Regressione lineare come strumento per il Machine Learning (Apprendimento Automatico)

AI artificial intelligence

Disciplina che si occupa della progettazione di software / hardware che possano prendere decisioni in maniera autonoma.

Machine Learning (ML) è una branca dell'AI.
(TOM MITCHELL)
ML si preoccupa di progettare algoritmi che si perfezionano in maniera automatica con l'esperienza.

L'idea è quella di fare acquisire alla macchina / o al software la capacità di eseguire un compito prefissato (Decidere se un'e-mail è spam o meno; predire il costo di un appartamento ...)

Un primo modo per raggiungere il nostro obiettivo è quello di

1. Scrivere delle regole ben precise da seguire.

2. Fornire dei dati al software e lasciare che il software impari in maniera autonoma dai dati a disposizione.

ML ~~ha~~ affronta due problemi principali

a) CLASSIFICAZIONE
   Es. Spam / Non Spam
       Melanoma / Non melanoma
       Frode / Non frode

b) REGRESSIONE
(Predire il valore di un dato)

Es costo di un appartamento

REGRESSIONE LINEARE

(Altezza padre)

Dati: x_1 —— x_{20} (x_1 —— x_N)

In generale N è molto grande

(Altezza studente)

y_1 —— y_{20} (y_1 —— y_N)

Cerchiamo l'equazione di una retta

$$y = mx + q$$

che approssimi l'andamento dei dati.

Dalla Google Sheet infatti ci è sembrato di riconoscere

una dipendenza di tipo
lineare delle Y dalle x
(scatter plot)

In realtà $\forall\, i = 1, \dots, 20$

$$(*) \quad Y_i = m x_i + q + \underline{\underline{e_i}}$$

e_i : errore per l'i-esimo dato

<u>Obiettivo</u> : cercare m e q

<u>IDEA</u> : li cerco in modo tale
da minimizzare gli errori
commessi.

<u>Def</u> Sia

$$E = \frac{1}{20} \sum_{i=1}^{20} \left(Y_i - (m x_i + q) \right)^2 =$$

$$\frac{1}{20} \sum_{i=1}^{20} e_i^2 \quad \left(\begin{array}{c} \text{Uso l'eq.ne} \\ (*) \end{array} \right)$$

E si dice errore quadratico medio.

OSSERVAZIONE

Non consideriamo gli e_i con il loro segno per evitare che errori con segno opposto si annullino. Prendiamo il quadrato (invece che ad esempio $|e_i|$) per pesare di più gli errori più grandi.

REGRESSIONE LINEARE AI MINIMI QUADRATI: Scegliamo m e q in modo tale da minimizzare E (errore quadratico medio)

Procedura: si considera E come funzione di m e q $E = E(m,q)$ e se ne calcolano le derivate parziali rispetto ad m e q.

Noi invece utilizziamo delle definizioni di statistica.

Def $x_1 — x_{20}$

Si dice media di $x_1 — x_{20}$

$$\overline{x} = \frac{1}{20} \sum_{i=1}^{20} x_i$$

(La media di $y_1 — y_{20}$: $\overline{y} = \frac{1}{20} \sum_{i=1}^{20} y_i$)

Def Si dice varianza di $x_1 — x_{20}$

$$S_x^2 = \frac{1}{20} \sum_{i=1}^{20} (x_i - \overline{x})(x_i - \overline{x})$$

S_x^2 è una misura di quanto gli x_i si discostano dalla media $\overline{x}$.

Def Si dice covarianza di

$x_1 — x_{20}$, $y_1 — y_{20}$

$$S_{xy} = \frac{1}{20} \sum_{i=1}^{20} (x_i - \overline{x})(y_i - \overline{y})$$

Se $\quad y_i = m x_i + q \quad i = 1, \dots, 20$

(se gli e_i fossero tutti nulli, se (x_i, y_i) sono tutti sulla retta di equazione $y = mx + q$)

1. $\bar{y} = \frac{1}{20} \sum_{i=1}^{20} y_i = \frac{1}{20} \sum_{i=1}^{20} (m x_i + q) =$

$= \frac{1}{20} \sum_{i=1}^{20} m x_i + \frac{1}{20} \sum_{i=1}^{20} q =$

$= m \cdot \frac{1}{20} \sum_{i=1}^{20} x_i + q = m \bar{x} + q$

2. $S_{xy} = \frac{1}{20} \sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}) =$

$y_i = m x_i + q$
$\bar{y} = m \bar{x} + q$

$\frac{1}{20} \sum_{i=1}^{20} (x_i - \bar{x})(m x_i + \cancel{q} - m \bar{x} - \cancel{q}) =$

$= \frac{1}{20} \sum_{i=1}^{20} (x_i - \bar{x}) m (x_i - \bar{x}) =$

$= m \frac{1}{20} \sum_{i=1}^{20} (x_i - \bar{x})(x_i - \bar{x}) =$

$= m S_x^2$

Dalla 2. $S_{XY} = m S_X^2$

$\Rightarrow \boxed{m = \dfrac{S_{XY}}{S_X^2}}$

Dall 1. $\bar{Y} = m\bar{X} + q$

$\Rightarrow$ Posso ricavare q da $\bar{Y}, m, \bar{X}$

$$\boxed{q = \bar{Y} - m\bar{X}}$$

Se $Y_i = mX_i + q$ possiamo ricavare m e q da $\bar{X}, \bar{Y}, S_X^2, S_{XY}$

TEOREMA (X_i, Y_i), $i = 1, \ldots, 20$

La retta di regressione lineare (quella che minimizza l'errore medio quadratico) è

$$Y = mX + q$$

con $\boxed{m = \frac{S_{xy}}{S_x^2}}$, $\boxed{q = \bar{y} - m\bar{x}}$

dove $\bar{x}, \bar{y}, S_{xy}, S_x^2$ sono state definite precedentemente.

Con i nostri dati $\bar{x} \simeq 179.85$

$S_x^2 \simeq 42.63$

$S_{xy} \simeq 37.63$

$\bar{y} \simeq 145.2$

$$m = \frac{S_{xy}}{S_x^2} \simeq 0.8828$$

$$q = \bar{y} - m\bar{x} \simeq 145.2 - 0.8828\,(179.85)$$
$$\simeq -13.56$$

$$y = m_1 x_1 + m_2 x_2 + q$$

Si risolve un opportuno sistema

lineare per trovare m_1 m_2 e q

(MINIMI QUADRATI
REGRESSIONE MULTILINEARE)

cinzia.elia@uniba.it