Metodi Statistici per le decisioni

2024-2025

Vincenzo Nardelli

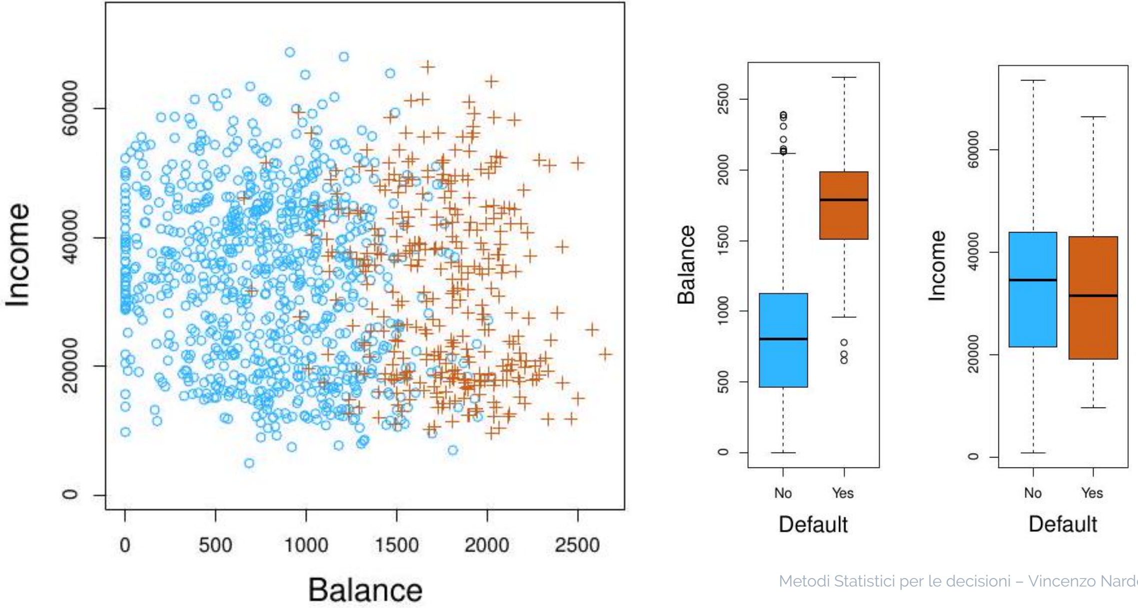
vincenzo.nardelli@unicatt.it

Classificazione

- Le variabili qualitative assumono valori in un insieme non ordinato $\mathcal{C}$, come ad esempio:
  colore degli occhi $\in$ { marrone, blu, verde } email $\in$ { spam, ham}.
- Dato un vettore di caratteristiche X e una risposta qualitativa Y che assume valori nell'insieme $\mathcal{C}$, il compito di classificazione consiste nel costruire una funzione C(X) che prenda in input il vettore X e ne predica il valore per Y; ovvero $C(X) \in \mathcal{C}$.
- Spesso siamo più interessati a stimare le probabilità che X appartenga a ciascuna categoria in $\mathcal{C}$.

Ad esempio, è più utile avere una stima della probabilità che una richiesta di assicurazione sia fraudolenta, piuttosto che una classificazione "fraudolenta" o "non fraudolenta".

Esempio Carta di Credito



Perchè la regressione lineare non basta?

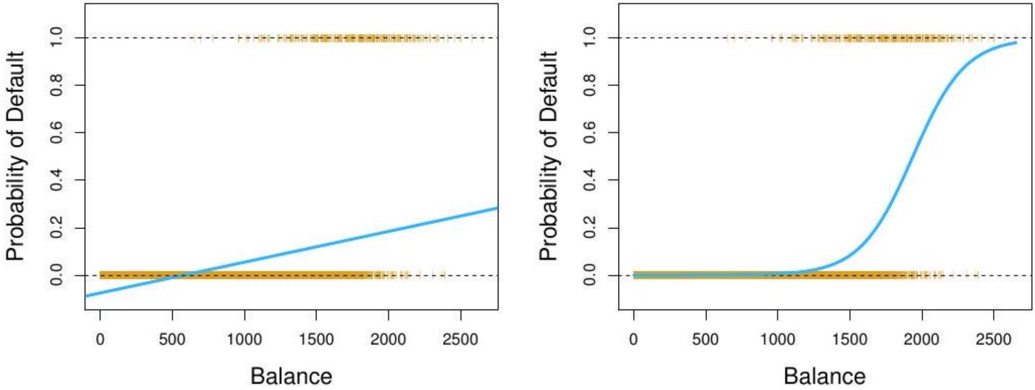
Supponiamo per il compito di classificazione Default di codificare

$$Y = \begin{cases} 0 & \text{se No} \\ 1 & \text{se Sì.} \end{cases}$$

Possiamo semplicemente eseguire una regressione lineare di Y su X e classificare come Sì se $\hat{Y} > 0.5$?

