Classificazione attraverso modelli di regressione e le misure di performance di un classificatore

Classificazione logistica e come valutare un classificatore

N. Torelli

2024

University of Trieste

Introduzione

Classificazione

- Nei problemi di classificazione supervisionata la variabile risposta Y è una variabile categoriale che può assumere valori in un insieme $\mathbb{C}$.
- Ad esempio:
  - **colore degli occhi** per cui $\mathbb{C}$=$\{$marroni,blu,verdi$\}$
  - **email** con $\mathbb{C}$=$\{$spam, mail regolare$\}$
  - **churn** con $\mathbb{C}$= $\{$resta cliente della compagnia, cambia compagnia$\}$.
- L'insieme $\mathbb{C}$ può essere costituito da solo due elementi, classificazione binaria, o da più di due elementi (classificazione multiclasse)

Costruire una regola di classificazione

Dato un vettore X di caratteristiche (covariate, inputs, features) e una variabile qualitativa Y che assume valori nell'insieme $\mathbb{C}$, il problema di classificazione è di costruire una regola $C(X)$ che in corrispondenza di un valore del vettore X fornisca una previsione $\hat{Y} \in \mathbb{C}$.

Si possono distinguere tre strategie per pervenire a una regola di classificazione:

1. stimare direttamente la $Pr(Y|X)$ e poi basarsi sulla semplice regola di classificare l'unità nella classe per cui la probabilità stimata è massima.
2. stimare la densità di X condizionatamente all'appartenenza dell'elemento a ciascuna categoria $P(X|Y)$. Usando la formula di Bayes si potrà poi ottenere una stima di $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

e poi usare la medesima regola di prevedere la categoria cui viene attribuita la massima probabilità. 3. infine, vi sono metodi che si propongono di ottenere direttamente una partizione dello spazio entro cui varia X e di attribuire direttamente a ciascun elemento della partizione un valore di $Y \in \mathbb{C}$.

Diversi metodi per la classificazione

- Fra i metodi che si propongono di costruire una regola di classificazione stimando direttamente $P(Y|X)$ vi sono:
  - regressione logistica (sia parametrica che semiparametrica)
  - alberi di classificazione (decisione)
  - reti neurali
- I metodi che invece stimano $P(X|Y)$ includono:
  - Analisi discriminante lineare e quadratica
  - modelli mistura
- I metodi che pervengono a una diretta attribuzione di un individuo con caratteristiche X sono:
  - Support vector machines
  - i metodi di insieme

Classificazione attraverso modelli di regressione

Regressione lineare: si può usare?

- In un problema di classificazione a due classi possiamo definire la variabile risposta come segue:

$$Y = \begin{cases} 0 \\ 1 \end{cases}$$

- Potremmo proporre per la classificazione semplicemente un modello di regressione lineare di Y su X e poi ottenuta la previsione $\hat{Y}$ classificare come 1 se questo è superiore a 0.5.
- In effetti poiche la variabile Y ha nella popolazione una distribuzione bernoulliana la funzione di regressione lineare $E(Y|X = x) = Pr(Y = 1|X = x)$ potrebbe anche funzionare. Essa ha tuttavia il difetto di produrre valori di $E(Y|X = x)$ maggiori di 1 o negativi.
- Sappiamo che in questo caso possiamo usare la regressione per variabili risposta binarie

