Università degli Studi di Palermo

dipartimento di ingegneria unipa

Stimatori, Stima e Campionamento

CORSO DI BIG DATA

a.a. 2022/2023

Prof. Roberto Pirrone

LABORATORIO DI INTERAZIONE UOMO-MACCHINA
CHILAB

Sommario

- Definizione di stimatore
- Stimatori polarizzati e non polarizzati
- Campionamento e stimatori
- Stima MLE e MAP

Definizione di stimatore

- Uno stimatore o una *statistica* è una funzione dei dati in nostro possesso la quale cerca di fornire la *migliore predizione possibile* di una quantità o funzione cui siamo interessati

  - Stima puntuale di un parametro $\boldsymbol{\vartheta}$ in funzione dei dati $\boldsymbol{x}^{(i)}$

$$\hat{\boldsymbol{\theta}}_m = g(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)})$$

  - Stima di una variabile $\boldsymbol{y}$ funzione dei dati $\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon}$
    - La relazione tra $\boldsymbol{y}$ ed $\boldsymbol{x}$ *non è* completamente descritta tramite f
    - Stimiamo una funzione che approssima f
    - È una stima puntuale nello spazio delle funzioni

Stimatori polarizzati e non polarizzati

- La polarizzazione o *bias* di uno stimatore si misura come la differenza tra il suo valore atteso ed il valore vero della quantità da stimare

$$\text{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}\left[\hat{\boldsymbol{\theta}}_m\right] - \boldsymbol{\theta}$$

  - Dipende dai dati che abbiamo per calcolare la stima

- Stimatore non polarizzato

$$\mathbb{E}\left[\hat{\boldsymbol{\theta}}_m\right] = \boldsymbol{\theta}$$

- Stimatore asintoticamente non polarizzato

$$\lim_{m \to \infty} \mathbb{E}\left[\hat{\boldsymbol{\theta}}_m\right] = \boldsymbol{\theta}$$

Stimatori campionari di media e varianza

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \hat{\mu}_m)^2$$

- Assumiamo che i campioni siano tratti da una distribuzione normale

$$\forall i,\ x^{(i)} \sim \mathcal{N}(x; \mu, \sigma)$$

Stimatori campionari di media e varianza

$$\text{bias}(\hat{\mu}_m) = \mathbb{E}\left[\hat{\mu}_m\right] - \mu =$$

$$= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \mu = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right] - \mu =$$

$$= \frac{m \cdot \mu}{m} - \mu = 0, \quad \mathbb{E}_{x^{(i)} \sim \mathcal{N}}\left[x^{(i)}\right] \triangleq \mu$$

Non polarizzato

- Assumiamo che i campioni siano tratti da una distribuzione normale

$$\forall i,\ x^{(i)} \sim \mathcal{N}(x; \mu, \sigma)$$

Stimatori campionari di media e varianza

$$\hat{\mu}_m - \mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} - \mu = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \mu\right);$$

$$m(\hat{\mu}_m - \mu) = \sum_{i=1}^{m}\left(x^{(i)} - \mu\right)$$

$$\mathbb{E}\left[\hat{\sigma}_m^2\right] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m + \mu - \mu\right)^2\right] =$$
$$= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\left(x^{(i)} - \mu\right) - \left(\hat{\mu}_m - \mu\right)\right)^2\right]$$

- Assumiamo che i campioni siano tratti da una distribuzione normale

$$\forall i,\ x^{(i)} \sim \mathcal{N}(x; \mu, \sigma)$$

Stimatori campionari di media e varianza

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\left(x^{(i)}-\mu\right)-\left(\hat{\mu}_m-\mu\right)\right)^2\right]=$$

$$=\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)}-\mu\right)^2-\frac{2}{m}\left(\hat{\mu}_m-\mu\right)\sum_{i=1}^{m}\left(x^{(i)}-\mu\right)+\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\mu}_m-\mu\right)^2\right]=$$

$$=\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)}-\mu\right)^2-\frac{2\cdot m}{m}\left(\hat{\mu}_m-\mu\right)^2+\frac{m}{m}\left(\hat{\mu}_m-\mu\right)^2\right]=$$

$$=\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)}-\mu\right)^2\right]-\mathbb{E}\left[\left(\hat{\mu}_m-\mu\right)^2\right]=$$

$$=\sigma^2-\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^{m}x^{(i)}\right]=\sigma^2-\frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}\left[x^{(i)}\right]=\sigma^2-\frac{m}{m^2}\sigma^2=\frac{m-1}{m}\sigma^2$$

