Regressione Logistica
(Logistic regression)

Classificazione / Analisi Multivariata

- ✓ La regressione lineare ($Y = X\beta$) assume che l'output sia quantitativo (in generale numeri reali)
- ✓ In alcuni casi, l'output è qualitativo (categoriale: 0-1, yes-no)

Predire una risposta qualitativa (Y) per un'osservazione può essere definito come classificare quell'osservazione, poiché implica assegnare l'osservazione a una categoria, o classe.

Regressione Logistica

Analisi Discriminante Lineare (LDA)

Analisi Discriminante Quadratica (QDA)

Metodi più avanzati, trattati in ISL, includono modelli additivi generalizzati, alberi, foreste casuali (random forests), boosting e macchine a vettori di supporto (support vector machines).

ESEMPIO 1:

Determinare se un paziente è a rischio di diabete (sì/no) in base a vari fattori clinici.

Variabili:

Variabile dipendente (outcome - Y):
Diabete: 1 = Sì (il paziente ha il diabete), 0 = No (il paziente non ha il diabete).

Variabili indipendenti (fattori di rischio - X):
Età del paziente (in anni)
BMI (Indice di Massa Corporea)
Livelli di glucosio a digiuno (mg/dL)
Storia familiare di diabete (1 = sì, 0 = no)

Obiettivo:

Usare una regressione logistica per stimare la probabilità che un paziente abbia il diabete, in base ai fattori di rischio indicati.

ESEMPIO 2:

Determinare se un paziente ha certi sintomi perché è a rischio di stroke, di overdose di farmaco, di manifestare una crisi epilettica.

$$Y = \begin{cases} 1 & stroke \\ 2 & drug\ overdose \\ 3 & epileptic\ seizure \end{cases}$$

Questa codifica implica un ordinamento degli esiti, ponendo l'overdose da farmaci tra l'ictus e la crisi epilettica, e insistendo sul fatto che la differenza tra ictus e overdose da farmaci sia la stessa della differenza tra overdose da farmaci e crisi epilettica

Supponiamo di poter semplificare: Per una risposta qualitativa binaria (a due livelli), la situazione è più semplice. Ad esempio, potrebbero esserci solo due possibilità per la condizione medica del paziente: ictus e overdose da farmaci.

$$Y = \begin{cases} 1 & if\ stroke \\ 0 & if\ drug\ overdose \end{cases}$$

La regressione logistica modella la probabilità che Y appartenga a una particolare categoria.
Nel nostro esempio, la regressione logistica modella la probabilità di stroke in questo modo:

La probabilità di sperimentare uno stroke dato lo stato (X) della persona può essere scritta come:

Probabilità (*stroke* = *Yes*|*stato*) o *Probabilità* (*stroke* = *1*|*stato*)

Come possiamo esplicitare in formule matematiche la:

$$Probabilità(Y = 1|X)\ e\ X\ ?$$

La regression logistica usa la seguente funzione:

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

Questo è un modello non lineare. Non possiamo usare l'estimatore dei minimi quadrati lineari, ma utilizziamo un estimatore non lineare (ad esempio: l'estimatore di massima verosimiglianza)

Proprietà:

- La distribuzione logistica vincola le probabilità stimate a trovarsi tra 0 e 1 ([0 1]).

- La probabilità viene stimata da:

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

- Se $\beta_0 + \beta_1 X = 0$, allora p = .50
- Se $\beta_0 + \beta_1 X \to +\infty$, p = 1
- se $\beta_0 + \beta_1 X \to -\infty$, p = 0

Tuttavia, possiamo manipolare questa equazione come segue:

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

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

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

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

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

Dove la quantità alla sinistra è chiamata **odds (probabilità).**

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

- Gli odds possono avere valore tra $[0, \infty]$
- Valori degli odds vicini a 0 e ∞ indicano, rispettivamente, probabilità molto basse o molto alte

Possiamo ulteriormente trasformare l'equazione in questo modo:

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

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

Dove ora il termine a sinistra è chiamato log-odds o logit. Si noti che il modello di regressione logistica ora è lineare in X.