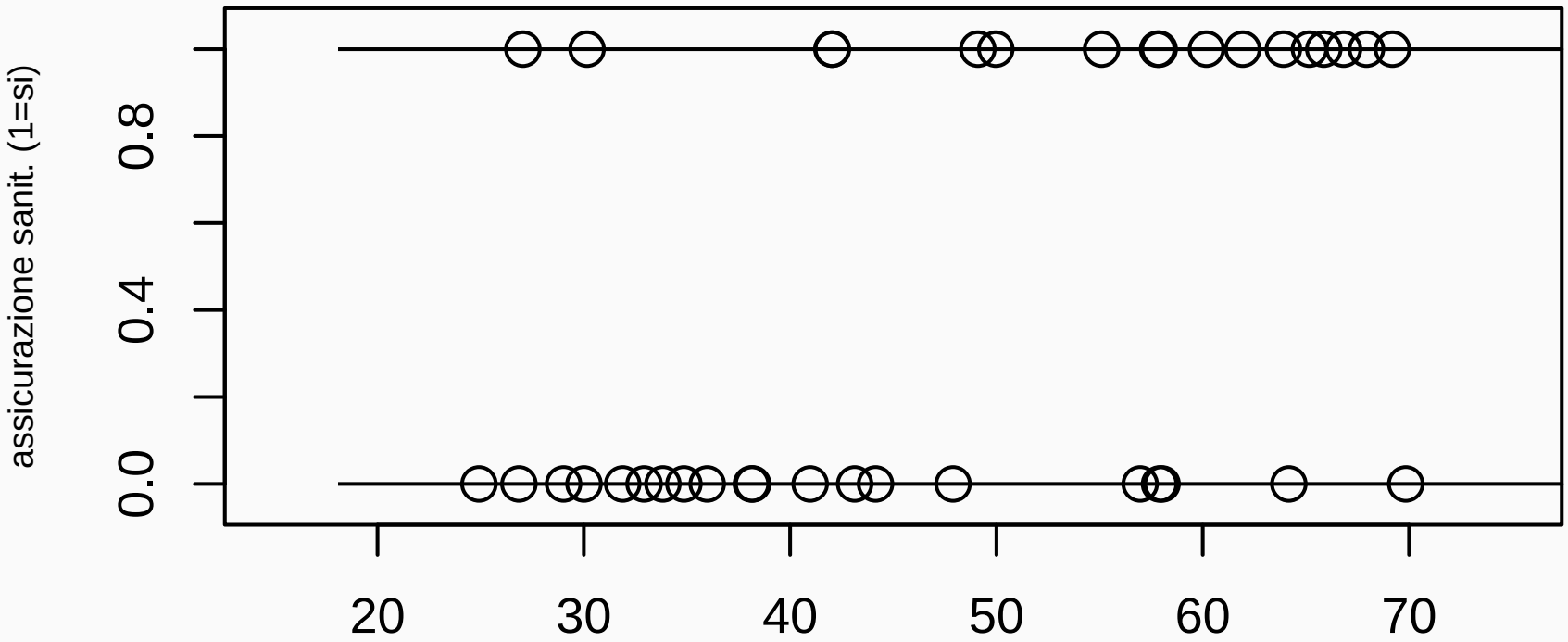
Modelli per variabile dipendente binaria

- Il set informativo di cui si suppone di disporre per l'unità i-esima è costituito quindi da
  - una variabile risposta y_i
  - un insieme di $p-1$ covariate $x_i = (x_{i1}, x_{i2}, \ldots, x_{i(p-1)})$
- Come per il caso della risposta quantitativa siamo interessati a costruire un modello (statistico) che ci permetta di (prevedere) la media della variabile y_i utilizzando le informazioni sulle covariate x_i. Quindi:
  - $y_i \sim Be(\pi_i)$ per cui $E(y_i) = \pi_i$
  - si definisce il **predittore lineare**

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

  - si ipotizza che $E(y_i) = \pi_i = r(\eta_i) = r(\boldsymbol{x}_i^T \boldsymbol{\beta})$
- La funzione $r()$ è detta **funzione risposta** e va scelta in modo opportuno fra quelle continue e invertibili.

Un banale esempio: un'assicurazione sanitaria



Si assuma di avere osservato 37 unità e per ciascuna di esse conosciamo l'età e se hanno o meno sottoscritto una polizza sanitaria. I grafico sembra suggerire che è più plausibile che nel campione osservato abbiano sottoscritto una polizza sanitaria i soggetti più anziani.

1. $y_i \sim \text{Be}(r(\eta_i))$ ove $\eta_i = \boldsymbol{x}_i^T \beta$.
2. una funzione ragionevole deve essere tale che $r(\cdot) \to [0, 1]$.

Possibili scelte per $r()$ e $g()$

È sempre possibile esprimere la relazione sopra invertendo la funzione $r()$. Si definisce quindi la **funzione legame** $g(\cdot) = r^{-1}(\cdot)$ e si ha equivalentemente

$$g(E(y_i)) = g(\pi_i) = \eta_i = \boldsymbol{x}_i^T \beta$$

Possiamo ottenere modelli alternativi scegliendo fra funzioni $g(\cdot))$ (oppure $r(\cdot)$) alternative. Le due opzioni più note sono quelle che conducono ai modelli **logit** e **probit**.

