UNIVERSITÀ DEGLI STUDI DELLA TUSCIA

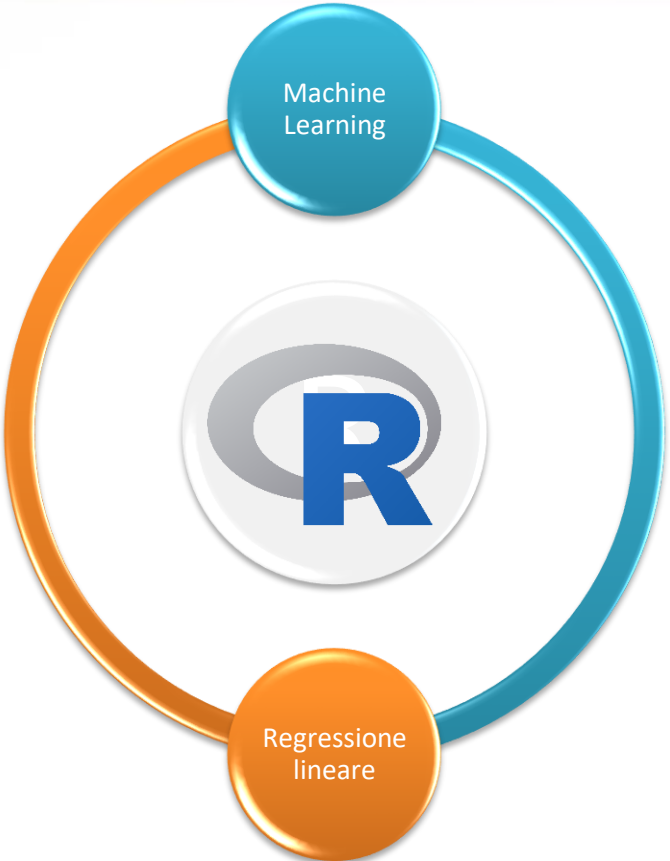
INFORMATICA

Linguaggio R
Fondamenti di Machine Learning

Dott. Franco Liberati
franco.liberati@unitus.it

LINGUAGGIO R

Argomenti del corso



LINGUAGGIO R

MACHINE LEARNING

- Il Machine Learning, o apprendimento automatico, è un ramo dell'Intelligenza Artificiale, cioè quella disciplina che studia e progetta sistemi che imitano l'intelligenza umana.

- L'apprendimento automatico è un ramo dell'informatica che studia i sistemi e gli algoritmi in grado di apprendere nuova conoscenza attraverso l'analisi di una collezione di dati detti caratteristiche (feature).

LINGUAGGIO R

MACHINE LEARNING

- Per valutare la bontà di un algoritmo di apprendimento automatico, si deve stimare una misura quantitativa della sua prestazione. Spesso la misura è specifica per un certo compito ma, nei casi più generali, si prende come parametro di riferimento il **tasso di errore**, definito come il rapporto tra predizioni sbagliate e dati analizzati

- Studiando il tasso di errore, l'autoapprendimento può essere calibrato con una continua opera di correzione in modo da ridurre gli errori di valutazione

LINGUAGGIO R

MACHINE LEARNING

- Un’altra distinzione che può essere operata tra gli algoritmi riguarda i principi su cui basano la tecnica di apprendimento. Tra i più noti ci sono:
  - Algoritmo di regressione (regression), basato cioè sulla tecnica statistica che si utilizza per studiare la relazione tra due o più variabili. In particolare questa famiglia di procedimenti individua il modo in cui da variabili indipendenti si possa prevedere il comportamento di variabili dipendenti.
  - Algoritmi di raggruppamento (clustering), che ammassano elementi in classi omogenee. Il loro scopo può avere una doppia valenza: da un lato permettono di individuare gruppi di oggetti accomunati da vari attributi, dall’altro consentono di riconoscere gli elementi non associabili a gruppi (outlier) rivelando, così, situazioni problematiche, casi anomali o aspetti interessanti da approfondire.