Coefficienti - significato

- Un'interpretazione del coefficiente logit che di solito è più intuitiva è il "rapporto di probabilità". Poiché

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

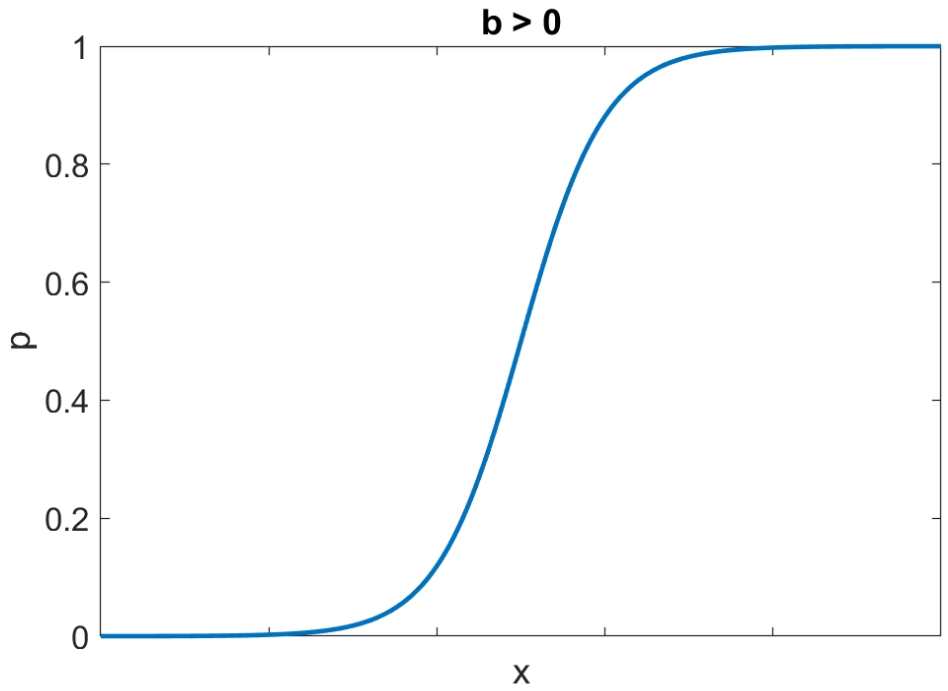
$e^{\beta 1}$ è l'effetto della variabile indipendente sul "rapporto di probabilità"

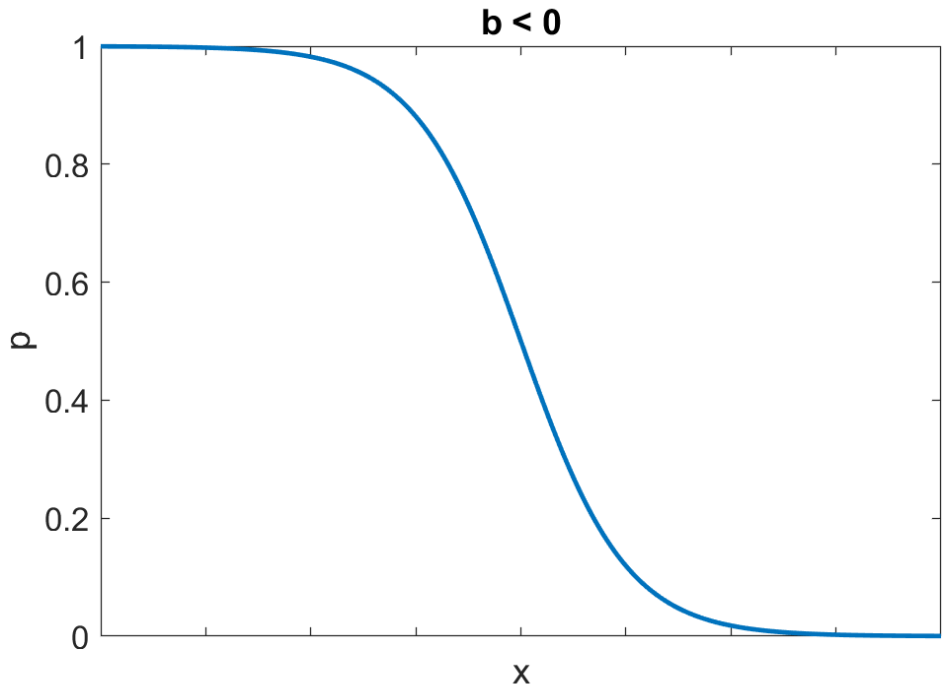
Confronto tra regressione lineare e logistica

- la probabilità non è mai inferiore, a zero.
- la probabilità non è mai superiore a uno.
- La funzione logistica produce sempre una curva a forma di S.

