UNIVERSITÀ DEGLI STUDI DI BERGAMO
Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione

Control Automation Lab

IDENTIFICAZIONE DEI MODELLI E ANALISI DEI DATI (IMAD)

Corso di Laurea Magistrale in INGEGNERIA INFORMATICA

Lezione 5: Regressione logistica

SPEAKER
Prof. Mirko Mazzoleni

PLACE
Università degli Studi di Bergamo

Syllabus

Parte I: sistemi statici

1. **Richiami di statistica**

2. **Teoria della stima**

   2.1 Proprietà degli stimatori

3. **Stima a minimi quadrati**

   3.1 Stima di modelli lineari

   3.2 Algoritmo del gradient descent

4. **Stima a massima verosimiglianza**

   4.1 Proprietà della stima

   4.2 Stima di modelli lineari

5. **Regressione logistica**

   5.1 Stima di un modello di regressione logistica

6. **Fondamenti di machine learning**

   6.1 Bias-Variance tradeoff

   6.2 Overfitting

   6.3 Regolarizzazione

   6.4 Validazione

7. **Cenni di stima Bayesiana**

   7.1 Probabilità congiunte, marginali e condizionate

   7.2 Connessione con Filtro di Kalman

IMAD

Parte I: sistemi statici

Stima parametrica $\hat{\theta}$

- **θ deterministico**
  - ***NO assunzioni su ddp dei dati***
    - ✓ Stima parametri popolazione
    - ✓ Stima modello lineare: minimi quadrati
  - ***SI assunzioni su ddp dei dati***
    - ✓ Stima massima verosimiglianza parametri popolazione
    - ✓ Stima modello lineare: massima verosimiglianza
    - ✓ Regressione logistica
- **θ variabile casuale**
  - ***SI assunzioni su ddp dei dati***
    - ✓ Stima Bayesiana

Machine learning

Parte II: sistemi dinamici

Stima parametrica $\hat{\theta}$

- **θ deterministico**
  - ***NO assunzioni su ddp dei dati***
    - ✓ Modelli lineari di pss
    - ✓ Predizione
    - ✓ Identificazione
    - ✓ Persistente eccitazione
    - ✓ Analisi asintotica metodi PEM
    - ✓ Analisi incertezza stima (numero dati finito)
    - ✓ Valutazione del modello

Outline

1. Il problema della classificazione
2. Perché non usare la regressione lineare?
3. Regressione logistica: formulazione del problema
4. Regressione logistica: funzione di costo
5. Riassunto
6. Esercizi con codice

Outline

1. **Il problema della classificazione**
2. Perché non usare la regressione lineare?
3. Regressione logistica: formulazione del problema
4. Regressione logistica: funzione di costo
5. Riassunto
6. Esercizi con codice

Il problema della classificazione

Il modello di **regressione lineare** discusso nella lezione precedente presuppone che la variabile di risposta sia **quantitativa (metrica)**

- in molte situazioni la variabile di risposta è invece **qualitativa (categorica)**

Le variabili qualitative assumono valori in un insieme non ordinato $\mathcal{C} = \{\text{"cat}_1\text{"}, \dots, \text{"cat}_C\text{"}\}$, come

- **`eye color`** $\in \{\text{"brown"}, \text{"blue"}, \text{"green"}\}$
- **`email`** $\in \{\text{"spam"}, \text{"not spam"}\}$

Dati metrici

- Descrivono una quantità
- È definito un ordine
- È definita una distanza

Dati categorici

- Descrivono «categorie di appartenenza»
- Non ha senso applicare un ordine
- Non ha senso calcolare le distanze

Il problema della classificazione

Il processo di stima di **output categorici**, utilizzando un insieme di regressori $\boldsymbol{\varphi}$, è chiamato **classificazione**

Spesso però siamo più interessati a **stimare le probabilità** che $\boldsymbol{\varphi}$ appartenga a ciascuna categoria in $\mathcal{C}$

Se si vuol ottenere una classificazione, la **categoria più probabile** viene scelta come **classe** (categoria) per l'osservazione $\boldsymbol{\varphi}$

Esempi di problemi di classificazione

- Una persona arriva al pronto soccorso con una **serie di sintomi** che potrebbero essere attribuiti a una delle **tre condizioni mediche**

  Quale delle tre condizioni affligge il paziente?

- Un sistema bancario online gestisce delle **transazioni**, memorizzando l'indirizzo IP dell'utente, la cronologia delle transazioni passate e così via

  La transazione è fraudolenta o no?

- Un biologo raccoglie dati su **sequenze di DNA** per un certo numero di pazienti **con e senza una determinata patologia**

  Quali mutazioni genetiche causano una patologia e quali no?