Polarizzato

- Assumiamo che i campioni siano tratti da una distribuzione normale

$$\forall i,\ x^{(i)} \sim \mathcal{N}(x;\mu,\sigma)$$

Stimatori campionari di media e varianza

$$\tilde{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2$$

$$\mathbb{E}\left[\tilde{\sigma}_m^2\right] = \mathbb{E}\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right]$$

$$= \frac{m}{m-1}\mathbb{E}\left[\hat{\sigma}_m^2\right] = \frac{m}{m-1}\left(\frac{m-1}{m}\sigma^2\right) = \sigma^2$$ *Non polarizzato*

- Assumiamo che i campioni siano tratti da una distribuzione normale

$$\forall i,\ x^{(i)} \sim \mathcal{N}(x; \mu, \sigma)$$

Varianza di uno stimatore

- È la misura utilizzata per quanto la stima sia stabile rispetto al variare dei campioni utilizzati

- La radice quadrata della varianza si definisce *standard error*

$$\mathrm{SE}(\hat{\theta}_m) = \sqrt{\mathrm{Var}(\hat{\theta}_m)}$$

Standard error della media campionaria

$$\mathrm{SE}\left(\hat{\mu}_m\right) = \sqrt{\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}}$$

Errore quadratico medio (MSE)

- Lo stimatore ideale ha il minimo bias, meglio se sia non polarizzato, e la minima varianza

- Per giudicare la bontà di una stima in questo senso si utilizza l'*errore quadratico medio* (mean square error – MSE)

$$\mathrm{MSE} = \mathbb{E}\left[(\hat{\theta}_m - \theta)^2\right] =$$

$$\mathrm{Bias}(\hat{\theta}_m)^2 + \mathrm{Var}(\hat{\theta}_m)$$

Campionamento e analisi di Monte Carlo

- Gli stimatori visti sinora sono semplici e rientrano nella famiglia di quelli detti di «Monte Carlo»

- I metodi di Monte Carlo restituiscono una stima della quantità ricercata con un certo ammontare casuale di errore

- In questo contesto il *campionamento* cioè la scelta dei campioni per formare la stima è essenziale

Campionamento e analisi di Monte Carlo

- In genere si campionano dati da una distribuzione per approssimare un integrale non trattabile numericamente con una serie finita di somme

- Tale integrale viene visto come il *calcolo di un valore atteso* stimato attraverso la corrispondente media

$$s = \int p(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} = \mathbb{E}_p \left[f(\mathbf{x}) \right]$$

Campionamento e analisi di Monte Carlo

- Campionamento con distribuzione uniforme dei campioni:
  - Fornisce gli stimatori visti in precedenza
  - La stima della varianza è polarizzata

Campionamento e analisi di Monte Carlo

- Campionamento stratificato:
  - La stima della media e della varianza dipendono dal numero degli strati *L*, che possono avere dimensioni diverse S_k ($S = \boldsymbol{\Sigma}_k S_k$) e dal numero di campioni per strato n_k

$$\hat{\mu} = \frac{1}{S}\sum_{k=1}^{L} S_k \mu_k, \;\; \hat{\sigma}^2 = \sum_{k=1}^{L} \left(\frac{S_k}{S}\right)^2 \frac{\sigma_k^2}{n_k}$$

$$S_k = \text{cost}, \;\; n_k = 1 \;\; \Rightarrow \hat{\sigma}^2 = \frac{1}{L}\sum_{k=1}^{L} \sigma_k^2, \;\; \hat{\mu} = \frac{1}{L}\sum_{k=1}^{L} \mu_k$$

