Fondamenti di Machine Learning

Data Preprocessing

Features Encoding

presentato da
Giuseppe Gullo

PROFESSION AI

Codifichiamo le variabili qualitative

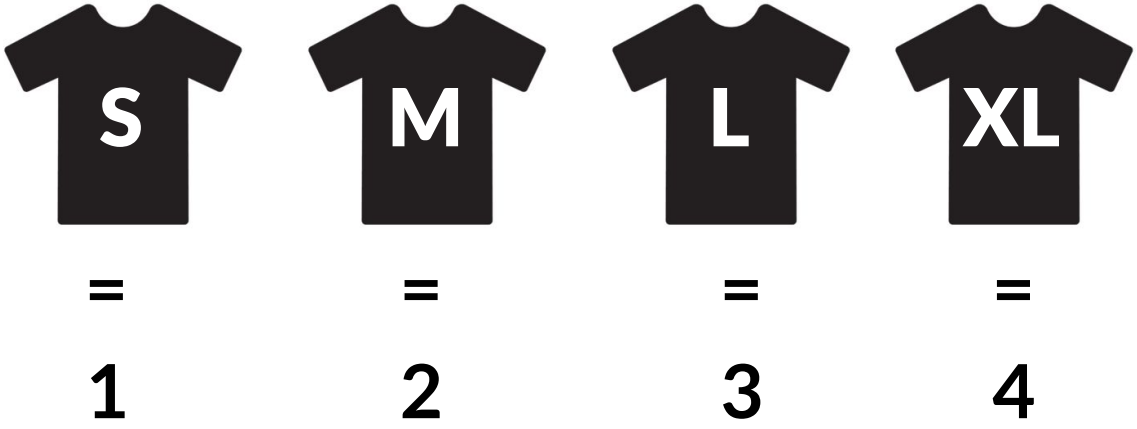
COLORE	TAGLIA	PREZZO
Rosso	S	9.90
Bianco	M	14.90
Verde	L	12.90
Rosso	XL	14.90

Label Encoding

I label sono i possibili valori non numerici che una variabile qualitative può avere.

Label Encoding

Il label encoding consiste nel mappare ogni label ad un valore numerico



Label Encoding

COLORE	TAGLIA	PREZZO
Rosso	S	9.90
Bianco	M	14.90
Verde	L	12.90
Rosso	XL	14.90

Label Encoding

COLORE	TAGLIA	PREZZO
Rosso	1	9.90
Bianco	2	14.90
Verde	3	12.90
Rosso	4	14.90

One-Hot Encoding

Sostituiamo la colonna con una nuova colonna per ogni possibile valore della variabile.

COLORE_ROSSO	COLORE_BIANCO	COLORE_VERDE	TAGLIA	PREZZO
			1	9.90
			2	14.90
			3	12.90
			4	14.90

One-Hot Encoding

Per ogni osservazione, inseriamo *vero* nella colonna corrispondente al valore, *falso* nelle altre

COLORE_ROSSO	COLORE_BIANCO	COLORE_VERDE	TAGLIA	PREZZO
vero	falso	falso	1	9.90
falso	vero	falso	2	14.90
falso	falso	vero	3	12.90
vero	falso	falso	4	14.90

One-Hot Encoding

In fase di costruzione del modello,
questi valori booleani vengono utilizzati come 0 o 1

COLORE_ROSSO	COLORE_BIANCO	COLORE_VERDE	TAGLIA	PREZZO
1	0	0	1	9.90
0	1	0	2	14.90
0	0	1	3	12.90
1	0	0	4	14.90

Fondamenti di Machine Learning

Data Preprocessing

Feature Scaling

presentato da
Giuseppe Gullo

PROFESSION AI

Dataset di vini

ALCOL	FLAVONOIDI
14.23	3.06
13.20	2.76
13.16	3.24
14.37	3.49
13.24	2.69

Dataset di vini

ALCOL	FLAVONOIDI
14.23	3.06
13.20	2.76
13.16	3.24
14.37	3.49
13.24	2.69

	ALCOL	FLAVONOIDI
MIN	13.16	2.69
MAX	14.37	3.49
MEAN	13.64	3.05

Perché portare i dati sulla stessa scala?

