UCI Heart Disease Dataset

16 columns, 920 rows

Columns

cp → chest pain type

trestbps → Resting blood pressure

chol → serum cholestrol

fbs → fasting blood sugar

restecg → Resting ecg resullts

thalch → Max. heart rate achieved

exang → exercise induced angina

num → disease severity

thal → Thalassemia status (Hereditary disease (haemoglobin)

Ca → No. of major vessels coloured by fluorosuppy

oldpeak — ST depression induced by exercise relative to rest

Slope — Trend of ST segment during peak exercise

Linear Regression

Model Eqn :- $\boxed{\hat{y} = b_0 + b_1 x}$ $\Rightarrow \hat{y} \in (-\infty, \infty)$

$\hat{y}$ = model prediction

b_0 = intercept

b_1 = regression coefficient / slope

For multiple features :-

$$\boxed{\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n}$$

Cost Function (MSE)

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right)^2$$

cost Function

$m \rightarrow$ total training samples

Gradient Descent Update Rule

$$\theta_j := \theta_j - \alpha \cdot \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) x_j^{(i)}$$

m = dataset size

α = learning rate in gradient descent

j^{th} feature of i^{th} example

Evaluation Metrics

$$MAE = \frac{1}{m} \sum |\hat{y} - y|$$

$$MSE = \frac{1}{m} \Sigma (\hat{y} - y)^2$$

$$RMSE = \sqrt{MSE}$$

$$R^2 = 1 - \frac{\Sigma (y - \hat{y})^2}{\Sigma (y - \bar{y})^2}$$

Logistic Regression

Sigmoid Function

$$\sigma(z) = \frac{1}{1 + e^{-z}} \Rightarrow \hat{y} \in [0, 1]$$

$\Downarrow$

Output

Logistic Hypothesis

$$\hat{y} = \sigma(b_0 + b_1 x_1 + \dots + b_n x_n)$$

Log Loss (Binary Cross Entropy)

$$J(\theta) = -\frac{1}{m} \sum \left[y \ln(\hat{y}) + (1-y) \ln(1-\hat{y}) \right]$$

Gradient Descent Update

$$\theta_j := \theta_j - \alpha \cdot \frac{1}{m} \sum (\hat{y} - y) x_j$$

Odds & Log - Odds

$$\text{Odds} = \frac{p}{1-p}$$

p = predicted probability

$$\log(\text{Odds}) = \ln\left(\frac{p}{1-p}\right)$$

Decision Boundary

$\hat{y} \geq 0.5 \Rightarrow 1$, $\hat{y} < 0.5 \Rightarrow 0$

Classification Metrics

$$\text{Precision} = \frac{TP}{TP + FP}$$

TP = True Positive

FP = False Positive

$$\text{Recall} = \frac{TP}{TP + FN}$$

TN = True Negative

FN = False Negative

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Random Forest

Gini Impurity

$$\text{Gini} = 1 - \sum p_i^2$$

p_i = probability of class i at a node

Entropy

$$\text{Entropy} = -\sum p_i \log_2 (p_i)$$

Information Gain

$$IG = \text{Entropy}(\text{parent}) - \sum \frac{n_j}{n} \text{Entropy}(\text{child}_j)$$

Number of Features Per Split

Classification : $\sqrt{P}$

Regression : $\frac{P}{3}$

P = total features

Bias-Variance Decomposition

Total Error = Bias^2 + Variance + Irreducible Error

Bagging (Bootstrap Aggregation)

Classification = Mode of all data

Regression = Mean of all data