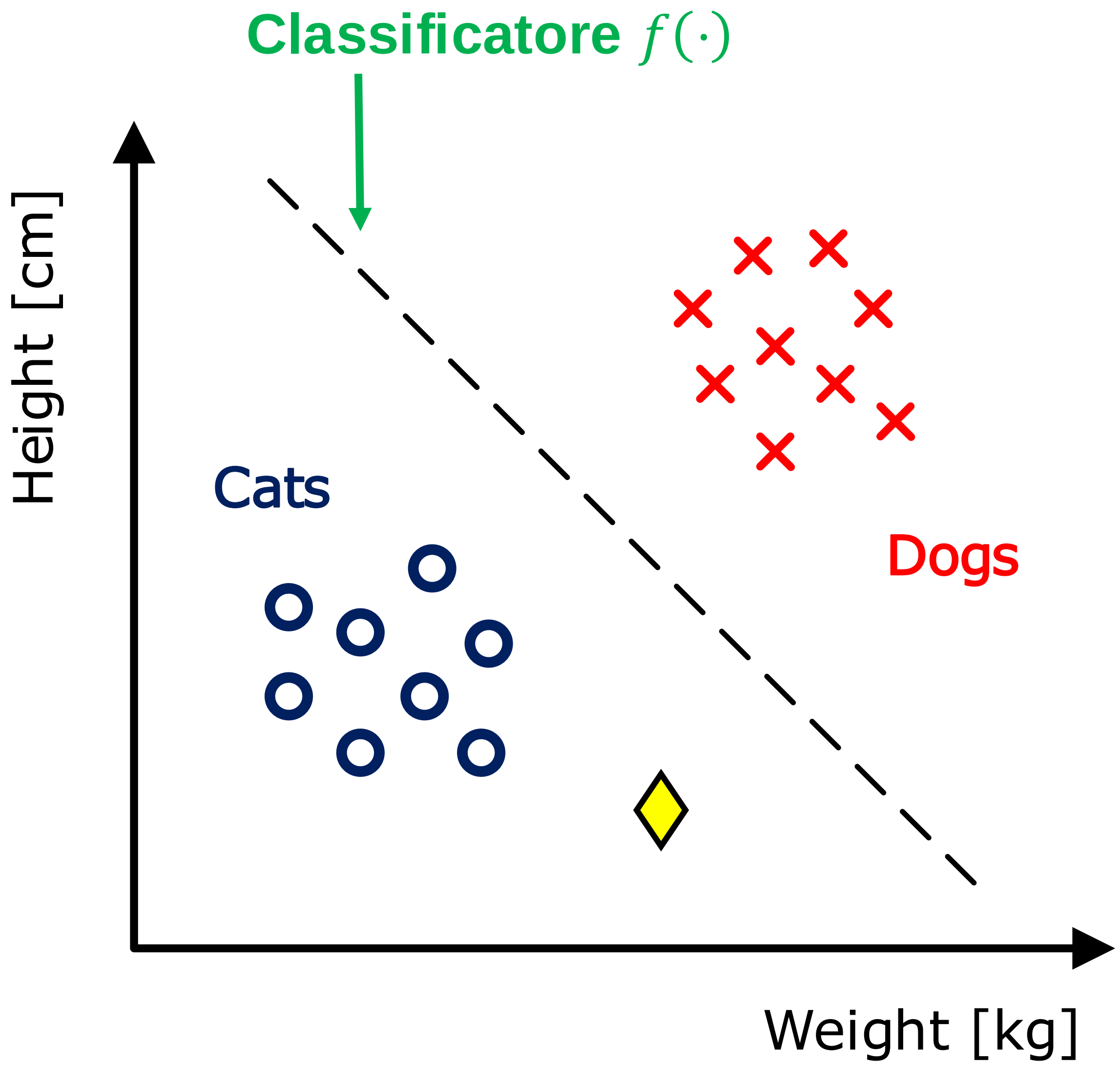
Esempio: cane vs. gatto

Supponiamo di misurare il **peso** e **l'altezza** di alcuni cani e gatti

Vogliamo imparare la funzione $f(\cdot)$ che ci dica se $\boldsymbol{\varphi} = [\varphi_1, \varphi_2]^\top$ è un cane o un gatto

- φ_1: peso
- φ_2: altezza

DOMANDA: Il punto  come è classificato dal modello? ________



QUIZ!

Consideriamo un'azienda che produce cancelli scorrevoli. I cancelli possono avere quattro pesi $\{300\ \mathrm{kg}, 400\ \mathrm{kg}, 500\ \mathrm{kg}, 600\ \mathrm{kg}\}$. Vogliamo rilevare il peso del cancello. Questo è un:

- ❑ Problema di regressione
- ❑ Problema di classificazione
- ❑ Sia un problema di classificazione che un problema di regressione

Outline

Perché non usare la regressione lineare?

Supponiamo di volere stimare la condizione di una paziente sulla base dei suoi sintomi. Ci sono tre possibilità: `stroke`, `drug overdose` and `epileptic seizure`

Potremmo considerare di codificare questi valori come una variabile **quantitativa**:

$$y = \begin{cases} 1 & \text{if} \quad \texttt{stroke} \\ 2 & \text{if} \quad \texttt{drug overdose} \\ 3 & \text{if} \quad \texttt{epileptic seizure} \end{cases}$$

Tuttavia, stiamo implicitamente dicendo che la «differenza» tra `drug overdose` e `stroke` è la medesima che tra `epileptic seizure` e `drug overdose`, il che **non ha molto senso**

Perché non usare la regressione lineare?

Potremmo anche cambiare la codifica in:

$$y = \begin{cases} 1 & \text{if} \quad \texttt{epileptic seizure} \\ 2 & \text{if} \quad \texttt{stroke} \\ 3 & \text{if} \quad \texttt{drug overdose} \end{cases}$$

Questo implicherebbe un **relazione totalmente differente** tra le tre condizioni

- ognuna di queste codifiche produrrebbe modelli lineari fondamentalmente diversi...
- …che alla fine porterebbe a diverse stime per nuove osservazioni

In generale, non esiste un modo naturale per convertire una variabile di risposta qualitativa con più di due livelli in una risposta quantitativa che sia adatta alla regressione lineare

Perché non usare la regressione lineare?