- *Modello logit*: si usa la funzione risposta

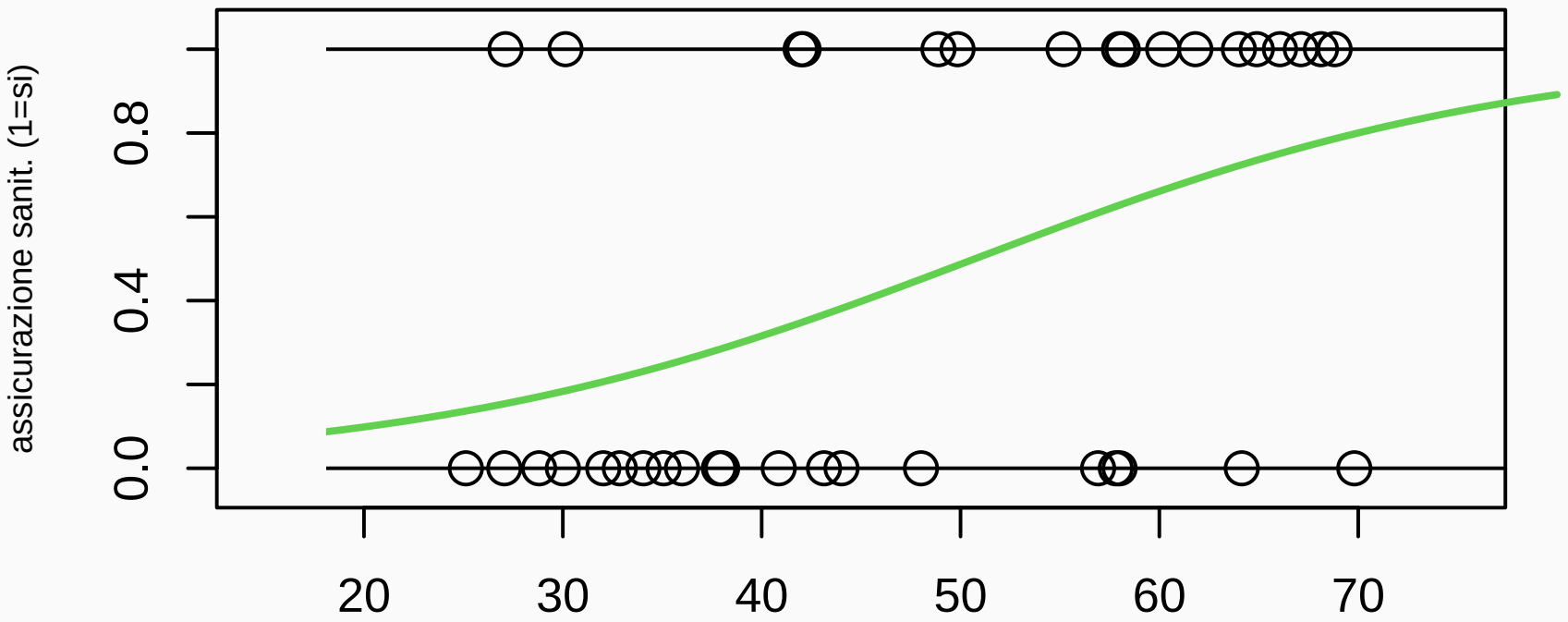
  $$\pi_i = r(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{e^{\boldsymbol{x}_i^T \beta}}{1 + e^{\boldsymbol{x}_i^T \beta}} \text{ oppure } g(\pi_i) = log(\frac{\pi_i}{1 - \pi_i}) = \eta_i = \boldsymbol{x}_i^T \beta$$

- *Modello probit*: si usa la funzione risposta

  $$\pi_i = \Phi(\eta_i) = \Phi(\boldsymbol{x}_i^T \beta)$$

  ove $\Phi(\cdot)$ è la funzione di ripartizione della Gaussiana standard

La regressione logistica per i dati dell'assicurazione sanitaria



1. $y_i \sim \text{Be}(r(\eta_i))$ ove $\eta_i = \boldsymbol{x}_i^T \beta$.
2. $\pi_i = r(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\boldsymbol{x}_i^T \beta}}{1+e^{\boldsymbol{x}_i^T \beta}}$

La regressione logistica è quella più comune

L'utilizzo della funzione legame logit dà origine alla regressione logistica che è di gran lunga la più popolare e la più usata. I motivi del suo successo sono:

- è computazionalmente più agevole e gode di alcune interessanti proprietà teoriche
- è più facile interpretare i risultati (effetto lineare delle covariate sul logaritmo degli odds-ratio, effetto moltiplicativo sugli odds-ratio)
- fornisce stime adeguate dei parametri anche in presenza di schemi di campionamento sbilanciati

Campionamento sbilanciato per le due classi

Il tema è stato sollevato dapprima nell'ambito bio statistico ma si applica in realtà a tutti i casi in cui il campione da analizzare vede sovrarappresentata una delle due modalità della variabile dipendente rispetto a quanto accade nella popolazione (nel caso ad esempio che una delle due classi sia più rara).

- Si immagini che nella popolazione la classe per cui $y = 1$ per un individuo con caratteristiche x si presenti con una frequenza pari a $p(x)$ molto bassa e quindi tali casi sono rari.
- Si immagini che sia π_1 la probabilità di includere un individuo nel campione condizionatamente al fatto che sia di tipo 1 e che sia π_0 la probabilità di includere un individuo nel campione condizionata al fatto che sia di tipo 0 .
- se si fa un campione casuale abbiamo che $\pi_1 = \pi_0$. Tuttavia se decidiamo di sovrarappresentare nel campione individui rari (di tipo 1) allora sarà $\pi_1 >> \pi_0$

Sovracampionamento della classe rara e legame logit

Definiamo la probabilità $p^*(x)$ come la probabilità che un individuo con caratteristiche x sia di tipo 1 condizionatamente alla sua inclusione nel campione e vediamo in che relazione è con la probabilità $p(x)$ che sia di tipo 1 (ovvero non condizionata alla scelta del campione e quindi relativa a un campione casuale in cui però gli 1 saranno rari).