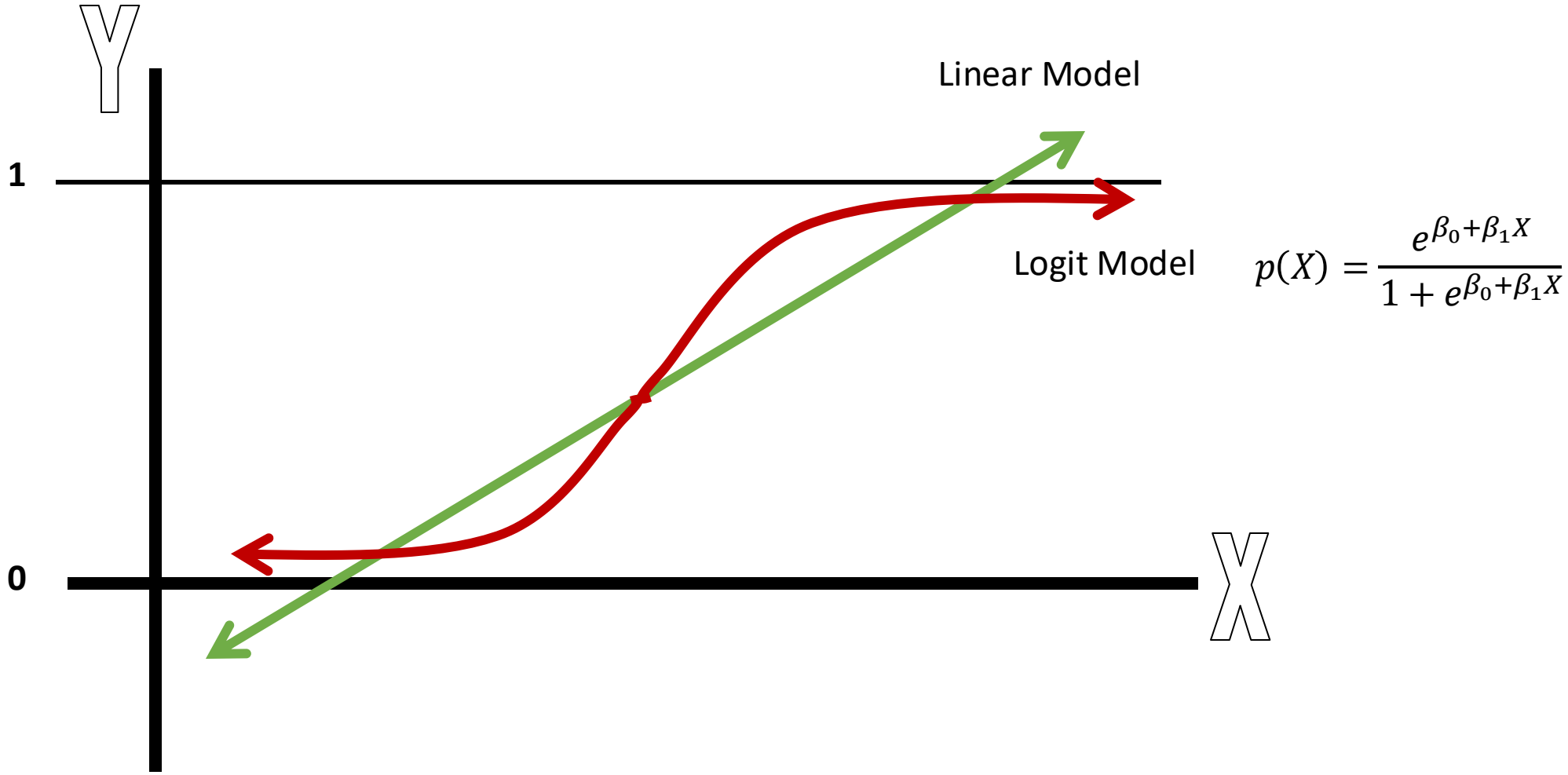
- Com’è fatta la quantità p predetta dalla regressione logistica?

