Algorithm

Concept: Uses entropy and information gain to select the best feature for splitting.

Entropy: $H(S) = -\sum p_i \log_2(p_i)$

Information Gain: $IG(A) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$

CART Algorithm

Concept: Uses Gini index for splitting in classification tasks.

Gini Index: $Gini = 1 - \sum(p_i)^2$

Lower Gini index indicates better split.

Random Forest and Ensemble Learning

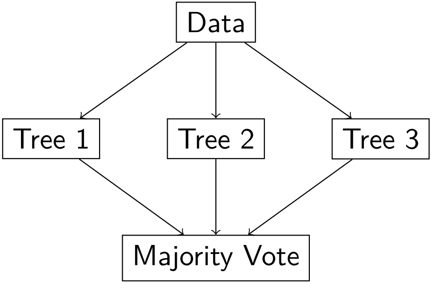
Concept: Combines multiple decision trees to improve accuracy and reduce overfitting.

Bagging: Bootstrap aggregating creates diverse subsets of data for each tree.

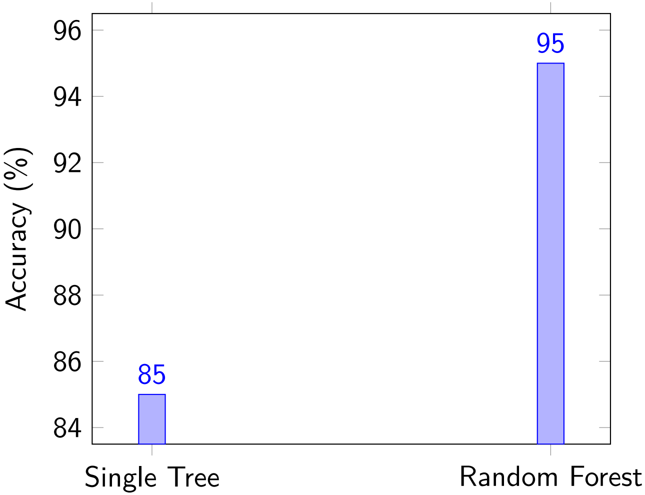
Applications:

- Fraud detection
- Stock market prediction

Random Forest Structure



Random Forest Performance



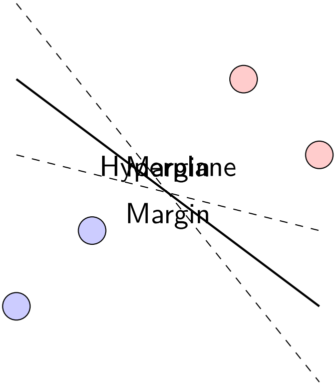
Introduction to SVM

Concept: Finds the optimal hyperplane that maximizes the margin between classes.

Applications:

- Text classification
- Face recognition

SVM Hyperplane and Margin



Model Persistence

Saving Models: Store trained models for future use.

Loading Models: Reuse pre-trained models for predictions.

Example (Python):

- Save: `joblib.dump(model, 'model.pkl')`
- Load: `model = joblib.load('model.pkl')`

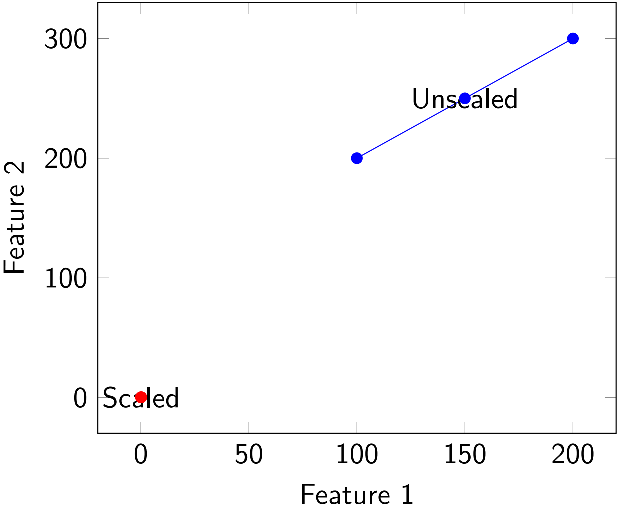
Importance of Data Scaling

Concept: Ensures features contribute equally to the model.

Techniques:

- **Min-Max Scaling:** Scales features to a range (e.g., [0,1]).
- **Standardization:** Centers data with mean 0 and variance 1.

Effect of Scaling



k-NN Implementation

Implementation (Python):

- Use `sklearn.neighbors.KNeighborsClassifier`
- Key parameters: `n_neighbors`, `metric`

Use Cases: Image recognition, recommender systems.

Decision Tree Implementation

ID3: Entropy-based, suitable for categorical data.

CART: Uses Gini index, supports regression and classification.

Implementation (Python): `sklearn.tree.DecisionTreeClassifier`

Random Forest Implementation

Building: Combine multiple trees using bagging.

Tuning Parameters:

- `n_estimators`: Number of trees
- `max_depth`: Maximum tree depth

Implementation (Python):

`sklearn.ensemble.RandomForestClassifier`

Practical Applications:

- Text classification (e.g., sentiment analysis)
- Bioinformatics

Implementation (Python): `sklearn.svm.SVC`

Conclusion

Supervised learning algorithms like k-NN, Decision Trees, Random Forests, and SVMs are powerful tools for classification and regression tasks.

Proper data scaling and model persistence enhance their practical utility.

References

- Scikit-learn Documentation: Supervised Learning
- GeeksforGeeks: Machine Learning Algorithms
- Towards Data Science: Understanding SVM