Week 00b Basic Handout: Supervised Learning - Predicting from Data

Machine Learning for Smarter Innovation

1 Week 00b Basic Handout: Supervised Learning - Predicting from Data

1.1 Overview

Learn prediction algorithms without advanced math. Focus on intuition and practical use.

1.2 Key Concepts

1.2.1 What is Supervised Learning?

Learning to predict outputs from inputs when you have labeled examples.

Examples: - Predict house prices from features (regression) - Classify emails as spam/not spam (classification) - Diagnose diseases from symptoms (classification)

1.2.2 When to Use

- Have input-output pairs
- Want to predict new cases
- Pattern is consistent

1.3 Algorithms at a Glance

1.3.1 1. Linear Regression

Use: Predict continuous values (price, temperature) **Pro**: Simple, interpretable, fast **Con**: Only captures linear relationships

1.3.2 2. Logistic Regression

Use: Binary classification (yes/no decisions) **Pro**: Probability outputs, interpretable **Con**: Linear decision boundary

1.3.3 3. Decision Trees

Use: Both regression and classification **Pro**: Human-readable, handles non-linear **Con**: Overfits easily

1.3.4 4. Random Forest

Use: Most tasks, especially tabular data **Pro**: Robust, handles overfitting, accurate **Con**: Slower, less interpretable

1.3.5 5. Gradient Boosting (XGBoost)

Use: Kaggle competitions, production systems **Pro**: State-of-art accuracy **Con**: Requires tuning, slow training

1.4 Decision Guide

Start with: Random Forest (most forgiving) **Need speed**: Logistic/Linear Regression **Need interpretability**: Decision Tree **Need max accuracy**: XGBoost **Non-linear + small data**: SVM with kernel

1.5 Common Pitfalls

- Using accuracy on imbalanced data
- Ignoring feature scaling
- No train/test split
- Overfitting to training set

1.6 Next Steps

- Week 00c: Unsupervised Learning
- Try: Kaggle Titanic competition