- In questo caso di risultato binario, la regressione lineare svolge un buon lavoro come classificatore ed è equivalente all'analisi discriminante lineare che discuteremo più avanti.
- Poiché nella popolazione $E(Y \mid X = x) = \Pr(Y = 1 \mid X = x)$, si potrebbe pensare che la regressione sia perfetta per questo compito.
- Tuttavia, la regressione lineare potrebbe produrre probabilità inferiori a zero o superiori a uno. La regressione logistica è più appropriata.

Regressione Lineare VS Regressione Logistica



I segni arancioni indicano la risposta Y, 0 o 1. La regressione lineare non stima bene Pr(Y = 1 | X). La regressione logistica sembra adatta al compito.

Regressione Logistica

Scriviamo $p(X) = \Pr(Y = 1 \mid X)$ per semplicità e consideriamo l'uso del bilancio per prevedere il default. La regressione logistica usa la forma

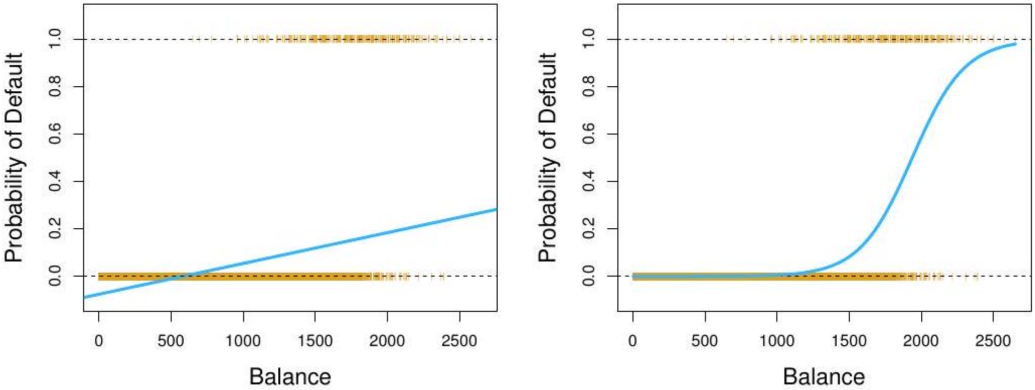
$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

($e \approx 2.71828$ è una costante matematica [numero di Eulero]). È facile vedere che, indipendentemente dai valori di β_0, β_1 o X, p(X) avrà valori compresi tra 0 e 1.
Un po' di riorganizzazione dà

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

Questa trasformazione monotona è chiamata trasformazione log-odds o logit di p(X).
(con log intendiamo logaritmo naturale: ln).

Regressione Lineare VS Regressione Logistica



La regressione logistica garantisce che la nostra stima per p(X) sia compresa tra 0 e 1.

Massima Verosimiglianza

Usiamo la massima verosimiglianza per stimare i parametri.

$$\ell\left(\beta_0, \beta\right) = \prod_{i:y_i=1} p\left(x_i\right) \prod_{i:y_i=0} \left(1 - p\left(x_i\right)\right)$$

Questa funzione di verosimiglianza dà la probabilità degli zeri e degli uno osservati nei dati. Scegliamo β_0 e β_1 per massimizzare la probabilità dei dati osservati.

Massima Verosimiglianza

La maggior parte dei pacchetti statistici è in grado di fare il fit di modelli di regressione logistica lineare usando la massima verosimiglianza. In R, utilizziamo la funzione 'glm'.

	Coefficiente	Errore Std.	Statistica Z	Valore-P
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Log-Odds: Coefficienti, Formula e Calcolo

Coefficienti:

$\text{log-odds} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

- Cambiamenti nelle log-odds associati a incrementi unitari nelle variabili.

Log-odds:

$$\text{log-odds} = \ln\left(\frac{P(y=1)}{1 - P(y=1)}\right)$$

- Permettono di modellare la probabilità su una scala lineare.

Odds

$$\text{odds} = e^{\text{log-odds}}$$

- Il rapporto tra la probabilità di successo e di insuccesso.

Perché il modello logistico produce log-odds?