$$p = \frac{e^{\beta_0 + \beta \cdot X}}{1 + e^{\beta_0 + \beta \cdot X}}$$





Comparing LP and Logit Models

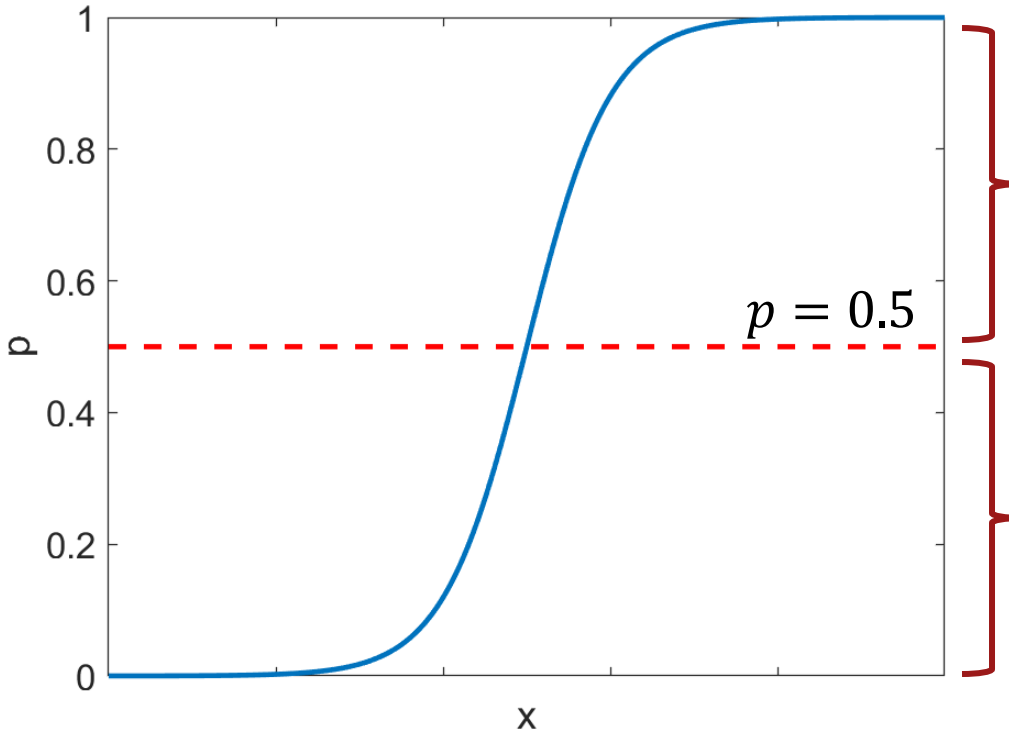


$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

SOGLIA DI CLASSIFICAZIONE

- Applichiamo una soglia th sul valore di p:
  - Se $p < th \rightarrow \hat{Y} = 0$
  - Se $p \geq th \rightarrow \hat{Y} = 1$

$$p = P(Y = 1|X, \beta)$$



$p \geq 0.5 \rightarrow \hat{Y} = 1$
Prediciamo la classe 1

$p < 0.5 \rightarrow \hat{Y} = 0$
Prediciamo la classe 0

Di fatto il problema che approcciamo con la regressione logistica è un problema di **classificazione binaria**: cerchiamo di costruire un modello matematico per predire il valore di una variabile binaria Y

Regressione logistica contro regressione lineare

Ricordiamoci che in un modello di regressione lineare, β_1 è il valore modifica il valore di Y associato con l'incremento di una unità di X.

In un modello di regressione logistica, l'incremento di una unità di X cambia gli **log odds** del valore β_1, o equivalentemente gli odds di $e^{\beta 1}$.

