Supervised Learning - Classification 1

Course:

INFO-6145 Data Science and Machine Learning

Developed by:
Mohammad Noorchenarboo

October 1, 2024

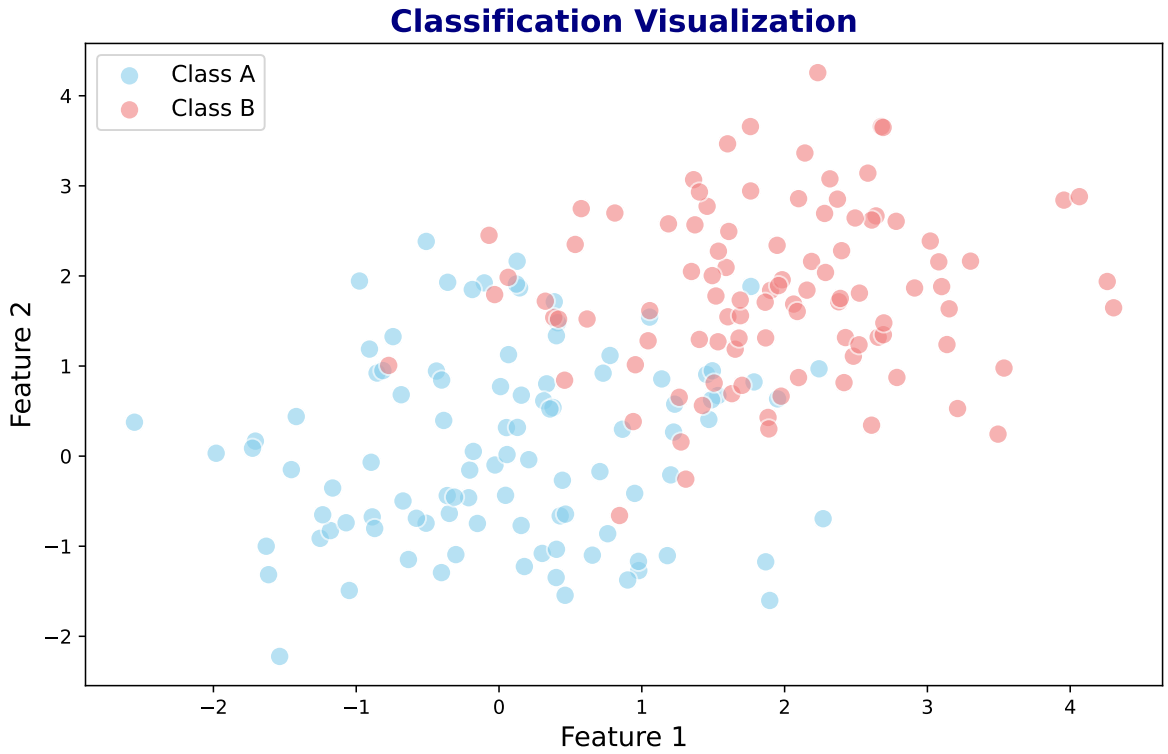
Contents

Current Section

1. Classification Basics
   - Introduction to Classification
   - The Problem of Classification
   - Examples of Classification
   - Logistic Regression
   - Scores and Thresholds
   - Confusion Matrix
   - Binary Performance Metrics

Classification Basics

Classification is about determining the category an object belongs to, and it is a supervised learning method. Unlike regression, which predicts continuous values, classification assigns discrete labels.



The Problem of Classification

The formal problem statement of classification involves mapping objects to classes based on a hypothesis that connects features to their respective classes.

Goal

The goal is to approximate the *target concept*, which defines the true class for each object based on the feature set.

Example

Suppose we have two features: X_1 (age) and X_2 (income). We classify whether a person buys a product or not (y).