- Nella regressione logistica, il nostro obiettivo è modellare la **probabilità** che un evento accada, $P(y = 1)$.
- Tuttavia, $P(y = 1)$ è limitato tra 0 e 1, quindi non può essere modellata direttamente come combinazione lineare delle variabili.
- Per superare questa limitazione:
  - Si calcolano le **odds**, cioè il rapporto tra la probabilità di successo e quella di insuccesso:
  $$\text{odds} = \frac{P(y = 1)}{1 - P(y = 1)}$$
  - Si applica il logaritmo naturale alle odds, ottenendo le **log-odds**, che hanno un intervallo illimitato:
  $$\text{log-odds} = \ln\left(\frac{P(y = 1)}{1 - P(y = 1)}\right)$$
- La regressione logistica modella le log-odds come combinazione lineare delle variabili esplicative:
$$\text{log-odds} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

Interpretazione dell'Intercetta nella Regressione Logistica

Definizione dell'Intercetta (β_0):

- L'intercetta rappresenta le **log-odds** dell'evento di successo ($y = 1$) quando tutte le variabili esplicative sono uguali a 0.
- Può essere trasformata in una probabilità:

$$P(y = 1) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

Interpretazione:

- $\beta_0 > 0$: L'evento è più probabile ($P(y = 1) > 0.5$) quando $x_1, x_2, \cdots = 0$.
- $\beta_0 = 0$: L'evento ha probabilità $P(y = 1) = 0.5$ quando $x_1, x_2, \cdots = 0$.
- $\beta_0 < 0$: L'evento è meno probabile ($P(y = 1) < 0.5$) quando $x_1, x_2, \cdots = 0$.

Considerazioni:

- L'intercetta ha senso solo se $x = 0$ è un valore interpretabile per tutte le variabili.
- Se $x = 0$ non ha significato pratico, l'intercetta è difficile da interpretare.

Interpretazione dei risultati

Supponiamo di avere i seguenti risultati di regressione logistica:

	Coefficiente	Errore Std.	Statistica Z	Valore-P
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

- **Intercept (-10.6513)**:
  - Rappresenta le log-odds quando tutte le variabili esplicative sono uguali a 0.
  - Essendo molto negativo, le probabilità di successo ($P(y = 1)$) sono molto basse quando balance = 0.
- **balance (0.0055)**:
  - Il coefficiente positivo indica che all'aumentare di balance, aumentano le log-odds, quindi la probabilità di successo.
  - Interpretazione delle odds: $e^{0.0055} \approx 1.0055$. Per ogni aumento unitario di balance, le odds aumentano dello 0.55%.
  - Il valore-p molto basso (< 0.0001) indica che l'effetto è statisticamente significativo.

Fare Previsioni per Balance = $1000

Log-odds:

$$\text{log-odds} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X = -10.6513 + 0.0055 \cdot 1000 = -5.1524$$

Odds:

$$\text{odds} = e^{\text{log-odds}} = e^{-5.1524} \approx 0.0058$$

Probabilità:

$$\hat{p}(X) = \frac{\text{odds}}{1 + \text{odds}} = \frac{0.0058}{1 + 0.0058} \approx 0.006$$

Interpretazione:

- **Log-odds**: Le log-odds di default sono −5.1524, indicando una probabilità estremamente bassa.
- **Odds**: Le odds sono 0.0058, ovvero circa 1 possibilità su 173 che il default avvenga.
- **Probabilità**: La probabilità stimata di default è 0.006, cioè lo 0.6%.

Fare Previsioni per Balance = \$2000

Log-odds:

$$\text{log-odds} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X = -10.6513 + 0.0055 \cdot 2000 = 0.3465$$

Odds:

$$\text{odds} = e^{\text{log-odds}} = e^{0.3465} \approx 1.414$$

Probabilità:

$$\hat{p}(X) = \frac{\text{odds}}{1 + \text{odds}} = \frac{1.414}{1 + 1.414} \approx 0.586$$

Interpretazione:

- **Log-odds**: Le log-odds di default sono 0.3465, indicando una probabilità maggiore del 50%.
- **Odds**: Le odds sono 1.414, ovvero il default è circa 1.41 volte più probabile rispetto al non default.
- **Probabilità**: La probabilità stimata di default è 0.586, cioè il 58.6%.

La devianza in un modello logistico

Formula della devianza:

$$D = -2 \cdot \left(\ell_{\text{modello}} - \ell_{\text{modello di saturazione}}\right)$$

Dove:

- ℓ_{modello}: La log-verosimiglianza del modello stimato.
- $\ell_{\text{modello di saturazione}}$: La log-verosimiglianza di un modello perfetto che si adatta esattamente ai dati (massima possibile).

