Documentazione degli Algoritmi di Machine Learning Implementati

Davide [Cognome]

November 8, 2024

Contents

Introduzione

Questa documentazione descrive l'implementazione di una libreria di base per algoritmi di machine learning, creata per dimostrare conoscenze in algoritmi supervisionati e non supervisionati. Vengono inclusi esempi pratici per ogni algoritmo e dettagli sulla loro implementazione.

Chapter 1

Regressione Lineare

1.1 Descrizione

La regressione lineare è un modello statistico che cerca di predire il valore di una variabile dipendente y in base a una o più variabili indipendenti X.

1.2 Funzione di Costo

La funzione di costo $J(\theta)$ misura la differenza tra le predizioni del modello e i valori reali. Nella regressione lineare, la funzione di costo utilizzata è il *Mean Squared Error* (MSE):

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2$$

dove $h(x) = \theta_0 + \theta_1 x$.

1.3 Algoritmo di Gradient Descent

Per minimizzare la funzione di costo, si utilizza il metodo di *Gradient Descent*:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

dove α è il tasso di apprendimento.

Chapter 2

K-Nearest Neighbors (KNN)

2.1 Descrizione

K-Nearest Neighbors (KNN) è un algoritmo supervisionato che classifica un dato punto in base alle classi dei suoi k vicini più prossimi.

2.2 Calcolo della Distanza

Una comune misura di distanza è la distanza euclidea:

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

Chapter 3

Albero di Decisione

3.1 Descrizione

Gli alberi di decisione sono modelli supervisionati che dividono i dati in base ad attributi che massimizzano l'informazione (es., usando l'entropia o l'indice di Gini).

3.2 Calcolo dell'Entropia

L'entropia misura il grado di disordine di un set:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

Chapter 4

Regressione Logistica

4.1 Descrizione

La regressione logistica è un algoritmo supervisionato usato per la classificazione, in cui si predice la probabilità che un esempio appartenga a una certa classe.

4.2 Funzione Sigmoid

La funzione di attivazione sigmoid è usata per ottenere un valore di probabilità:

$$h(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Chapter 5

Support Vector Machine (SVM)

5.1 Descrizione

La Support Vector Machine (SVM) cerca di trovare un iperpiano ottimale che separi le classi, massimizzando il margine tra i dati.

Chapter 6

K-means Clustering

6.1 Descrizione

K-means è un algoritmo non supervisionato che raggruppa i dati in k cluster, minimizzando la distanza dei punti dai centroidi.

6.2 Aggiornamento dei Centroidi

Ad ogni iterazione, i centroidi vengono aggiornati calcolando la media dei punti in ciascun cluster.

Chapter 7

PCA (Principal Component Analysis)

7.1 Descrizione

L'Analisi delle Componenti Principali (PCA) è un algoritmo non supervisionato per la riduzione della dimensionalità, che cerca di massimizzare la varianza proiettando i dati lungo gli assi principali.

7.2 Calcolo della Matrice di Covarianza

Per trovare i componenti principali, si calcola la matrice di covarianza:

$$\text{Cov}(X) = \frac{1}{m} \sum_{i=1}^{m} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T$$