1. Avere le feature su una scala comune
   potrebbe rendere più veloce il processo di addestramento.
2. A feature con magnitudine maggiore
   potrebbe essere associato un peso maggiore.

Perché portare i dati sulla stessa scala?

1. Avere le feature su una scala comune
   potrebbe rendere più veloce il processo di addestramento.
2. A feature con magnitudine maggiore
   potrebbe essere associato un peso maggiore.

Modelli scale-invariant
La scala dei dati non ha alcuna rilevanza.

Normalizzazione

Portiamo tutte le variabili in una scala che va da 0 a 1

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{\min}}{x_{\max} - x_{min}}$$

Normalizzazione

Portiamo tutte le variabili in una scala che va da 0 a 1

ALCOL	FLAVONOIDI
14.23	3.06
13.20	2.76
13.16	3.24
14.37	3.49
13.24	2.69

Normalizzazione

Portiamo tutte le variabili in una scala che va da 0 a 1

ALCOL	FLAVONOIDI
14.23	3.06
13.20	2.76
13.16	3.24
14.37	3.49
13.24	2.69

$$x^{(i)}_{norm} = \frac{x^{(i)} - x_{\min}}{x_{\max} - x_{min}}$$

Normalizzazione

ALCOL	FLAVONOIDI
14.23	3.06
13.20	2.76
13.16	3.24
14.37	3.49
13.24	2.69

$$\frac{14.23 - 13.16}{14.37 - 13.16} = 0.88$$

Normalizzazione

ALCOL	FLAVONOIDI
0.88	3.06
13.20	2.76
13.16	3.24
14.37	3.49
13.24	2.69

$$\frac{14.23 - 13.16}{14.37 - 13.16} = 0.88$$

Normalizzazione

ALCOL	FLAVONOIDI
0.88	0.46
0.33	0.09
0	0.69
1	1
0.067	0

Standardizzazione

Portiamo tutte le variabili in una distribuzione normale, cioè con media 0 e deviazione standard 1.

$$x_{std}^{(i)} = \frac{x^{(i)} - x_{mean}}{x_{sd}}$$

Standardizzazione

ALCOL	FLAVONOIDI
14.23	3..06
13.20	2.76
13.16	3.24
14.37	3.49
13.24	2.69

$$x_{std}^{(i)} = \frac{x^{(i)} - x_{mean}}{x_{sd}}$$

Standardizzazione

ALCOL	FLAVONOIDI
14.23	3.06
13.20	2.76
13.16	3.24
14.37	3.49
13.24	2.69

$$\frac{14.23 - 13.64}{0.54} = 1.09$$

Standardizzazione

ALCOL	FLAVONOIDI
1.09	3.06
13.20	2.76
13.16	3.24
14.37	3.49
13.24	2.69

$$\frac{14.23 - 13.64}{0.54} = 1.09$$

Standardizzazione

ALCOL	FLAVONOIDI
1.09	0.04
-0.81	-0.97
-0.89	0.64
1.35	1.49
-0.74	-1.20

Quale metodo utilizzare? Dipende

Normalizzazione

- Utile per algoritmi che non fanno assunzioni sulla distribuzione (es. KNN e Reti Neurali).
- Utile per le immagini, dove ogni pixel va da 0 a 1.

Standardizzazione

- Utile quando i dati seguono una distribuzione Gaussiana.
- Mantiene le informazioni sugli outlier.