Con due livelli, la situazione è migliore. Ad esempio, forse ci sono solo due possibilità per le condizioni mediche del paziente: **stroke** e **drug overdose**

$$y = \begin{cases} 0 & \text{if} \quad \texttt{stroke} \\ 1 & \text{if} \quad \texttt{drug overdose} \end{cases}$$

Potremmo fittare una regressione lineare e classificare come **drug overdose** se $\hat{y} > 0.5$ e **stroke** altrimenti, interpretando $\hat{y}$ come una **probabilità di overdose**

Tuttavia, se usiamo la regressione lineare, alcune delle nostre stime potrebbero **essere al di fuori dell'intervallo [0, 1]**, il che non ha senso come probabilità. Non c'è nulla che "satura" l'uscita tra 0 e 1.

➡ **Logistic function (Sigmoid)**

Outline

Regressione logistica: formulazione del problema

Obiettivo: **Stimare la probabilità** che le osservazioni $\boldsymbol{\varphi} \in \mathbb{R}^{d\times 1}$ appartengano ad **una di due classi** $y \in \{0, 1\}$

Definiamo la combinazione lineare:

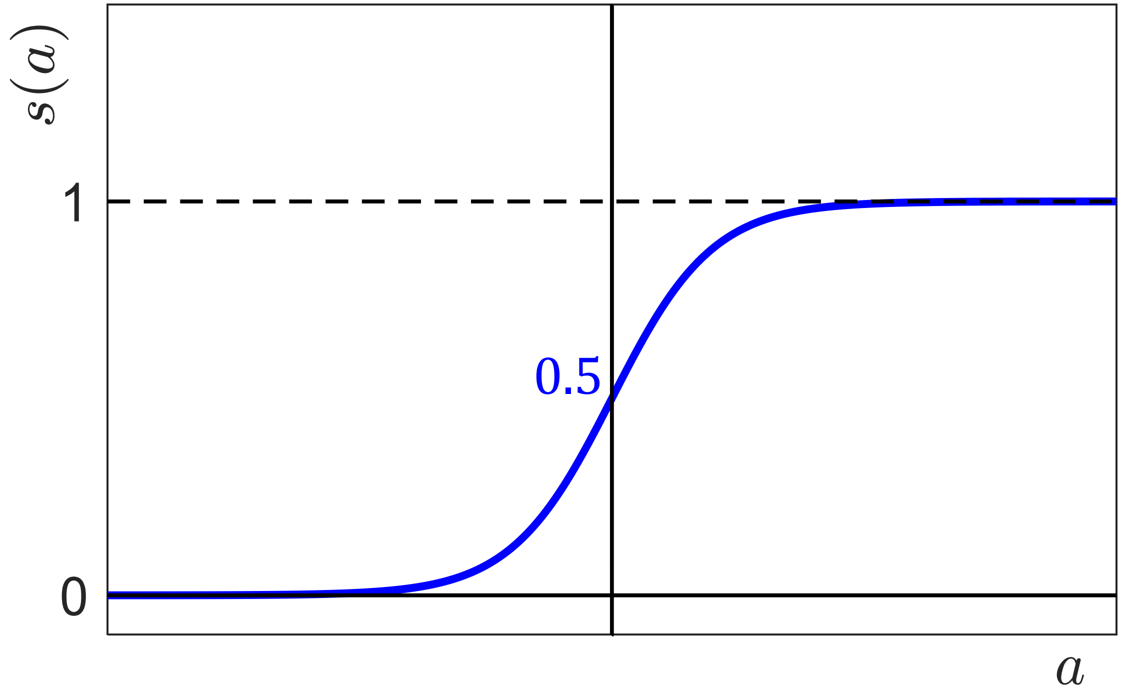
$$a = \sum_{j=0}^{d-1} \varphi_j \cdot \theta_j = \underset{1\times d}{\boldsymbol{\varphi}^\top} \cdot \underset{d\times 1}{\boldsymbol{\theta}}$$

La formula $s(a)$ è la **funzione logistica:**

$$s(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{1+e^a}$$

- $a \gg 0 \Rightarrow s(a) \approx 1$
- $a \ll 0 \Rightarrow s(a) \approx 0$

Funzione logistica (Sigmoide)



Regressione logistica: formulazione del problema

In particolare, il modello di regressione logistica modella la probabilità che $y = 1$ **tramite un modello lineare**

$$\boxed{P(y=1|\boldsymbol{\varphi}) = s(a) = s(\underset{1\times d}{\boldsymbol{\varphi}^\top}\underset{d\times 1}{\boldsymbol{\theta}}) = \frac{1}{1+e^{-\boldsymbol{\varphi}^\top\boldsymbol{\theta}}}}$$

L'output di $s(\boldsymbol{\varphi}^\top\boldsymbol{\theta})$ è **interpretato come una probabiltà**