Possiamo applicare la formula di Bayes:

$$p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0 (1 - p(x))}$$

alcuni semplici passaggi mostrano che per un generico individuo con caratteristiche x

$$log\left(\frac{p^*(x)}{1 - p^*(x)}\right) = log\frac{\pi_1}{\pi_0} + log\left(\frac{p(x)}{1 - p(x)}\right) = log\frac{\pi_1}{\pi_0} + x^T\beta$$

Quindi i parametri β associati alle variabili x non cambiano (cambia solo l'intercetta che sarà pari a $log\frac{\pi_1}{\pi_0} + \beta_0$). Occorre però tenere conto del rapporto $log\frac{\pi_1}{\pi_0}$ quando si tratta di stimare la $p(x)$ a scopo previsivo.

Inferenza

Stime dei parametri del modello

- Il modello di regressione logistica è un particolare modello lineare generalizzato (GLM): con risposta binomiale e legame (e corrispondete funzione risposta) logit.
- Il metodo della massima verosimiglianza è adeguato in questo e se si assume che il campione casuale (da variabili iid) di n osservazioni.
- Nel caso del modello Bernoulliano, la log verosimiglianza è $log(L(\beta)) = \ell(\beta)$

$$\ell(\beta) = \sum_{i=1}^{n} [y_i log(\pi_i) - y_i log(1-\pi_i) + log(1-\pi_i)] = \sum_{i=1}^{n} \left[y_i \frac{\pi_i}{1-\pi_i} + log(1-\pi \right.$$

- la funzione score risulta pari a

$$s(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \pi_i)$$

da cui le equazioni di verosimiglianza $s(\beta) = 0$. La cui soluzione va cercata numericamente (in genere l'algoritmo iterativo usato è quello tipico dei GLM ovvero IWLS)

Inferenza sui parametri

1. Per l'inferenza sui singoli parametri è utile ricordare le proprietà asintotiche delle stime di ML:

per n elevato si ha $\hat{\beta} \dot{\sim} \mathcal{N}(\beta, I(\beta)^{-1})$ ove $I(\beta)$ è la matrice di informazione attesa che in questo caso è pari a

$$I(\beta) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \pi_i (1 - \pi_i)$$

ove $\pi_i = r(\boldsymbol{x}_i^T \beta)$.

2. Tale matrice dipende dai parametri incogniti in β ma una sua stima consistente si ottiene sostituendo a β la sua stima $\hat{\beta}$.

3. L'elemento sulla diagonale $I(\beta)_{jj}^{-1}$ è quindi una stima della varianza di β_j.

4. Pertanto il rapporto $\frac{\beta_j}{\sqrt{I(\hat{\beta})_{jj}^{-1}}}$ relativo all'ipotesi nulla $H_0 : \beta_j = 0$ è asintoticamente distribuito come una Gaussiana standard se vera H_0.

Classificazione logistica

Classificazione a partire dal modello di regressione logistica

- Ottenute le stime di $\hat{\beta}$ è possibile ottenere delle stime della probabilità che un'unità i con vettore di covariate $\boldsymbol{x}_i$ sia classificata nella classe 1 semplicemente come

$$\hat{P}[(Y_i = 1|\boldsymbol{x}_i)] = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}} = \frac{e^{\boldsymbol{x}_i^T \hat{\beta}}}{1 + e^{\boldsymbol{x}_i^T \hat{\beta}}}$$

- A questo punto la classificazione può semplicemente ottenersi ponendo

$\hat{Y}_i = 1$ se $\hat{P}[(Y_i = 1|\boldsymbol{x}_i)] > k$

- A quanto va fissato k?

Soglia per la decisione

- Seguendo quanto già visto nell'introduzione si dovrebbe fissare $k = 0.5$

e in effetti è questa la scelta più comune

- Tuttavia se siamo a conoscenza che nella popolazione la proporzione di valori $Y = 1$ e pari a p si potrebbe argomentare dicendo che se non avessi alcuna informazione su x allora la mia regola di classificazione potrebbe essere posta a $k = p$
- In realtà vi sono molti motivi per variare la soglia: ad esempio
  - i due errori di classificazione (FP e FN) hanno costi diversi
  - si è sovracampionata una delle due classi

Classificazione multiclasse: logit multinomiale

- Finora abbiamo discusso la regressione logistica con due classi. Essa può facilmente essere generalizzata a più di due classi. Una versione ha la forma simmetrica

$$\hat{P}[(Y_i = k|\boldsymbol{x}_i)] = \frac{e^{\boldsymbol{x}_i^T \hat{\beta}_k}}{1 + e^{\boldsymbol{x}_i^T \hat{\beta}_k}}$$