Campionamento e analisi di Monte Carlo

- Campionamento per importanza
  - L'idea di base è quella di approssimare la distribuzione di probabilità di cui si vuole calcolare il valore atteso (l'integrale che è il vero obiettivo della stima)
  - Tale distribuzione è non nota, perché è non noto tutto l'integrando, ma si può introdurre una nuova distribuzione di probabilità, *nota*, dalla quale trarre i campioni
  - Chiamiamo questa distribuzione *funzione importanza* perché condiziona la scelta dei campioni che da questa sono tratti; auspicabilmente dovrebbe essere tale da approssimare la distribuzione originaria

Campionamento e analisi di Monte Carlo

- Campionamento per importanza

$$s = \int p(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} = \int q(\boldsymbol{x}) \frac{p(\boldsymbol{x}) f(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$$

$$s = \mathbb{E}_{\mathbf{x} \sim q} \left[\frac{pf}{q} \right]$$

Campionamento e analisi di Monte Carlo

- Campionamento per importanza
  - Analizziamo il comportamento degli stimatori

$$\hat{s}_q = \frac{1}{n} \sum_{i=1, \mathrm{x}^{(i)} \sim q}^{n} \frac{p\left(\boldsymbol{x}^{(i)}\right) f\left(\boldsymbol{x}^{(i)}\right)}{q\left(\boldsymbol{x}^{(i)}\right)}$$

$$\mathbb{E}_q\left[\hat{s}_q\right] = \frac{1}{n} \sum_{i=1, \mathrm{x}^{(i)} \sim q}^{n} q(\boldsymbol{x}^{(i)}) \frac{p\left(\boldsymbol{x}^{(i)}\right) f\left(\boldsymbol{x}^{(i)}\right)}{q\left(\boldsymbol{x}^{(i)}\right)} = \mathbb{E}_p\left[\hat{s}_p\right] = s$$

Campionamento e analisi di Monte Carlo

- Campionamento per importanza
  - Analizziamo il comportamento degli stimatori

$$\mathrm{Var}\left[\hat{s}_q\right] = \mathrm{Var}\left[\frac{p(\mathbf{x})f(\mathbf{x})}{q(\mathbf{x})}\right]/n$$

Legge dei grandi numeri

$$q^*(\boldsymbol{x}) = \frac{p(\boldsymbol{x})|f(\boldsymbol{x})|}{Z} \Rightarrow \mathrm{Var}\left[\hat{s}_q^*\right] = 0$$

Basta *un solo campione* per
La stima esatta

  - Si possono utilizzare versioni non normalizzate di *p* e *q* che comportano una stima asintoticamente non polarizzata

Campionamento e analisi di Monte Carlo

- Campionamento per importanza
  - La scelta di q^* è critica rispetto alla reale magnitudo di $p|f|$ perché porta a sovrastimare/sottostimare pesantemente l'integrale e ad ottenere varianze troppo elevate
  - Possono insorgere problemi di stima in elevata dimensionalità di $\boldsymbol{x}$
  - Il campionamento per importanza è comunque molto utilizzato in diversi ambiti del machine learning
    - *Stochastic Gradient Descent*

Stima a massima verosimiglianza (MLE)

- La MLE è uno dei principi che possono essere utilizzati per definire se una particolare funzione *rappresenta un buon stimatore per una certa classe di modelli di apprendimento*

- Di conseguenza è uno dei principi fondanti del machine learning poiché se, dato il problema da risolvere, si riesce a identificare il modello più adatto, la MLE ci fornisce un criterio certo per poter stimare l'errore commesso e l'affidabilità della predizione.

Stima a massima verosimiglianza (MLE)

$$\mathbb{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}, \; p_{\text{data}}(\mathbf{x})$$

non nota

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \; p_{\text{model}} \left(\mathbb{X}; \boldsymbol{\theta}\right)$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{model}} \left(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}\right)$$

Assunzione di indipendenza statistica dei campioni

Stima a massima verosimiglianza (MLE)

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\mathrm{model}} \left(\boldsymbol{x}^{(i)}; \boldsymbol{\theta} \right) =$$

$$= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\mathrm{data}}} \log p_{\mathrm{model}} \left(\boldsymbol{x}; \boldsymbol{\theta} \right)$$

nota empiricamente
dai dati osservati → distribuzione empirica

Stima a massima verosimiglianza (MLE)

- Possiamo assumere di voler minimizzare la distanza tra la distribuzione empirica sui dati e quella del modello in termini della loro D_{KL}

non dipende dai parametri del modello

$$D_{\mathrm{KL}}\left(\hat{p}_{\mathrm{data}} \| p_{\mathrm{model}}\right) = \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}}\left[\log \hat{p}_{\mathrm{data}}(\boldsymbol{x}) - \log p_{\mathrm{model}}(\boldsymbol{x})\right] =$$

$$= -\mathbb{E}_{\mathbf{x}\sim\hat{p}_{\mathrm{data}}}\left[\log p_{\mathrm{model}}(\boldsymbol{x})\right]$$

Cross-entropia della distribuzione del modello Rispetto a quella dei dati!!

