Università degli Studi di Urbino Carlo Bo

Implementazione di Regressione Lineare e K-Nearest Neighbors in Haskell e Prolog

Corso: Programmazione Logica e Funzionale

29 luglio 2023

Docente
Prof. Marco Bernardo

Corsista
Andrea De Lorenzis, 308024

Indice

1 Specifica del problema

Linear regression e K-Nearest neighbors (KNN) sono due algoritmi che trovano applicazione in diversi campi, tra cui statistica, analisi di dati e intelligenza artificiale. La prima consente di prevedere il valore di una variabile sconosciuta mediante un modello basato su un'equazione lineare. La seconda è una tecnica utilizzata per la classificazione di oggetti basandosi sulle caratteristiche degli oggetti vicini a quello considerato. L'obiettivo di questo progetto è quello di implementare le due tecniche nei linguaggi di programmazione Haskell e Prolog.

2 Analisi del problema

2.1 Dati di ingresso del problema

I dati di ingresso del problema sono inseriti dall'utente e consistono in un insieme di punti bidimensionali per allenare e valutare i modelli. L'utente inizialmente deve fornire un valore numerico per scegliere l'operazione da svolgere (0 - Linear regression, 1 - KNN, 2 - Terminare il programma). Una volta scelta una delle prime due operazioni, all'utente è richiesto di inserire un insieme di punti bidimensionali per allenare il modello. Successivamente, potrà inserire un ulteriore punto da valutare sul modello allenato, ottenendo poi il risultato. In particolare, i dati di ingresso per la linear regression sono:

- Un insieme di punti 2D separati da spazio, ognuno dei quali è costituito da una coordinata x e una coordinata y

$$x_1 \ y_1 \ x_2 \ y_2 \ \dots$$

- Un punto 2D da valutare sul modello lineare allenato.

I dati di ingresso per il KNN sono:

- Un insieme di punti 2D etichettati e separati da andata a capo, ognuno dei quali è costituito da una coordinata x, una coordinata y e una classe o etichetta numerica che rappresenta la categoria di appartenenza del punto

$$x_1 \ y_1 \ \text{label1}$$
$$x_2 \ y_2 \ \text{label2}$$
$$\dots$$

- Il valore k di vicini da valutare per il punto.
- Il punto 2D di test, senza etichetta, da classificare tramite KNN.

2.2 Dati di uscita del problema

I dati di uscita per la linear regression sono:

- Coefficienti della retta di regressione, ossia il valore della pendenza (m) e dell'intercetta (b), i quali definiscono l'equazione della retta

$$y = mx + b$$

- Risultati predittivi stimando nuovi valori di y per determinati valori di x sulla base del modello di regressione lineare ottenuto.

Invece, l'operazione di KNN produce in uscita:

- Classificazione del punto di test, al quale verrà assegnato una classe in base alle etichette della maggioranza dei suoi k punti più vicini.
- Etichette dei k punti più vicini al punto di test, informazioni utile a fini di comprendere le ragioni della classificazione assegnata al punto di test.

2.3 Relazioni intercorrenti

Nella linear regression si cerca di modellare la relazione tra la variabile indipendente x e la variabile dipendente y attraverso una retta. Pertanto la relazione che intercorre tra i dati di ingresso ed uscita per questa operazione è:

$$y = mx + b$$

dove:

- y è la variabile dipendente (o di output)
- x è la variabile indipendente (o di input)
- m è il coefficiente di pendenza, che rappresenta il cambiamento in y rispetto a una variazione unitaria in x
- b è l'intercetta, che rappresenta il valore di y quando x è uguale a zero

Per il calcolo dei coefficienti m e b vengono utilizzate le seguenti formule:

$$m = \frac{Cov(x, y)}{Var(x)}$$
$$b = \bar{y} - m \cdot \bar{x}$$

dove:

- $Cov(c, y)$ rappresenta la covarianza tra x e y, calcolata come media delle deviazioni dei punti (x, y) rispetto alle loro medie. Si calcola usando questa formula:

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- $Var(x)$ rappresenta la varianza di x, data da:

$$\text{var}(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- $\bar{x}$ e $\bar{y}$ sono le medie di x e y, rispettivamente.

Dopo aver calcolato i coefficienti m e b, possiamo utilizzare l'equazione della retta per effettuare delle previsioni per dei nuovi punti

$$\hat{y} = mx + b$$

Dove $\hat{x}$ è il valore predetto di y per il nuovo valore di x.

Per quanto riguarda il KNN, questo si basa su una misura di distanza tra i punti nel dataset. Useremo la distanza euclidea tra due punti (x_1, y_2) e (x_2, y_2), data da:

$$Dist(x_1, y_2, x_2, y_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Il parametro k influisce sulla classificazione dei punti e rappresenta il numero di vicini che verranno considerati. Un valore di k più piccolo comporta una classificazione più influenzata dai punti vicini, mentre una valore più grande porta ad una classificazione più generale basata su una maggiore diversità di vicini. Il punto di test viene poi assegnato alla classe di maggioranza presa dai suoi k punti più vicini. Vengono cioè conteggiate le etichette dei punti vicini, e si assegna al punto l'etichetta avente il maggior numero di occorrenze.

3 Progettazione dell'algoritmo

3.1 Scelte di progetto