- Una funzione logit-lineare per ciascuna classe (si noti che basta definirne k-1 di tali funzioni) e se K$=2$ si ottiene la regressione logistica
- La regressione logit-multinomiale è quindi una generalizzazione della regressione logistica.
- Si noti che si tratterrebbe di una versione multivariata del modello di regerssione logistica
- Non si può usare per essa direttamente la funzione `glm()` di R

Costruire un algoritmo di classificazione

- Il passo successivo è quello di costruire un algoritmo di classificazione a partire da un modello di regressione logistica.
- In questo caso ancora una volta giudicheremo sulla base delle performance del modello in un insieme di dati "nuovi"
- Si procede quindi a suddividere l'insieme di dati in training e test set
- Si stimano i parametri sul training set
- Si ottengono le previsioni sul test set

Misurare la performance per un classificatore a due vie

La Matrice di confusione

- La qualità di un algoritmo di apprendimento viene valutata osservando le sue prestazioni sul test set.
- Le misure di performance più semplici si basano sul confronto tra le previsioni del classificatore e i valori realmente osservati (matrice di confusione).
- Nella tabella le due modalità della variabile risposta sono indicate come "Positivi" e "Negativi"

		previsti		
		positivi	negativi	totale
	positivi	TP	FN	POS
Osservati	negativi	FP	TN	NEG
	totale	PrevPOS	PrevNEG	N

- Sulla diagonale principale si trovano TN (i casi classificati correttamente come negativi) e TP (osservati positivi)
- FP sono i casi detti falsi positivi mentre FN conta i falsi negativi

Le misure di performance: l'accuratezza

- La misura più semplice, la precisione (ACC) è definita come

$$ACC = \frac{TP + TN}{N}$$

- si noti che l'ACC **non** può essere utilizzato come misura della performance in un insieme di dati sbilanciato (con una delle classi con pochissimi casi): un classificatore banale che classifichi sempre i nuovi casi nella classe maggioritaria avrà un'accuratezza pari alla proporzione della classe maggioritaria nel campione
- Se abbiamo lo 0.1% del campione di casi con un tipo di tumore raro, una strategia (banale) che prevede sempre che non sei malato otterrà comunque ACC pari al 99.9%

Le misure di performance: il richiamo (recall) e la precisione (precision)

- True positive rate (recall or sensitività) $= \frac{TP}{POS}$
- True negative rate (specificità) $= \frac{TN}{NEG}$
- Positive predictive value (precisione) $= \frac{TP}{predPOS}$
- $F1 = (1 + \beta^2)\frac{\text{precisione}\cdot\text{recall}}{(\beta^2\text{precisione})+\text{recall}}$

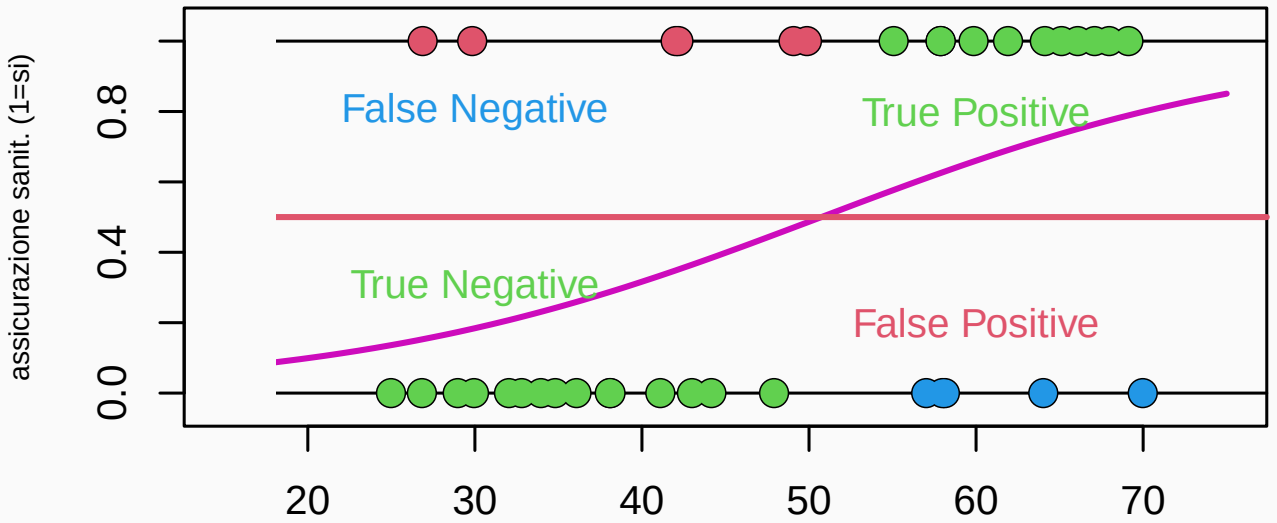
β è spesso posto pari a 1 (ovvero precisione e recall hanno lo stesso peso)

- La classificazione si ottiene fissando una soglia per i metodi che stimano $P(Y = 1|X)$ (regressione logistica, alberi) e, come già visto, questa soglia può essere arbitraria e a volte andrebbe modificata in modo appropriato.