- $\boldsymbol{\varphi}^\top\boldsymbol{\theta} \gg 0 \Rightarrow s(\boldsymbol{\varphi}^\top\boldsymbol{\theta}) \gg 0.5 \Rightarrow P(y=1|\boldsymbol{\varphi}) \approx 1$ $\Rightarrow$ **$\boldsymbol{\varphi}$ è classificato nella classe «1»**
- $\boldsymbol{\varphi}^\top\boldsymbol{\theta} \ll 0 \Rightarrow s(\boldsymbol{\varphi}^\top\boldsymbol{\theta}) \ll 0.5 \Rightarrow P(y=1|\boldsymbol{\varphi}) \approx 0$ $\Rightarrow$ **$\boldsymbol{\varphi}$ è classificato nella classe «0»**

Regressione lineare e regressione logistica

La regressione lineare e la regressione logistica fanno parte di una caterogia di modelli più generale detti **Generalized Linear Models (GLM)**

Regressione lineare

$$\mu = \boldsymbol{\varphi}^\top \boldsymbol{\theta} = \theta_0 + \theta_1 \varphi_1 + \dots + \theta_{d-1} \varphi_{d-1}$$

$$y \sim \mathcal{N}(\mu, \sigma^2)$$

Regressione logistica

$$\pi = s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) = s(\theta_0 + \theta_1 \varphi_1 + \dots + \theta_{d-1} \varphi_{d-1})$$

Link function

$$y \sim \mathrm{Bernoulli}(\pi) = \pi^y \cdot (1 - \pi)^{1-y}$$

Probabilità che $y = 1$

In questi modelli, un parametro di «tendenza centrale» di una distribuzione di probabilità è modellato tramite un modello lineare. I dati sono poi modellati come realizzazioni di questa distribuzione

Outline

Regressione logistica: funzione di costo

Supponiamo di avere a disposizione un dataset $\mathcal{D} = \{(\boldsymbol{\varphi}(1), y(1)), \dots, (\boldsymbol{\varphi}(N), y(N))\}$ dove $\boldsymbol{\varphi} \in \mathbb{R}^{d\times 1}$ e $y(i) \in \{0,1\}, i = 1, \dots, N,\ \text{i.i.d.}$ → Notiamo che y è una **variabile categorica**

Vogliamo modellare i dati tramite una regressione logisica $$P(y = 1|\boldsymbol{\varphi}) = \frac{1}{1 + e^{-\boldsymbol{\varphi}^\top \boldsymbol{\theta}}} \equiv \pi$$

Interpretiamo i dati come **realizzazioni** di una distribuzione di **Bernoulli** $y \sim \mathrm{Bernoulli}(\pi)$

Procederemo nel modo seguente:

- Calcolo della meno-log-likelihood $J(\boldsymbol{\theta}) = -\ln \mathcal{L}(\pi|\mathcal{D})$
- Calcolo del gradiente $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
- Ottimizzazione per trovare il minimo di $J(\boldsymbol{\theta})$

Regressione logistica: funzione di costo

Calcoliamo la verosimiglianza

$$\pi(i) \equiv P(y(i) = 1|\boldsymbol{\varphi}(i)) = \frac{1}{1 + e^{-\boldsymbol{\varphi}^\top(i)\boldsymbol{\theta}}}$$

$$\mathcal{L}(\pi|Y) = \prod_{i=1}^{N} \pi(i)^{y(i)} \cdot \big(1 - \pi(i)\big)^{1-y(i)} \Rightarrow$$ Calcolo la meno-log-likelihood $\Rightarrow$

$$-\ln[\mathcal{L}(\pi|Y)] = -\ln\left[\prod_{i=1}^{N} \pi(i)^{y(i)} \cdot \big(1 - \pi(i)\big)^{1-y(i)}\right] = -\sum_{i=1}^{N} \ln\left[\pi(i)^{y(i)} \cdot \big(1 - \pi(i)\big)^{1-y(i)}\right]$$

$$= -\sum_{i=1}^{N} \left(\ln\left[\pi(i)^{y(i)}\right] + \ln\left[\big(1 - \pi(i)\big)^{1-y(i)}\right]\right) \quad \boxed{= -\sum_{i=1}^{N} \Big(y(i) \cdot \ln \pi(i) + \big(1 - y(i)\big) \cdot \ln[1 - \pi(i)] \Big)}$$