Regressione Logistica (generalizzazione)

Consideriamo ora il problema di predire una risposta binaria utilizzando più predittori.

Generalizziamo questo come segue :

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

dove $X = (X_1, \ldots, X_p)$ sono p predittori.
Possiamo anche usare la versione non lineare:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

RISULTATI DEL PROCESSO DI STIMA

- **Stime dei coefficienti**: $\hat{\beta}_j, j = 0, \dots, m$
- **Standard error** sulle stime dei coefficienti: $SE_j, j = 0, \dots, m$

- **Intervallo di confidenza** al 95% sulle stime dei parametri:

$$\hat{\beta}_j \pm 1.96 \cdot SE_j, \qquad j = 0, \dots, m$$

- **Odds ratio** delle stime dei coefficienti:

$$OR_j = e^{\hat{\beta}_j}$$

- Intervallo di confidenza al 95% sugli odds ratio:

$$e^{\hat{\beta}_j \pm 1.96 \cdot SE_j}$$

ESEMPIO (2)

- Vogliamo investigare se sussiste una relazione tra il successo/fallimento di un impianto dentale (Y) e due variabili esplicative: la lunghezza dell'impianto dentale (X_1) e l'età del paziente al momento dell'impianto (X_2).

- Disponiamo di una dataset contenente 200 osservazioni per le 3 variabili in gioco.
  - Per 50 osservazioni l'esito è fallimentare (Y=0)
  - Per 150 osservazioni l'esito è successo (Y=1)
- Applichiamo il modello di regressione logistica
- Equazione del modello:

$$p = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$$

- Risultato della stima di massima verosimiglianza:

Variabile	Stima del coefficiente $\hat{\beta}_j$	Standard error SE_j	Intervallo di confidenza al 95% $\hat{\beta}_j \pm 1.96 \cdot SE_j$	Odds ratio $e^{\hat{\beta}_j}$	Intervallo di confidenza al 95% $e^{\hat{\beta}_j \pm 1.96 \cdot SE_j}$
Intercetta	-15.437	2.9884	[-21.29 -9.58]	1.9756×10^{-7}	[5.6×10^{-10} 6.9×10^{-5}]
Lunghezza impianto [mm]	1.5842	0.2559	[1.08 2.09]	4.8752	[2.95 8.05]
Età [anni]	-0.0231	0.0371	[-0.096 0.048]	0.9771	[0.91 1.05]

VALUTAZIONE DELLA BONTA' DEL MODELLO

- Valutazione dell'errore di classificazione
- Deviance e likelihood ratio test

ERRORE DI CLASSIFICAZIONE

Data una soglia th per la classificazione, dobbiamo confrontare:

- $\hat{y}_i$: valori dell'outcome predetti dal modello (0 o 1)
  - $\hat{\mathbf{y}}_\mathbf{i} = \mathbf{1}$ → valore predetto **positivo**
  - $\hat{\mathbf{y}}_\mathbf{i} = \mathbf{0}$ → valore predetto **negativo**
- y_i: valori reali dell'outcome (0 o 1)
- **Matrice di confusione**:

	# valori positivi $\mathbf{y_i = 1}$	**# valori negativi** $\mathbf{y_i = 0}$
# valori predetti positivi $\hat{\mathbf{y}}_\mathbf{i} = \mathbf{1}$	# VERI POSITIVI o TRUE POSITIVES (TP)	# FALSI POSITIVI o FALSE POSITIVES (FP)
# valori predetti negativi $\hat{\mathbf{y}}_\mathbf{i} = \mathbf{0}$	# FALSI NEGATIVI o FALSE NEGATIVES (FN)	# VERI NEGATIVI o TRUE NEGATIVES (TN)

METRICHE DI CLASSIFICAZIONE

- **Accuratezza**: frazione di predizioni corrette sul totale dei valori predetti

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

- **Sensibilità** (o Recall): frazione di valori positivi correttamente predetti

$$Sensitivity = \frac{TP}{TP + FN}$$

- **Specificità**: frazione di valori negativi correttamente predetti

$$Specificity = \frac{TN}{TN + FP}$$

- **Precisione** (o positive predictive value): frazione di predizioni positive corrette

$$Precision = \frac{TP}{TP + FP}$$

	$\mathbf{y_i = 1}$	$\mathbf{y_i = 0}$
$\mathbf{\hat{y}_i = 1}$	TP	FP
$\mathbf{\hat{y}_i = 0}$	FN	TN

Tutte queste metriche presentano valori compresi tra 0 e 1 e il loro valore è tanto migliore quanto più è vicino a 1.