Ancora sull'assicurazione sanitaria. Soglia e errori di classificazione

Table 1: Confusion Matrix

	Predicted Negative	Predicted Positive
Negative	15	5
Positive	6	11

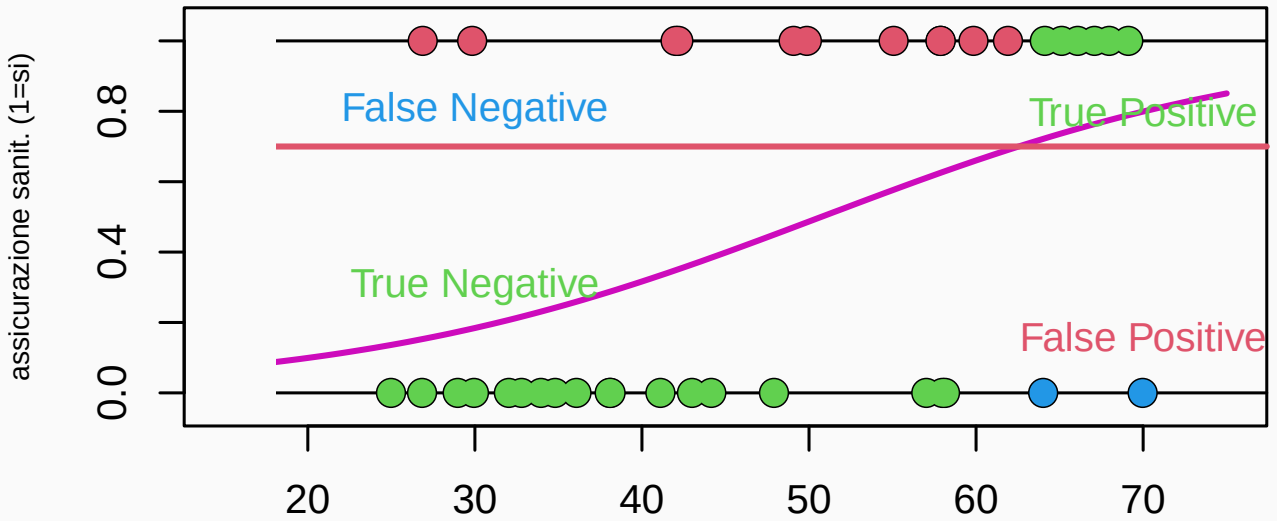


Il modello presentato in questo esempio è probabilmente leggermente sovradattato. Si ricorda che la matrice di confusione deve essere calcolata su un nuovo insieme di dati (test set)

Ancora l'esempio sull'assicurazione sanitaria. L'effetto di un cambiamento di soglia (k=0.7)

Table 2: Confusion Matrix

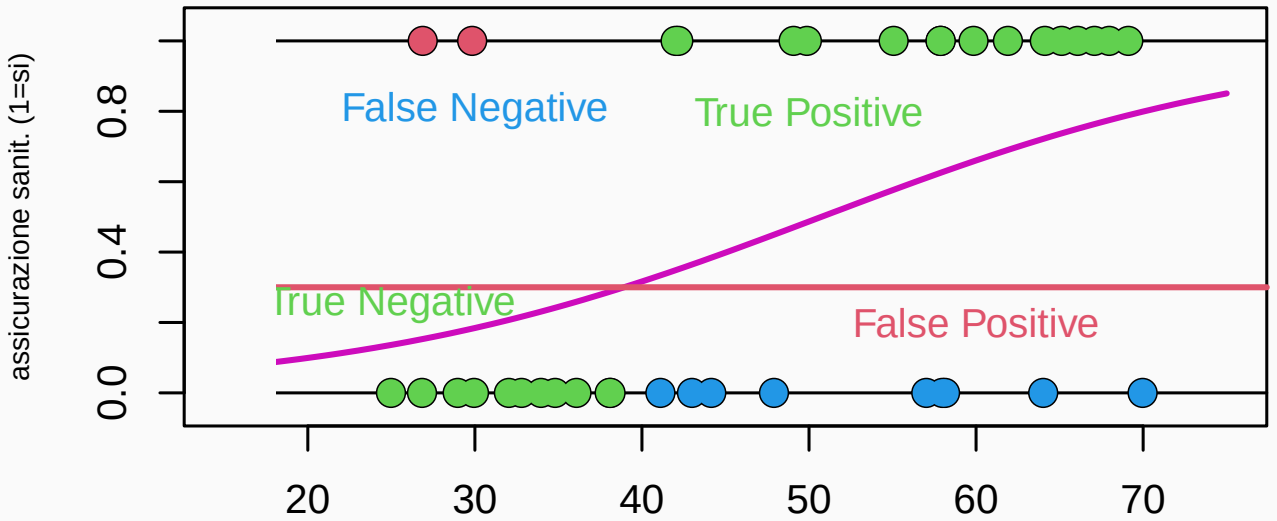
	Predicted Negative	Predicted Positive
Negative	18	2
Positive	11	6