LINGUAGGIO R

MACHINE LEARNING

- Un’altra distinzione che può essere operata tra gli algoritmi riguarda i principi su cui basano la tecnica di apprendimento. Tra i più noti ci sono:
  - Alberi decisionali, cioè dei grafi particolari che collegano i vertici dei dati ai vertici risultati mediante percorsi generati in relazione ad alcune condizioni.
  - Algoritmi Bayesiani, che determinano (mediante funzioni statistiche) la probabilità di appartenenza di un elemento ad una determinata classe.
  - Reti neurali (denoising), che generano grafi in cui ogni nodo è un elemento di computazione più o meno complesso capace di apprendere e di riprodurre uscite per ingressi sconosciuti (cioè di acquisire capacità di generalizzazione). Il loro valore si apprezza particolarmente nei casi con dati caratterizzati da forte ‘rumore’ o da una molteplicità di informazione eterogenea che ne ostacola una classificazione.

LINGUAGGIO R

MACHINE LEARNING

- Il processo di apprendimento automatico richiede una serie di passi ben distinti:
  1. la scelta dell'algoritmo di auto apprendimento e la definizione dei parametri iniziali;
  2. la preparazione dei dati (es.: applicare un processo di normalizzazione, cioè selezionare alcuni valori in relazione ad un certo intervallo numerico o precise caratteristiche; ridurre il rumore; discretizzare variabili continue; …);
  3. suddivisione dei dati per la fase di apprendimento (training set) e quelli inerenti alla verifica del modello prodotto (test set);
  4. l'applicazione nella fase di training;
  5. la prova del modello derivato dall'auto apprendimento (applicando il test set).

- Le fasi possono essere ripetute più volte nel tentativo di migliorare l'accuratezza del modello cambiando l'algoritmo o modulando diversamente i parametri.

Regressione Lineare

LINGUAGGIO R

REGRESSIONE LINEARE

- La regressione lineare è una tecnica statistica che mira ad individuare la correlazione tra una variabile dipendente ed una o più variabili indipendenti
- Acquisendo (un numero enorme) di dati, attraverso la regressione lineare, si possono ottenere i parametri che permettono di individuare il valore atteso della variabile dipendente rispetto a quella o quelle indipendenti.

LINGUAGGIO R

REGRESSIONE LINEARE

- Nella forma più semplice si cercano i valori a_0 e a_1 di una equazione del tipo

$$y=a_0x+a_1 \text{ (con x variabile indipendente e y dipendente)}$$

detta retta approssimante, che accosta i valori di x_i correlati a y_i appartenenti al training set e consente di derivare y_a da un x_a del test set con un errore accettabile.

LINGUAGGIO R

REGRESSIONE LINEARE: esempio (dominio)

- Come dimostrazione si considera la collezione di dati faithful disponibile nel dataset del Linguaggio R, che ha informazioni sulla durata di emissione di un gayser [eruptions] e il tempo di quiescienza dopo il soffio [waiting]

Dai valori campionati si può essere interessati a comprendere se esiste un modo di determinare il tempo di quiescenza in relazione al tempo di emissione della sorgente di calore, in pratica trattando il primo parametro come variabile dipendente dalla seconda (ad esempio, comprendere se più l'eruzione è duratura allora è maggiore il tempo di riposo; o viceversa)

```
print(faithful)
   eruptions waiting
01     3.600      79
02     1.800      54
03     3.333      74
04     2.283      62
05     4.533      85
06     2.883      55
07     4.700      88
08     3.600      85
09     1.950      51
10     4.350      85
```

LINGUAGGIO R

REGRESSIONE LINEARE: esempio (dominio)

- Un modo per analizzare il tipo di relazione è la modalità grafica (sfruttando, ad esempio, una rappresentazione a nuvola di punti delle due variabili). Qualora sussista la condizione, allora si deve individuare la retta approssimante (cioè la retta intorno alla quale sono presenti i valori) la cui equazione è del tipo $y=a_0+a_1x$, dove a_0 è l'**intercetta**, cioè il punto in cui la retta interseca l'asse delle y, mentre a_1 è il **coefficiente angolare**, cioè la sua pendenza.