METRICHE DI CLASSIFICAZIONE PER IL CLASSIFICATORE RANDOM

Un modello che assegna i valori predetti di Y a caso avrà le seguenti metriche di classificazione:

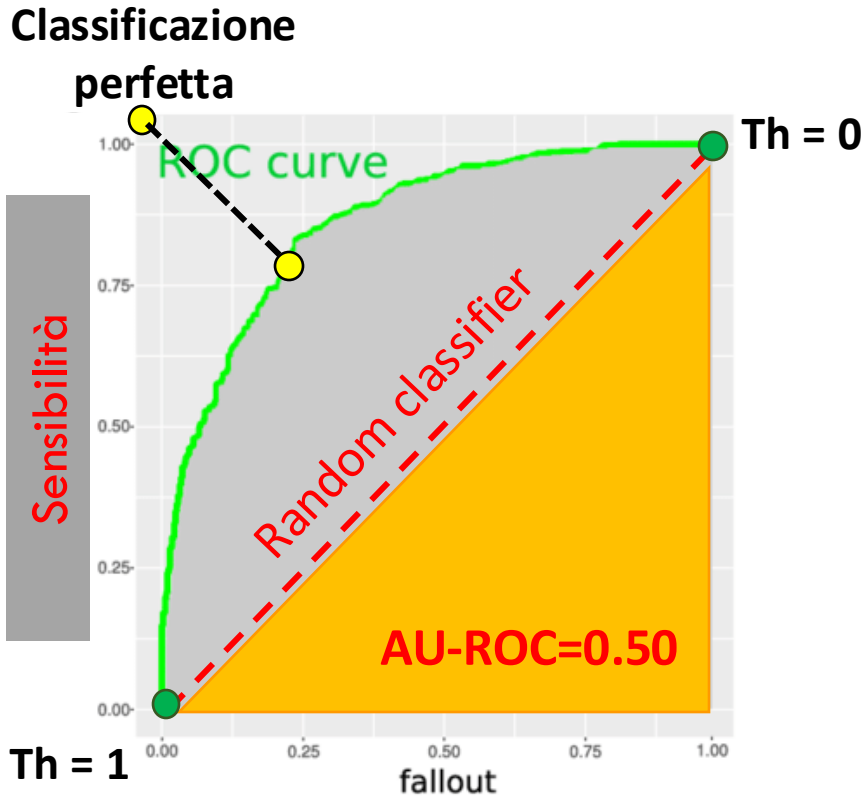
- Accuratezza = 0.5
- Sensibilità = 0.5
- Specificità = 0.5
- Precisione = (# osservazioni con Y =1)/# totale di osservazioni

SCELTA DELLA SOGLIA DI CLASSIFICAZIONE

- Come scegliamo la soglia di classificazione th?
- La scelta più intuitiva è 0.5, ma questa potrebbe non essere la scelta ottimale, specie se la prevalenza delle classi è sbilanciata (tante più osservazioni con Y=0 rispetto a Y=1 o viceversa).
- Per una scelta ottimale si possono testare diversi valori di th e scegliere il migliore sulla base delle metriche di classificazione.
- All'aumentare di th, tipicamente sensibilità diminuisce e specificità aumenta
  - Se $th = 0$ → tutti i valori predetti sono positivi → sensitività = 1, specificità = 0
  - Se $th = 1$ → tutti i valori predetti sono negativi → sensitività = 0, specificità = 1
- Tipicamente si cerca il valore di th per cui si ha un buon compromesso tra sensibilità e specificità.