$$\equiv J(\boldsymbol{\theta})$$

Interpretazione della funzione di costo

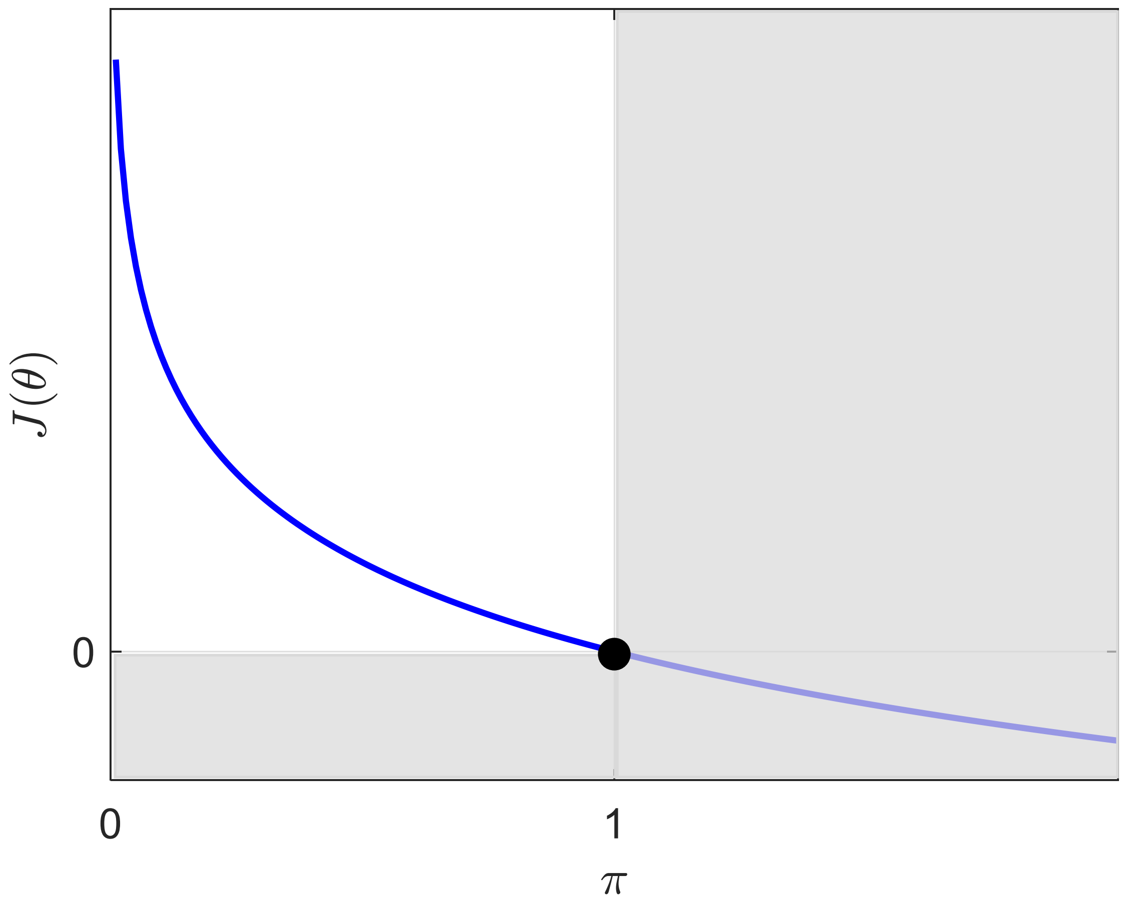
Assumiamo ci sia **un solo dato** $\mathcal{D} = \{(\boldsymbol{\varphi}, y)\}$

$$\Rightarrow J(\boldsymbol{\theta}) = \begin{cases} -\ln \pi & \text{se} \quad y = 1 \\ -\ln[1-\pi] & \text{se} \quad y = 0 \end{cases}$$

Caso $y = 1$

$$\boxed{J(\boldsymbol{\theta}) = -\ln \pi}$$

- $J(\boldsymbol{\theta}) \approx 0$ se $y = 1$ e $\pi \approx 1$
- $J(\boldsymbol{\theta}) \approx +\infty$ se $y = 1$ e $\pi \approx 0$



Interpretazione della funzione di costo

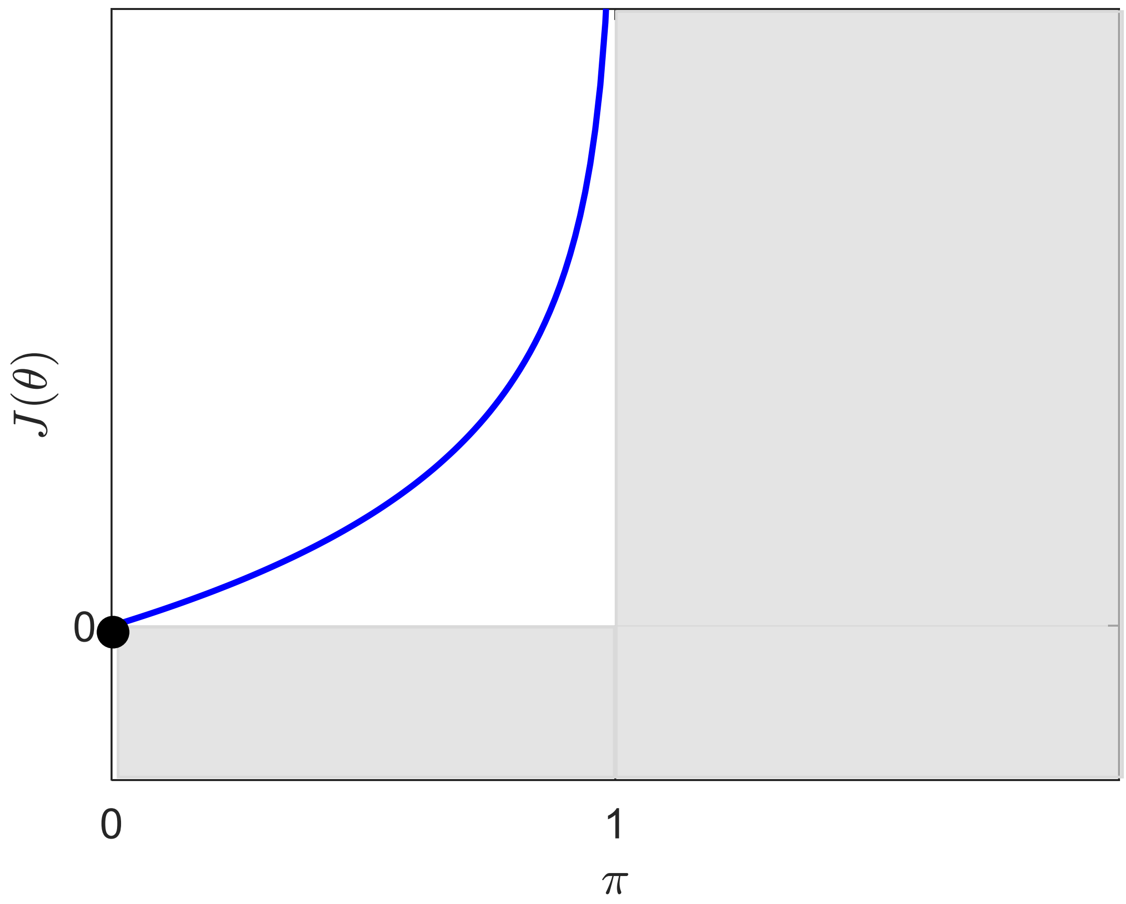
Assumiamo ci sia **un solo dato** $\mathcal{D} = \{(\boldsymbol{\varphi}, y)\}$