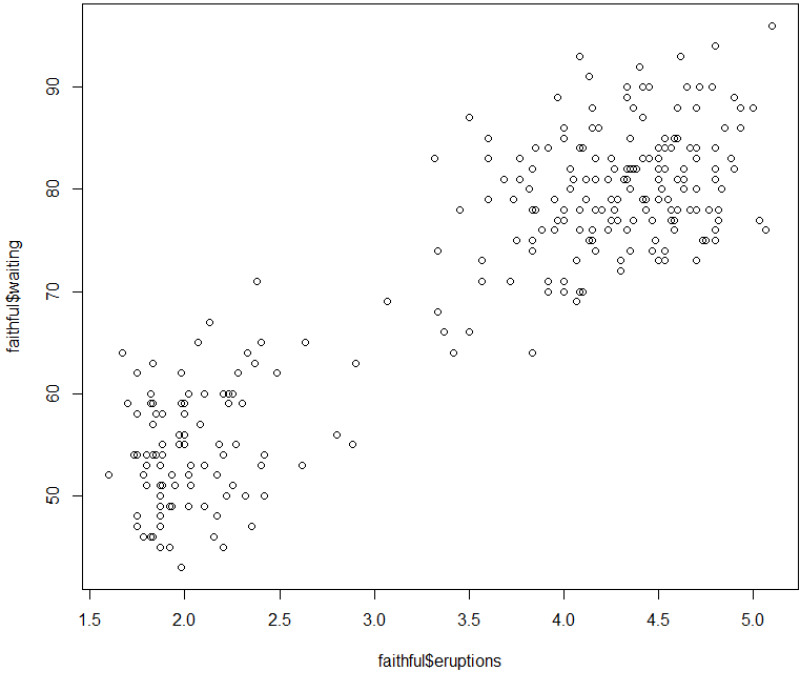
- La regressione lineare è applicabile se sussiste una relazione pressoché 'lineare' tra le due variabili
- Lo scopo della regressione lineare, infatti, è quello di individuare il coefficiente angolare e la intercetta della retta approssimante. Una volta ottenuti questi dati, si possono predire valori al di fuori dei dati a disposizione

LINGUAGGIO R

REGRESSIONE LINEARE: esempio (I fase)

- Come prima fase si analizza la linearità dei dati

```
#Uso del package
library (caret)
plot(faithful$eruptions,faithful$waiting)
```



LINGUAGGIO R

REGRESSIONE LINEARE: esempio (II fase)

- I valori delle due variabili hanno una disposizione lineare, pertanto si procede ad individuare i parametri della retta di regressione attraverso un meccanismo di training e quindi selezionando preliminarmente i dati del set di apprendimento
- Nel caso in esame si selezionano i primi novanta campioni, circa un terzo dei rilevamenti, omettendo i casi in cui dei campioni non siano rilevati

```
#Selezione dei primi 90 record
good_data<-c(1:90)
#Esclusione dei record con valori non
significativi o non campionati
training_set<-na.omit((faithful[good_data,]))
```

LINGUAGGIO R

REGRESSIONE LINEARE: esempio (II fase)

- Dopo aver selezionato i dati si attiva la fase di auto-apprendimento con il quale l'algoritmo 'impara' dai dati (Script XI.2a) sfruttando la funzione:

train(form, data, ..., method="lm", subset, na.action = na.fail, contrasts = NULL**)**

dove form è la formula applicata, cioè si usa l'operatore ~ per stabilire quali sono le variabili dipendenti (cioè la y, a sinistra del simbolo) e quelle indipendenti (cioè la x. alla destra del simbolo; se le variabili indipendenti sono più di una si usa il simbolo +); mentre data sono il vettore dei dati da analizzare; e il parametro method specifica quale algoritmo di learning è applicato (es.: lm, sta per linear model).