Fondamenti di Machine Learning

Data Preprocessing

Gestire dati mancanti

presentato da
Giuseppe Gullo

PROFESSION AI

T-shirt disponibili in un negozio di abbigliamento

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	14.90
Dolci & Gabbiani	Verde	L	12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde	M	12.90
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	XL	14.90
Dolci & Gabbiani	Rosso	S	12.90

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso		9.90
Cucci	Bianco	M	14.90
Dolci & Gabbiani	Verde	L	12.90
Cucci	Rosso		14.90
Cucci	Bianco		14.90
Dolci & Gabbiani	Verde		12.90
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	XL	14.90
Dolci & Gabbiani	Rosso		12.90

Oltre il 50% dei valori mancanti per una colonna

MARCA	COLORE	PREZZO
Alvaro Vitali	Rosso	9.90
Cucci	Bianco	14.90
Dolci & Gabbiani	Verde	12.90
Cucci	Rosso	14.90
Cucci	Bianco	14.90
Dolci & Gabbiani	Verde	12.90
Alvaro Vitali	Rosso	9.90
Cucci	Bianco	14.90
Dolci & Gabbiani	Rosso	12.90

Droppiamola!

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	14.90
	Verde		12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde	M	12.90
		S	9.90
Cucci	Bianco	XL	14.90
Dolci & Gabbiani			12.90

Oltre il 50% dei valori mancanti per una riga

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	14.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde	M	12.90
Cucci	Bianco	XL	14.90

Droppiamola!

Problema con la rimozione

Potremmo perdere tanta informazione

Variabili continue: Sostituzione con media/mediana

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	
Dolci & Gabbiani	Verde	L	12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde	M	
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	XL	14.90
Dolci & Gabbiani	Rosso	S	12.90

Variabili continue: Sostituzione con media/mediana

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	
Dolci & Gabbiani	Verde	L	12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde	M	
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	XL	14.90
Dolci & Gabbiani	Rosso	S	12.90

$$\frac{9.90 + 12.90 + 14.90 + 14.90 + 9.90 + 14.90 + 12.90}{7} = 12.90$$

Variabili continue: Sostituzione con media/mediana

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	**12.90**
Dolci & Gabbiani	Verde	L	12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde	M	**12.90**
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	XL	14.90
Dolci & Gabbiani	Rosso	S	12.90

$$\frac{9.90 + 12.90 + 14.90 + 14.90 + 9.90 + 14.90 + 12.90}{7} = 12.90$$

Variabili qualitative: Sostituzione con valore più frequente

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	14.90
Dolci & Gabbiani	Verde		12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde		12.90
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	XL	14.90
Dolci & Gabbiani	Rosso	S	12.90

Variabili qualitative: Sostituzione con valore più frequente

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	14.90
Dolci & Gabbiani	Verde		12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde		12.90
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	XL	14.90
Dolci & Gabbiani	Rosso	S	12.90

S: 3
M: 1
L: 1
XL: 2

Variabili qualitative: Sostituzione con valore più frequente

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	14.90
Dolci & Gabbiani	Verde	S	12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde	S	12.90
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	XL	14.90
Dolci & Gabbiani	Rosso	S	12.90

S: 3
M: 1
L: 1
XL: 2

Variabili qualitative: Creazione di una nuova categoria

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	14.90
Dolci & Gabbiani	Verde		12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde		12.90
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco		14.90
Dolci & Gabbiani	Rosso		12.90

Variabili qualitative: Creazione di una nuova categoria

MARCA	COLORE	TAGLIA	PREZZO
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	M	14.90
Dolci & Gabbiani	Verde	**U**	12.90
Cucci	Rosso	XL	14.90
Cucci	Bianco	L	14.90
Dolci & Gabbiani	Verde	**U**	12.90
Alvaro Vitali	Rosso	S	9.90
Cucci	Bianco	**U**	14.90
Dolci & Gabbiani	Rosso	**U**	12.90

Problema con la sostituzione

Data leakage

Diamo al modello informazioni che non dovrebbe avere