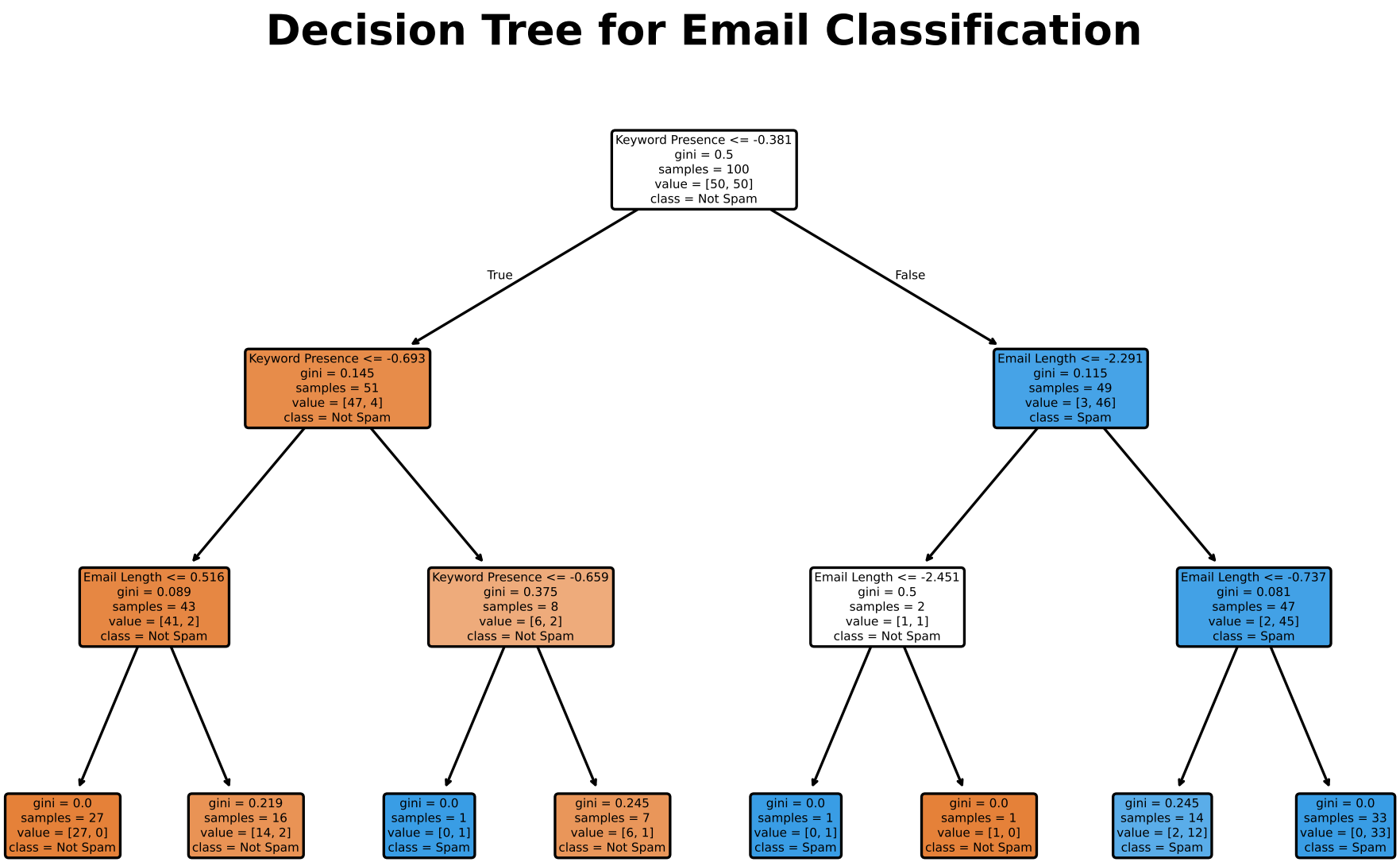
Examples of Classification

Hypotheses are formed based on visible features. For example:

Example

In email classification, an email can be classified as spam or not spam based on features such as the presence of certain keywords.

Examples of Classification



Logistic Regression

Logistic Regression is a commonly used classification algorithm that models the probability of a class label as a logistic function of input features.

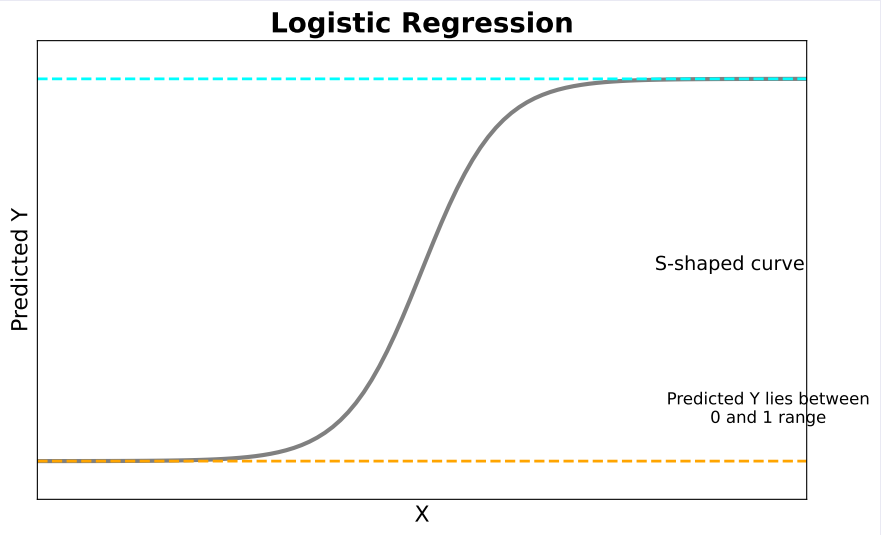
Logistic Regression

Key Idea

Logistic regression estimates the probability that a given input belongs to a certain class using a sigmoid function:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}}$$

where $X_1, \ldots, X_p$ are features, and $\beta_0, \beta_1, \ldots, \beta_p$ are model parameters.



Scores and Thresholds

Classification can also be done by generating probability scores and using different thresholds to assign class labels.