LINGUAGGIO R

REGRESSIONE LINEARE: esempio (II fase)

```
good_data<-c(1:90)
training_set<-na.omit((faithful[good_data,]))
#Fase di training
modello <- train(waiting~eruptions, training_set, method="lm")
```

La funzione restituisce una lista di informazione analitiche in cui l'elemento modello contiene coefficiente e intercetta

```
intercetta <- coef(modello$finalModel)[1]
coef <- coef(modello$finalModel)[2]
print(intercetta)
print(coef)
(Intercept)
  37.9371
   eruptions
9.5533
```

LINGUAGGIO R

REGRESSIONE LINEARE: esempio (II fase)

I parametri completi presenti nella lista sono reperibili con la funzione summary() o dal pannello Enviroment di R-Studio

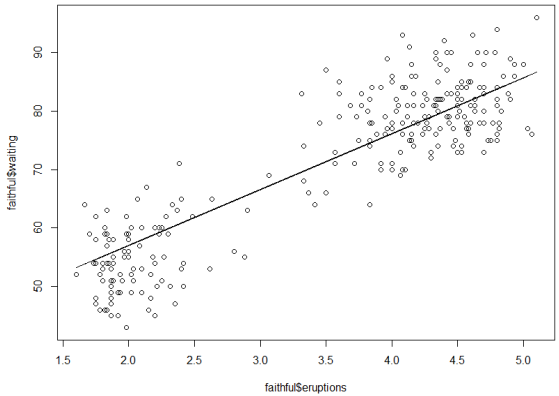
```
#Selezione dei primi 60 record con valori significativi
good_data<-c(1:90)
training_set<-na.omit((faithful[good_data,]))
#Fase di training
modello <- train(waiting~eruptions, training_set, method="lm")
print(summary(modello))
(Intercept)
   37.93711
eruptions
 9.553265
Call:
lm(formula = .outcome ~ ., data = dat)
Residuals:
     Min       1Q   Median       3Q      Max
-10.6390  -4.1080  -0.2335   3.4871  13.3747
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.9371     1.9157    19.8   <2e-16 ***
eruptions     9.5533     0.5248    18.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.784 on 88 degrees of freedom
Multiple R-squared:  0.7902,                    Adjusted R-squared:  0.7878
F-statistic: 331.4 on 1 and 88 DF,  p-value: < 2.2e-16
```

LINGUAGGIO R

REGRESSIONE LINEARE: esempio (III fase)

Una possibile rappresentazione grafica si può ottenere con la funzione plot(), che riporta i campioni, e con lines(), che consente la sovrapposizione della retta di approssimazione ricavata dal procedimento di regressione lineare

```
library (caret)
print(faithful)
#Selezione dei primi 90 record con valori significativi
good_data<-c(1:90)
training_set<-na.omit((faithful[good_data,]))
#Fase di training
modello <- train(waiting~eruptions, training_set, method="lm")
intercetta <- coef(modello$finalModel)[1]
coef <- coef(modello$finalModel)[2]
#Sovrapposizione linea di regressione
plot(faithful$eruptions,faithful$waiting)
lines(faithful$eruptions, (faithful$eruptions*coef)+intercetta)
```



LINGUAGGIO R

REGRESSIONE LINEARE: esempio (III fase)

La retta di approssimazione, come si vede, rappresenta un compromesso dell'andamento derivato dai campioni rilevati.

È possibile usare i risultati del modello predittivo per stimare eventuali comportamenti applicando la formula geometrica.

$$\text{quiescenza} = (\text{coef} \cdot \text{tempo\_eruzione}) + \text{intercetta}$$

Ad esempio il record 226 di faithful (eruptions 4.117 waiting 79) è in linea con quanto appreso: $(4.117 \cdot 9.5533) + 37.9371 = 77.27$ a fronte del valore reale campionato 79


Fine