Stima a massima verosimiglianza (MLE)

- La MLE può estendersi anche alle probabilità condizionali p($\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\vartheta}$) che rappresentano ciò che viene stima dai classificatori

- Siamo sotto l'ipotesi base del machine learning sui dati e cioè che essi siano *indipendenti e identicamente distribuiti* (*i.i.d*. assumption)

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\theta}) =$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

ipotesi i.i.d.

Stima a massima verosimiglianza (MLE)

- Il MSE corrisponde alla MLE se la distribuzione è gaussiana

$$p(y|\boldsymbol{x}) = \mathcal{N}(y; \hat{y}(\boldsymbol{x}; \boldsymbol{w}), \sigma^2)$$

$$\sum_{i=1}^{m} \log p\left(y^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{w}\right)$$

$$= -m \log \sigma - \frac{m}{2} \log(2\pi) - \sum_{i=1}^{m} \frac{\left\|\hat{y}^{(i)} - y^{(i)}\right\|^2}{2\sigma^2}$$

- Si assuma di avere un modello che forma una stima ŷ (***x***;***w***) di una variabile *y*

$$\mathrm{MSE}_{\mathrm{train}} = \frac{1}{m} \sum_{i=1}^{m} \left\|\hat{y}^{(i)} - y^{(i)}\right\|^2$$

Minimizzare il MSE corrisponde a massimizzare la MLE

Stima Maximum A Posteriori (MAP)

- La MLE rappresenta l'approccio alla massimizzazione della stima nell'ottica della *statistica frequentista*
  - La stima del parametro $\boldsymbol{\vartheta}$ è una variabile statistica che è funzione dei dati, anch'essi visti come casuali
  - L'incertezza sulla stima è rappresentata dalla sua varianza

- Nella *statistica Bayesiana*, al contrario, i dati sono considerati come già osservati e non sono casuali mentre $\boldsymbol{\vartheta}$ è esatto, ma non noto

- La nostra conoscenza imperfetta su $\boldsymbol{\vartheta}$ è rappresentata dal *prior* e cioè la sua distribuzione di probabilità $p(\boldsymbol{\vartheta})$ prima di osservare i dati

Stima Maximum A Posteriori (MAP)

- Il processo di stima si avvale della regola di Bayes che stima una *distribuzione di probabilità a posteriori su* $\boldsymbol{\vartheta}$

posterior *likelihood* *prior*

$$p\left(\boldsymbol{\theta}|x^{(1)},\ldots,x^{(m)}\right) = \frac{p\left(x^{(1)},\ldots,x^{(m)}|\boldsymbol{\theta}\right)p(\boldsymbol{\theta})}{p\left(x^{(1)},\ldots,x^{(m)}\right)}$$

Costante perché I dati sono noti

- In questo contesto è importante la scelta del prior:
  - Deve riflettere l'incertezza a priori sul valore di $\boldsymbol{\vartheta}$
  - Distribuzione uniforme o gaussiana sul dominio di variazione
  - Elevata entropia del prior, cioè *elevato contenuto informativo*

Stima Maximum A Posteriori (MAP)

- La stima MAP non è altro che la scelta di un valore puntuale di $\boldsymbol{\vartheta}$ al posto di un'intera distribuzione

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

  - Si sceglie il valore di $\boldsymbol{\vartheta}$ che massimizza il posterior
  - Si tratta della somma del MLE e del prior
  - La stima MAP riduce la varianza *dello stimatore* rispetto a quella MLE
  - Utile quando abbiamo una conoscenza su $\boldsymbol{\vartheta}$ che non possiamo trovare nei dati