Thresholds

Altering thresholds can affect the classification result:

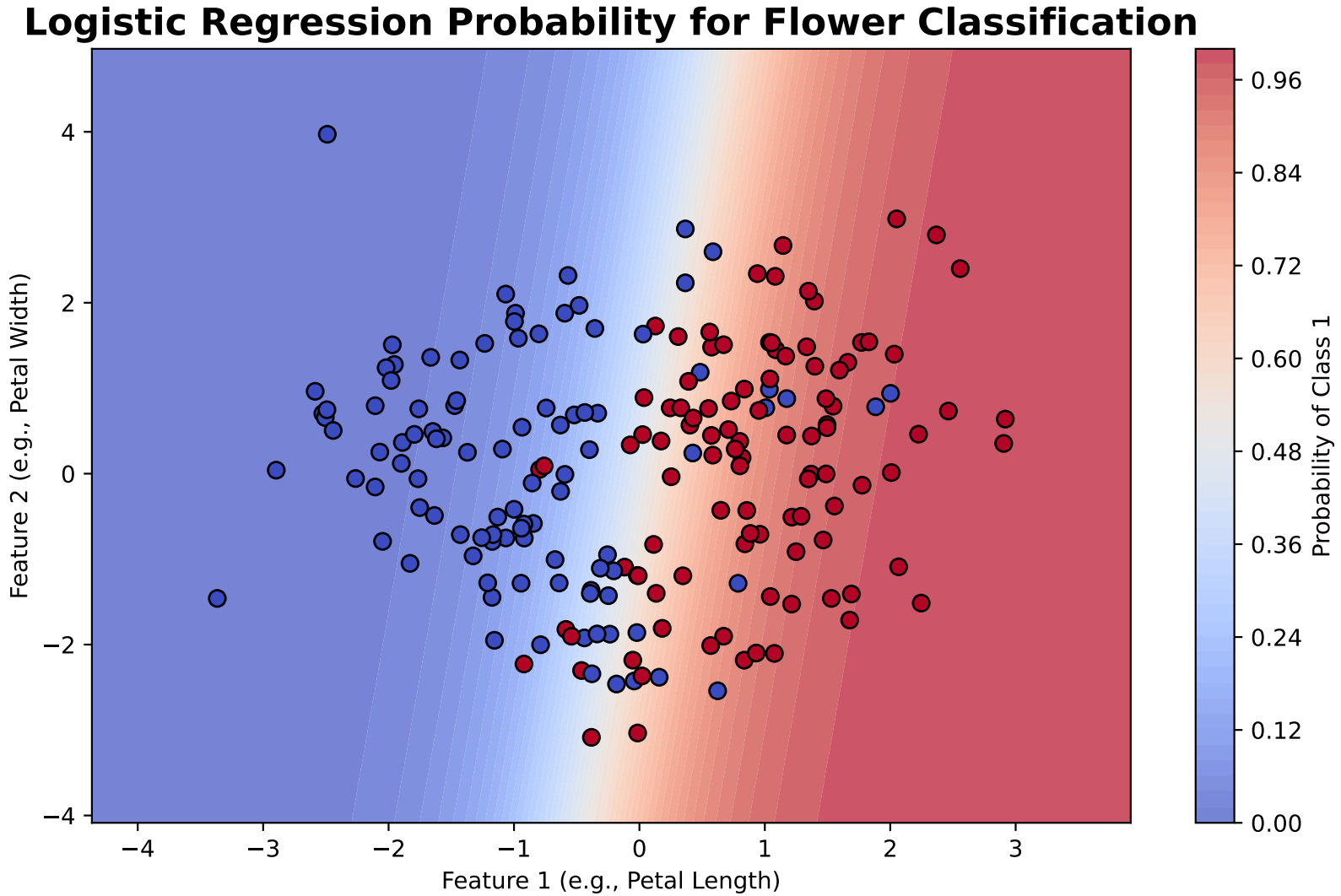
- A lower threshold increases sensitivity but may decrease specificity.
- A higher threshold decreases sensitivity but increases specificity.

Numerical Example

For the logistic regression output $P(y = 1|X) = 0.7$:

- Threshold $t = 0.5$: Class label is 1.
- Threshold $t = 0.8$: Class label is 0.

Scores and Thresholds



Confusion Matrix

A confusion matrix is used to describe the performance of a classification model, especially in binary classification.

Matrix Components

- **True Positive (TP):** Correctly predicted positive cases.
- **False Positive (FP):** Incorrectly predicted positive cases.
- **True Negative (TN):** Correctly predicted negative cases.
- **False Negative (FN):** Incorrectly predicted negative cases.

Confusion Matrix

Numerical Example

For a spam classifier:

	Predicted Positive	Predicted Negative
Actual Positive	$TP = 80$	$FN = 20$
Actual Negative	$FP = 10$	$TN = 90$

Binary Performance Metrics

Several performance metrics can be derived from the confusion matrix:

- **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{80 + 90}{200} = 85\%$$

- **True Positive Rate (Recall):**

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{80}{80 + 20} = 80\%$$

- **Specificity:**

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{90}{90 + 10} = 90\%$$

Binary Performance Metrics

- **False Positive Rate:**

$$\text{False Positive Rate} = \frac{FP}{FP + TN} = \frac{10}{90 + 10} = 10\%$$

- **Precision** Precision evaluates the correctness of positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{80}{80 + 10} = 88.9\%$$