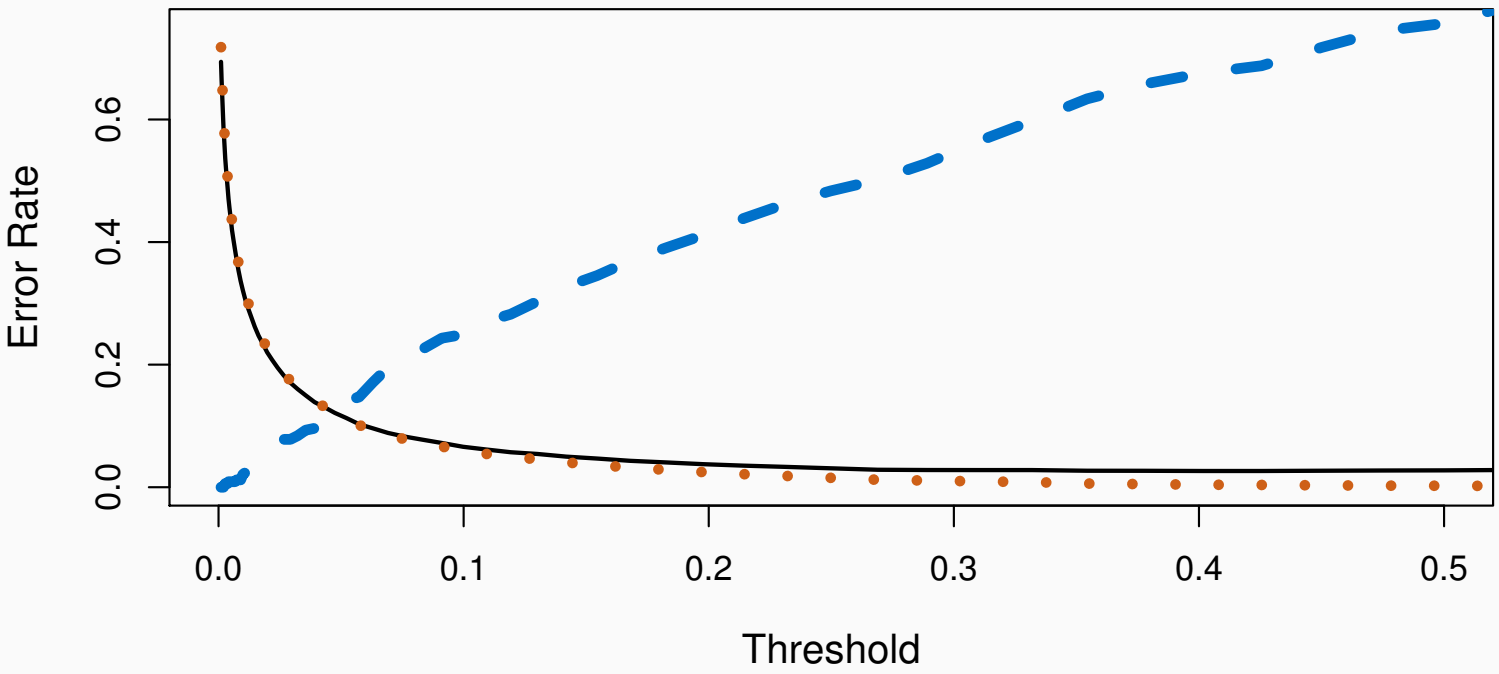
Ancora l'esempio sull'assicurazione sanitaria. L'effetto di un cambiamento di soglia (k=0.3)