$$\Rightarrow J(\boldsymbol{\theta}) = \begin{cases} -\ln \pi & \text{se} \quad y = 1 \\ -\ln[1-\pi] & \text{se} \quad y = 0 \end{cases}$$

Caso $y = 0$

$$\boxed{J(\boldsymbol{\theta}) = -\ln[1-\pi]}$$

- $J(\boldsymbol{\theta}) \approx 0$ se $y = 0$ e $\pi \approx 0$
- $J(\boldsymbol{\theta}) \approx +\infty$ se $y = 0$ e $\pi \approx 1$



QUIZ!

Nella funzione di costo della regressione logistica, dove sono i parametri $\boldsymbol{\theta}$ che vogliamo stimare?

- ❑ Nei termini $y(i)$
- ❑ Nei termini $\ln$
- ❑ Nei termini $\pi(i)$

$$J(\boldsymbol{\theta}) = -\sum_{i=1}^{N}\Big(y(i) \cdot \ln \pi(i) + \big(1 - y(i)\big) \cdot \ln[1 - \pi(i)] \Big)$$

APPROFONDIMENTO

Calcolo del gradiente

Dobbiamo calcolare il gradiente di $J(\boldsymbol{\theta})$ rispetto a $\boldsymbol{\theta} \in \mathbb{R}^{d\times 1}$. Per prima cosa, calcoliamo la derivate di $s(a) = \frac{1}{1+e^{-a}}$ rispetto allo scalare $a \in \mathbb{R}$

$$\frac{ds(a)}{da} = \frac{d}{da}\left[\frac{1}{1+e^{-a}}\right] = \frac{d}{fa}\left[(1+e^{-a})^{-1}\right] = \frac{1}{(1+e^{-a})} \cdot \frac{e^{-a}}{(1+e^{-a})} = \frac{1}{(1+e^{-a})} \cdot \frac{(1+e^{-a})-1}{1+e^{-a}} =$$

$$= \frac{1}{1+e^{-a}} \cdot \left(\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right) = \boxed{s(a) \cdot [1 - s(a)]}$$

Nel caso in cui $a = \boldsymbol{\varphi}^\top \boldsymbol{\theta}$, abbiamo

$$\underbrace{\nabla_{\boldsymbol{\theta}} s(\boldsymbol{\varphi}^\top \boldsymbol{\theta})}_{d\times 1} = \underbrace{\boldsymbol{\varphi}}_{d\times 1} \cdot \underbrace{s(\boldsymbol{\varphi}^\top \boldsymbol{\theta})}_{1\times 1} \cdot \underbrace{[1 - s(\boldsymbol{\varphi}^\top \boldsymbol{\theta})]}_{1\times 1} = \underbrace{\boldsymbol{\varphi} \cdot \pi \cdot [1-\pi]}_{d\times 1}$$

APPROFONDIMENTO

Calcolo del gradiente

Possiamo ora calcolare il gradiente di $J(\boldsymbol{\theta})$

$$J(\boldsymbol{\theta}) = -\sum_{i=1}^{N}\Big(y(i)\ln \pi(i) + \big(1 - y(i)\big)\ln[1-\pi(i)] \Big) \qquad \pi(i) = \frac{1}{1+e^{-\boldsymbol{\varphi}(i)^\top \boldsymbol{\theta}}}$$

$$\underset{d\times 1}{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})} = -\sum_{i=1}^{N}\left(y(i)\frac{\pi'(i)}{\pi(i)} + \big(1-y(i)\big)\frac{-\pi'(i)}{1-\pi(i)}\right)$$

$$= -\sum_{i=1}^{N}\left(y(i)\frac{\boldsymbol{\varphi}(i)\cancel{\pi(i)}[1-\pi(i)]}{\cancel{\pi(i)}} + \big(1-y(i)\big)\frac{-\boldsymbol{\varphi}(i)\pi(i)\cancel{[1-\pi(i)]}}{\cancel{1-\pi(i)}}\right)$$