Spiegazione:

- La devianza misura quanto il modello corrente si discosta da un modello perfetto.
- Valori più bassi di devianza indicano un miglior adattamento ai dati.
- Si confrontano due tipi di devianza:
  - **Devianza nulla**: Modello con solo l'intercetta (nessun predittore).
  - **Devianza residua**: Modello con i predittori inclusi.
- La differenza tra devianza nulla e devianza residua indica quanto i predittori migliorano il modello.

Akaike Information Criterion (AIC)

Formula dell'AIC:

$$\text{AIC} = -2 \cdot \ell_{\text{modello}} + 2 \cdot k$$

Dove:

- ℓ_{modello}: La log-verosimiglianza del modello stimato.
- k: Il numero di parametri stimati nel modello (inclusa l'intercetta).

Spiegazione:

- L'AIC valuta un modello bilanciando:
  - **Adattamento**: Quanto bene il modello si adatta ai dati (log-verosimiglianza).
  - **Semplicità**: Penalizza modelli troppo complessi (numero di parametri).
- Modelli con un AIC più basso sono preferibili.
- L'AIC è utile per confrontare modelli: un modello con AIC più basso è considerato migliore.

Nota: L'AIC non misura la qualità assoluta di un modello, ma è un criterio comparativo tra modelli con gli stessi dati.

Confusion Matrix e Accuracy

Definizione della Confusion Matrix:

	Predetto: No	Predetto: Yes
Effettivo: No	True Negative (TN)	False Positive (FP)
Effettivo: Yes	False Negative (FN)	True Positive (TP)

Terminologia chiave:

- **True Positive (TP):** Predetto "Yes" e l'evento è effettivamente "Yes".
- **False Positive (FP):** Predetto "Yes" ma l'evento è effettivamente "No".
- **True Negative (TN):** Predetto "No" e l'evento è effettivamente "No".
- **False Negative (FN):** Predetto "No" ma l'evento è effettivamente "Yes".

Accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Totale delle osservazioni}}$$

Confusion Matrix e Accuracy

Risultati ottenuti:

	Effettivo: No	**Effettivo: Yes**
Predetto: No	9625	233
Predetto: Yes	42	100

Calcolo dell'Accuracy:

$$\text{Accuracy} = \frac{9625 + 100}{9625 + 233 + 42 + 100} = 0.9725\ (97.25\%)$$

LAB Puntualità delle Consegne nell'E-Commerce

La puntualità delle consegne è un aspetto fondamentale per garantire la soddisfazione del cliente nel settore dell'e-commerce. Attraverso l'analisi dei dati delle spedizioni, è possibile identificare i fattori che influenzano maggiormente i ritardi, consentendo all'azienda di ottimizzare le operazioni logistiche e migliorare l'efficienza.

Obiettivo dell'analisi:

- Identificare i fattori chiave che portano a ritardi nelle consegne.
- Creare un modello predittivo per anticipare i ritardi.
- Fornire raccomandazioni operative per mitigare i ritardi e migliorare la soddisfazione del cliente.

Descrizione delle Variabili - Dataset Shipping

Il dataset contiene informazioni raccolte su spedizioni e clienti per analizzare i ritardi nelle consegne. Le principali variabili includono:

- **mode_of_shipment**: Modalità di spedizione (Nave, Aereo, Strada).
- **customer_rating**: Valutazione del cliente (1 = peggiore, 5 = migliore).
- **cost_of_the_product**: Costo del prodotto (USD).
- **prior_purchases**: Numero di acquisti precedenti del cliente.
- **product_importance**: Importanza del prodotto (bassa, media, alta).
- **gender**: Genere del cliente (Maschio, Femmina).
- **discount_offered**: Sconto offerto sul prodotto.
- **weight_in_gms**: Peso del prodotto in grammi.
- **reached_time**: Variabile target (1 = in ritardo, 0 = puntuale).

LAB Puntualità delle Consegne: Attività

1. Esplora il dataset e verifica la distribuzione delle variabili. Controlla eventuali valori mancanti o incoerenti.
2. Costruisci un modello di regressione logistica con `Reached.on.Time_Y.N` come variabile dipendente e le altre variabili come predittori.
3. Analizza i coefficienti del modello per identificare i fattori significativi e il loro impatto sulla probabilità di ritardo.
4. Sviluppa raccomandazioni operative basate sui risultati del modello per ridurre i ritardi nelle consegne.