Lezione 5: Matrice di confusione e metriche

Davide Evangelista e Dario Lanzoni
dario.lanzoni6@unibo.it

Università di Bologna

3 Aprile 2025

Metriche

Il modo più semplice (e intuitivo) per classificare un modello è quello di definire una o più metriche, in grado di valutare l'abilità di un modello nel fare predizioni.

Definizione

Una metrica è una qualunque funzione $\rho : \mathbb{R}^m \times \mathbb{R}^m \longrightarrow \mathbb{R}$, *tale che:*

- $\rho(y, y') \geq 0$.
- $\rho(y, y') = \rho(y', y)$.
- $\rho(y, y') = 0 \iff y = y'$

Intuitivamente, una metrica rappresenta una distanza tra la l'output corretto e la predizione fatta da $h_\theta(x)$, ed è tanto più piccola quanto più il predittore è corretto.

MSE

Molte delle misure statistiche che avete visto (e vedrete) a lezione sono delle metriche.
Una delle più comuni, è il cosiddetto *Mean Squared Error (MSE)*, la quale, fissato un dataset $\mathbb{D} = \{x_i, y_i\}_{i=1}^{N}$, dove $\{x_i\}_{i=1,\dots,N}$ sono le variabili input e $\{y_i\}_{i=1,\dots,N}$ variabili di output, è data da

$$MSE(\mathbb{D}) = \frac{1}{N} \sum_{i=1}^{N} ||y_i - h_\theta(x_i)||_2^2, \tag{1}$$

dove h_θ è un predittore. Questa, sebbene molto utilizzata nella pratica, non è la più adatta per problemi di classificazione, poiché in questo caso la variabile di output y è di tipo Categorico, e l'MSE non sfrutta in maniera efficiente questa caratteristica.

Misclassification Error (ME)

Una misura molto più comune per problemi di classificazione è l'errore di Misclassificazione (ME), definito come il numero totale di punti mal classificati dal nostro predittore. Formalmente, definita la funzione

$$I(y \neq y') = \begin{cases} 1 & \text{se } y \neq y' \\ 0 & \text{se } y = y' \end{cases} \tag{2}$$

Si definisce l'errore di misclassificazione come

$$ME(\mathbb{D}) = \sum_{i=1}^{N} I(y_i \neq h_\theta(x_i)) \tag{3}$$

Misclassification Rate (MR)

In realtà, spesso, al posto dell'errore di misclassificazione, si utilizza la frequenza di misclassificazione (MR), definita come il numero medio di predizioni accurate che vengono fatte dal predittore. Formalmente

$$MR(\mathbb{D}) = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq h_\theta(x_i)) \tag{4}$$

Si osservi che, nel caso di classificazione binaria (in cui la variabile di output y assume valori in $\{0; 1\}$), la frequenza di misclassificazione e l'MSE sono equivalenti.
Moltiplicando per 100 la MR, si ottiene la cosiddetta percentuale di misclassificazione $M\% = 100 \cdot MR$, che dice quanti punti in percentuale sono stati mal classificati dal modello.

Accuratezza

Strettamente collegata con la frequenza di misclassificazione, è la cosiddetta Accuratezza, che pur non essendo propriamente una metrica (l'accuratezza è tanto più alta quanto migliore è il modello), viene spesso usata per valutare il modello. L'accuratezza è definita come:

$$Acc(\mathbb{D}) = \frac{1}{N} \sum_{i=1}^{N} I(y_i = h_\theta(x_i) = 1 - MR(\mathbb{D}) \tag{5}$$

Implementazione

Supponendo di avere un modello già addestrato, chiamato SVM, calcolare la frequenza di misclassificazione con i seguenti comandi

```
y_pred = model.predict(df[['x1','x2']])
y_true = df['class']

MR = np.mean(y_pred != y_true)
```

Similmente, l'accuratezza può essere facilmente calcolata come

```
acc = 1 - MR
```

Matrice di Confusione

In alcune situazioni (ad esempio, in problemi di classificazione medici), non tutti gli errori hanno la stessa importanza. Nell'esempio di classificazione medica, sbagliare una predizione e diagnosticare ad un paziente una malattia che in realtà non ha è poco grave: basteranno altri semplici esami per identificare l'errore e risolvere il problema, il paziente si prenderà al massimo un po' di paura.
Invece, sbagliare una predizione dicendo ad un paziente malato che è sano, potrebbe avere delle conseguenze pericolose: il paziente non si curerà e potrà incorrere in problemi di vario genere.