APPROFONDIMENTO

Calcolo del gradiente

$$= \sum_{i=1}^{N} \left(-y(i)\boldsymbol{\varphi}(i)[1-\pi(i)] - \big(1-y(i)\big)\big(-\boldsymbol{\varphi}(i)\pi(i)\big) \right)$$

$$= \sum_{i=1}^{N} \left(\boldsymbol{\varphi}(i) \cdot [-y(i) + y(i)\pi(i)] + \boldsymbol{\varphi}(i) \cdot [\pi(i) - y(i)\pi(i)] \right)$$

$$= \sum_{i=1}^{N} \left(\boldsymbol{\varphi}(i) \cdot [-y(i) + y(i)\pi(i) - y(i)\pi(i) + \pi(i)] \right)$$

Gradiente $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

$$\boxed{= \sum_{i=1}^{N} \underbrace{\boldsymbol{\varphi}(i)}_{d \times 1} \cdot \underbrace{\big(\pi(i) - y(i)\big)}_{1 \times 1}}$$

Outline

1. Il problema della classificazione
2. Perché non usare la regressione lineare?
3. Regressione logistica: formulazione del problema
4. Regressione logistica: funzione di costo
5. **Riassunto**
6. Esercizi con codice

Riassunto

Il modello di regressione logistica, nonostante il suo nome, non viene utilizzato per la regressione, ma per la **classificazione**

Una volta che il modello stima la probabilità di una classe, possiamo classificare un dato in una particolare classe se la probabilità per quella classe è **superiore a una soglia** (di solito 0.5)

La funzione che stiamo stimando è: $f(\boldsymbol{\varphi}) = P(y = 1|\boldsymbol{\varphi})$

La regressione logistica tenta di modellare f utilizzando il modello: $s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) = \dfrac{1}{1 + e^{-\boldsymbol{\varphi}^\top \boldsymbol{\theta}}}$

Il punto $\boldsymbol{\varphi}$ può quindi essere classificato alla classe $y = 1$ se $s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) \geq 0.5$

Riassunto

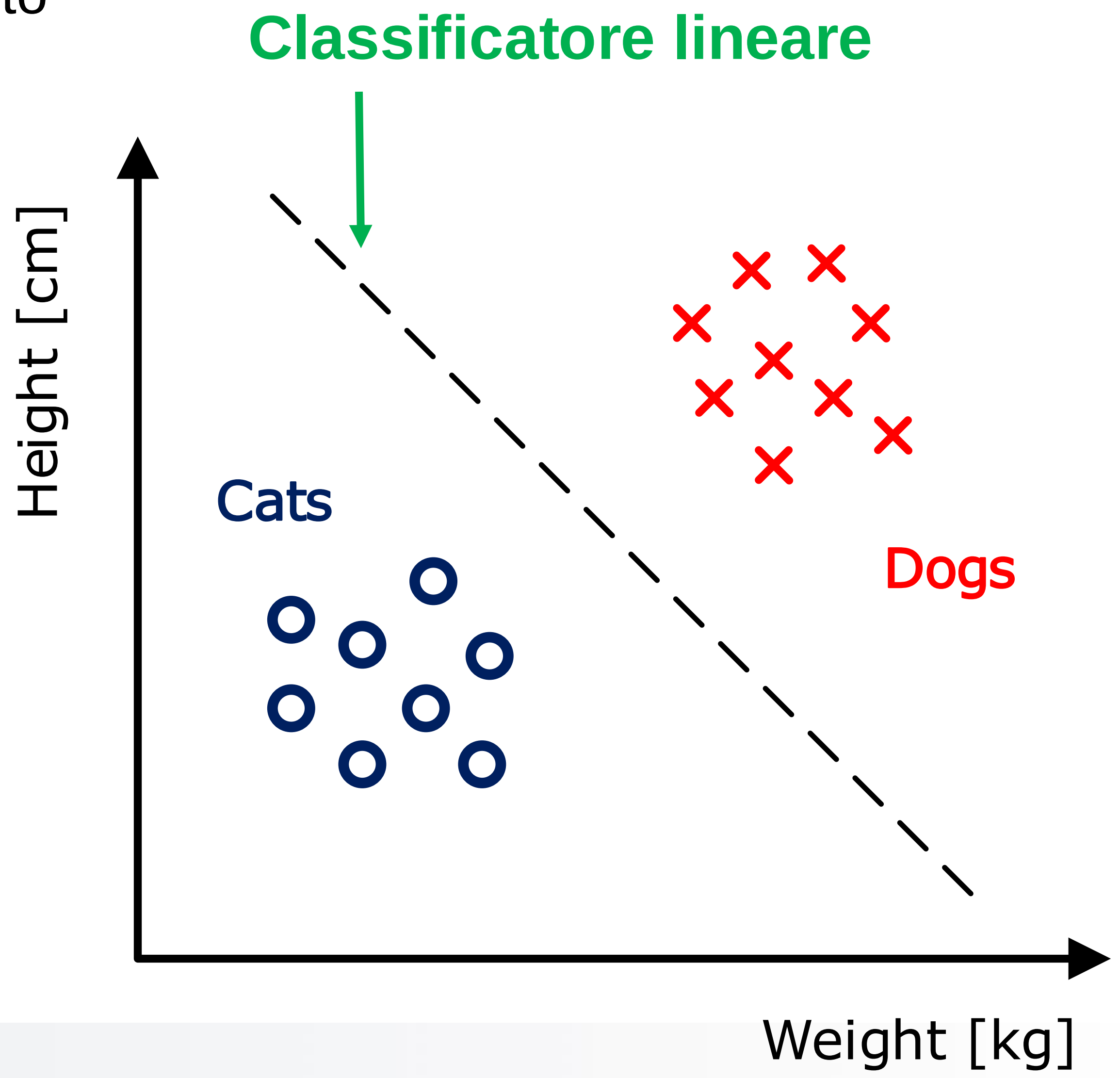
Il **confine di classificazione** che viene generato dalla regressione logistica è **lineare**