Table 3: Confusion Matrix

	Predicted Negative	Predicted Positive
Negative	11	9
Positive	2	15



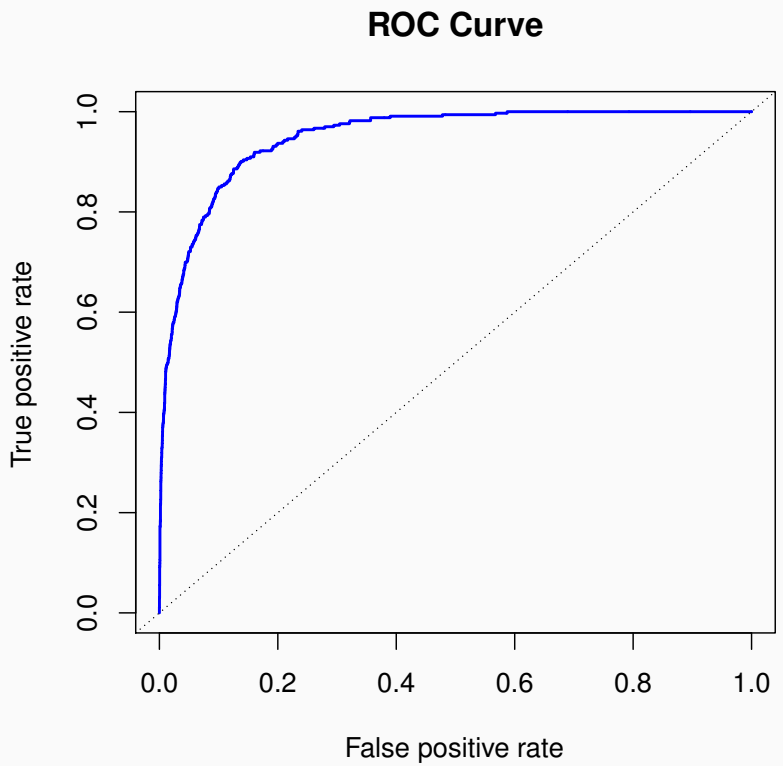
Errori e soglia per la classificazione



- Al variare della soglia sono riportate le proporzioni di falsi positivi (in blu) e di falsi negativi (in rosso)

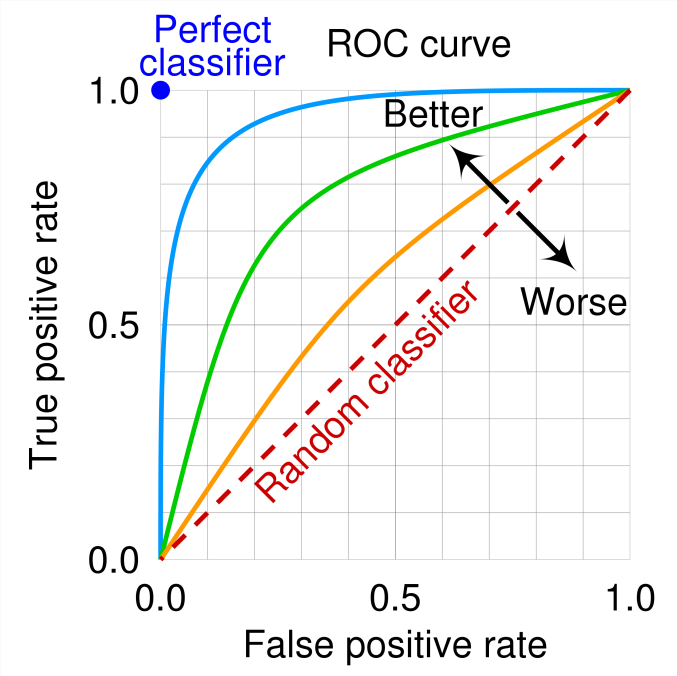
La curva ROC

- La curva ROC (Receiver Operating Characteristics) misura l'accuratezza di una previsione di classificazione e può essere valutata quando la previsione si presenta sotto forma di un punteggio numerico (nel caso di specie una probabilità).
- La curva ROC si ottiene tracciando la sensibilità rispetto alla specificità al variare della soglia tra 0 e 1.



AUC (area sotto la curva ROC)

- L'AUC è l'area che è sotto la curva ROC: essa non dipende da una determinata soglia ed è spesso una misura più appropriata per confrontare le prestazioni (ad esempio con un set di dati ri-bilanciato).



- L'AUC misura l'area sotto questa curva e più grande è il suo valore migliore è la prestazione di un classificatore (per un classificatore perfetto l'AUC è 1 per uno pessimo è pari a 0,5).
- Si noti che la curva ROC, e di conseguenza l'AUC, può essere molto instabile quando il set di dati del test è piccolo.

Determinare una soglia ottimale: l'indice di Youden

- Una soglia ottimale è quella che massimizza la somma del tasso di veri positivi e di veri negativi (ovvero la somma di sensibilità e specificità).
- Quindi si calcola la somma di specificità e sensibilità per ogni possibile k e si definisce il seguente indice (di Youden):

$$J = max_k[\text{sensibilità}(k) + \text{specificità}(k) - 1]$$

- Tale indice rappresenta anche un modo di sintetizzare la curva ROC essendo la massima distanza verticale fra la curva e la diagonale (che rappresenta il classificatore casuale) o anche la minima distanza fra il verice del quadrante in alto a sinistra.
- Altre proprietà di J sono le seguenti:
  - è compreso tra 0 e 1. Se è 0 indica una classificazione inefficace (equivalente a quella casuale) 1 è un classificatore perfetto.
  - Due classificatori che hanno lo stesso indice J thanno lo stesso tassi di errata classificazione.