LA CURVA ROC

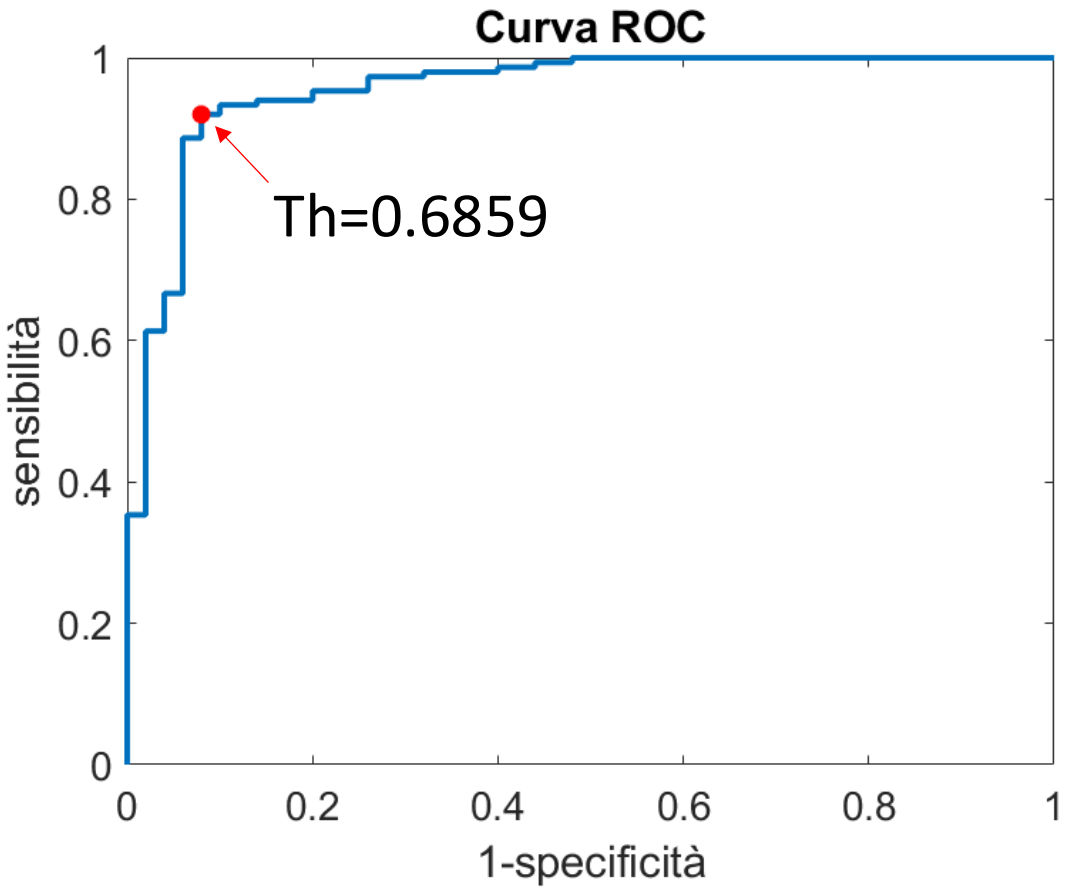
- **Curva ROC** (*Receiver Operating Characteristic* curve): grafico che mostra sensibilità vs. 1-specificità al variare di th.



- Possibile soglia ottima: il valore di th per cui la distanza della curva ROC dall'angolo in alto a sinistra è minima.
- L'**area sotto la curva ROC (AU-ROC)** è una metrica di classificazione indipendente da th
  - Valore compreso tra 0 e 1, tanto più è vicino a 1, tanto più il modello tende ad assegnare valori di probabilità elevati ad osservazioni per cui Y=1.
  - Il classificatore che assegna i valori predetti in modo random presenta AU-ROC=0.5.

ESEMPIO

- Valutiamo l'errore di classificazione per il modello dell'esempio precedente.



AU-ROC=0.9557

ESEMPIO

- Metriche di classificazione in corrispondenza della soglia ottima:

Matrice di confusione

	$y_i = 1$	$y_i = 0$
$\hat{y}_i = 1$	137	4
$\hat{y}_i = 0$	13	46

- Accuratezza: (137+46)/200 = 0.915
- Sensibilità: 137/(137+13) = 0.9133
- Specificità: 46/(46+4) = 0.92
- Precisione = 137/(137+4)=0.9716