Infatti, classificare con la regola:

$$y = 1 \quad \textbf{if} \quad s(\boldsymbol{\varphi}^\top \boldsymbol{\theta}) \geq 0.5$$

è **equivalente** a dire

$$y = 1 \quad \textbf{if} \quad \underbrace{\boldsymbol{\varphi}^\top \boldsymbol{\theta}}_{\text{modello lineare}} \geq 0$$



Outline

1. Il problema della classificazione
2. Perché non usare la regressione lineare?
3. Regressione logistica: formulazione del problema
4. Regressione logistica: funzione di costo
5. Riassunto
6. **Esercizi con codice**

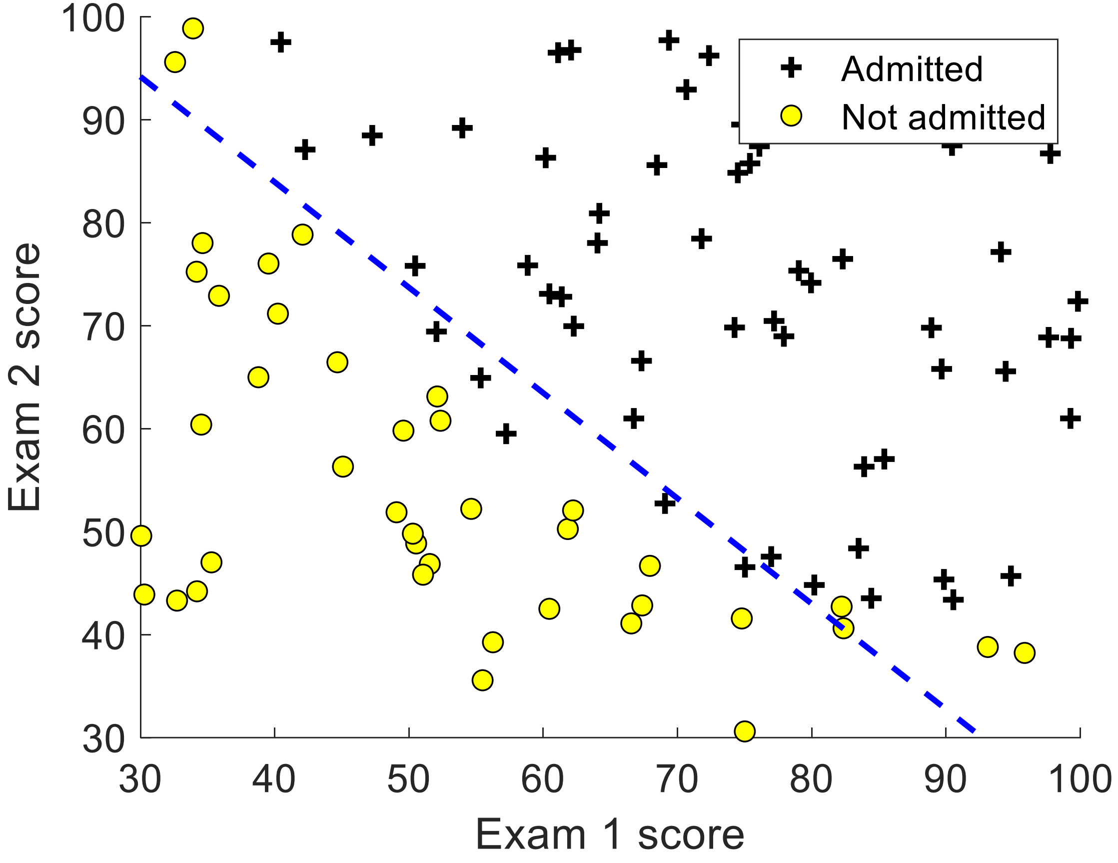
Esercizio: stima ammissione studenti

Vogliamo stimare la **probabilità di ammissione** $P(y=1)$ di uno studente (o studentessa) all'università, visti i risultati di due esami (φ_1, φ_2), tramite una regressione logistica

Il dataset consiste di $N = 100$ studenti con $\varphi_1(i), \varphi_2(i)$ e $y(i) \in \{0,1\}$, per $i = 1, \dots, N$

$$P(y=1|\boldsymbol{\varphi}) = s(a) = s(\underset{1\times 3}{\boldsymbol{\varphi}^\top}\underset{3\times 1}{\boldsymbol{\theta}}) = \frac{1}{1+e^{-\boldsymbol{\varphi}^\top\boldsymbol{\theta}}}$$

- Matrice dei dati $X \in \mathbb{R}^{100\times 3}$
- Vettore delle label $Y \in \mathbb{R}^{100\times 1}$
- Vettore dei parametri $\boldsymbol{\theta} \in \mathbb{R}^{3\times 1}$



UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Dipartimento
di Ingegneria Gestionale,
dell'Informazione e della Produzione