TP, FP, TN, FN

Identifichiamo quattro tipologie di predizioni per modelli binari (in cui l'output y assume valori 0 (Negativo) e 1 (Positivo)), due delle quali saranno predizioni corrette, mentre le altre due saranno errori.

- Veri Positivi (TP): Il modello predice 1 (Positivo), e l'output corretto è effettivamente 1 (Positivo).
- Falsi Positivi (FP): Il modello predice 1 (Positivo), ma l'output corretto è 0 (Negativo).
- Veri Negativi (TN): Il modello predice 0 (Negativo), e l'output corretto è effettivamente 0 (Negativo).
- Falsi Negativi (FN): Il modello predice 0 (Negativo), ma l'output corretto è 1 (Positivo).

TP, FP, TN, FN

Chiaramente, una predizione di tipo Vero Positivo o Vero Negativo (TP o TN) è una predizione corretta, quindi non ci crea alcun tipo di problema.
I Falsi Positivi e i Falsi Negativi, invece, rappresentano gli errori di predizione e possono avere valenza diversa. Nell'esempio predecente sui dati medici, i FN sono molto più pericolosi dei FP!
Queste definizioni di permettono di definire alcune metriche molto usate nella pratica e che permettono di distinguere la tipologia di errore compita dal modello.

Sensitività

Definiamo la **Sensitività (o TPR, True Positive Rate)** di un modello la grandezza

$$TPR = \frac{TP}{TP + FN} = 1 - FNR \tag{6}$$

Come la percentuale di Veri Positivi sul totale delle osservazioni Positive (calcolate come $TP + FN$). Il *contrario* della TPR è la **FNR, o False Negative Rate**, calcolata come $1 - TPR$, che rappresenta la percentuale di Falsi Negativi sul totale delle osservazioni Positive. Una TPR alta rappresenta un modello che, quando predice esito positivo, è molto probabilmente un Vero Positivo. La TPR non da alcuna informazione su come si comporta il modello con le osservazioni Negative.

Specificità

La **Specificità (o TNR, True Negative Rate)** è la grandezza equivalente alla Sensitività, ma nel caso negativo, ovvero

$$TNR = \frac{TN}{TN + FP} = 1 - FPR \tag{7}$$

Che conta la percentuale di Veri Negativi sul totale delle osservazioni Negative (calcolate come $TN + FP$. Il *contrario* della TNR è la **FPR, o False Positive Rate**, calcolata come $1 - TNR$, che rappresenta la percentuale di Falsi Positivi sul totale delle osservazioni Negative.
Una TNR alta rappresenta un modello che, quando predice esito Negativo, è molto probabilmente un Vero Negativo. La TNR non da alcuna informazione su come si comporta il modello con le osservazioni Positive.

Precisione

La **Precisione** di un modello è la grandezza

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

che conta la percentuale di Veri Positivi sul totale delle osservazioni predette come Positive. Una PPV alta indica un modello che predice un numero di Falsi Positivi bassi rispetto al numero dei Veri Positivi.

Matrice di Confusione

Come già osservato, la caratteristica negativa di queste metriche è il fatto che, sebbene esse riescano a distinguere le varie tipologie di errore, in genere una misura da sola non è sufficiente a spiegare la resa effettiva del modello, e bisogna confrontare tra loro varie grandezze per identificare la bontà della previsione.
Per risolvere questo problema, è spesso comodo visualizzare la Matrice di Confusione, che permette di visualizzare in un colpo d'occhio la resa del modello in termini di TP, FP, TN, FN.
La Matrice di Confusione viene implementata in R semplicemente con il comando

```
from sklearn.metrics import confusion_matrix
y_pred = model.predict(df[['x1','x2']])
y_true = df['class']

confusion_matrix = confusion_matrix(y_true, y_pred)
```

Matrice di Confusione

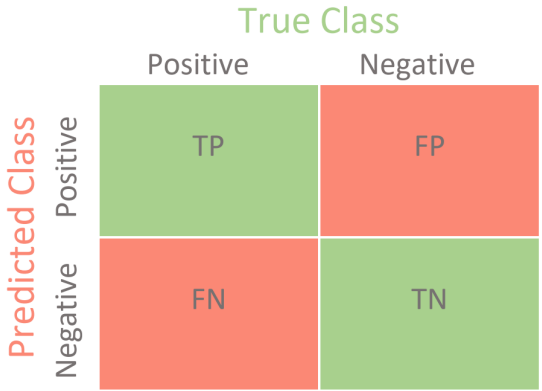

Figura: Matrice di Confusione