



UNIVERSITÀ
DEGLI STUDI
DI PALERMO



Stimatori, Stima e Campionamento

CORSO DI BIG DATA
a.a. 2019/2020

Prof. Roberto Pirrone

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 ϑ in funzione dei dati $\mathbf{x}^{(i)}$

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

- Stima di una variabile y funzione dei dati $y = f(x) + \varepsilon$
 - La relazione tra y ed 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

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

- Dipende dai dati che abbiamo per calcolare la stima
- Stimatore non polarizzato
- Stimatore asintoticamente non polarizzato

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

$$\lim_{m \rightarrow \infty} \mathbb{E} [\hat{\theta}_m] = \theta$$

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

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

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

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

- 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

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

$$\begin{aligned}\mathbb{E} [\hat{\sigma}_m^2] &= \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^m (x^{(i)} - \hat{\mu}_m)^2 \right] = \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^m (x^{(i)} - \hat{\mu}_m + \mu - \mu)^2 \right] = \\ &= \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^m ((x^{(i)} - \mu) - (\hat{\mu}_m - \mu))^2 \right] = \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)^2 \right] - \mathbb{E} [(\hat{\mu}_m - \mu)^2] = \\ &= \sigma^2 - \text{Var} \left[\frac{1}{m} \sum_{i=1}^m x^{(i)} \right] = \sigma^2 - \frac{1}{m^2} \sum_{i=1}^m \text{Var} [x^{(i)}] = \sigma^2 - \frac{m}{m^2} \sigma^2 = \frac{m-1}{m} \sigma^2 \quad \textcolor{red}{Polarizzato}\end{aligned}$$

- 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

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

$$\begin{aligned}\mathbb{E} [\tilde{\sigma}_m^2] &= \mathbb{E} \left[\frac{1}{m-1} \sum_{i=1}^m \left(x^{(i)} - \hat{\mu}_m \right)^2 \right] \\ &= \frac{m}{m-1} \mathbb{E} [\hat{\sigma}_m^2] = \frac{m}{m-1} \left(\frac{m-1}{m} \sigma^2 \right) = \sigma^2 \text{ *Non polarizzato*}\end{aligned}$$

- 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*

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

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)

$$\begin{aligned} \text{MSE} &= \mathbb{E} \left[(\hat{\theta}_m - \theta)^2 \right] = \\ &\quad \text{Bias}(\hat{\theta}_m)^2 + \text{Var}(\hat{\theta}_m) \end{aligned}$$

Campionamento e analisi di Monte Carlo

- Gli stimatori visti sinora sono semplici 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

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

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 varianza dipende dal numero degli strati, che possono avere dimensioni diverse, e dal numero di campioni per strato

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

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

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

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

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

Campionamento e analisi di Monte Carlo

- Campionamento per importanza
 - Analizziamo il comportamento degli estimatori

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

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

Campionamento e analisi di Monte Carlo

- Campionamento per importanza
 - Analizziamo il comportamento degli estimatori

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

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

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
 - Possono insorgere problemi di stima in elevata dimensionalità di 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)

$$\mathbb{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}, \quad p_{\text{data}}(\mathbf{x}) \quad \text{non nota}$$

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

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

Stima a massima verosimiglianza (MLE)

$$\begin{aligned}\theta_{\text{ML}} &= \arg \max_{\theta} \sum_{i=1}^m \log p_{\text{model}}(\mathbf{x}^{(i)}; \theta) = \\ &= \arg \max_{\theta} \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(\mathbf{x}; \theta)\end{aligned}$$

*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*

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

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

Stima a massima verosimiglianza (MLE)

- La MLE può estendersi anche alle probabilità condizionali $p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ 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)*

$$\begin{aligned}\boldsymbol{\theta}_{\text{ML}} &= \arg \max_{\boldsymbol{\theta}} P(\mathbf{Y}|\mathbf{X}; \boldsymbol{\theta}) = \\ &= \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^m \log P(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}; \boldsymbol{\theta})\end{aligned}$$

ipotesi i.i.d.

Stima a massima verosimiglianza (MLE)

- Il MSE corrisponde alla MLE se la distribuzione è gaussiana
- Si assume di avere un modello che forma una stima $\hat{y}(\mathbf{x}; \mathbf{w})$ di una variabile y

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

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

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

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

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 θ 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 θ è esatto, ma non noto
- La nostra conoscenza imperfetta su θ è rappresentata dal *prior* e cioè la sua distribuzione di probabilità $p(\theta)$ 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 θ*

$$p(\theta|x^{(1)}, \dots, x^{(m)}) = \frac{p(x^{(1)}, \dots, x^{(m)}|\theta) p(\theta)}{p(x^{(1)}, \dots, x^{(m)})}$$

posterior *likelihood* *prior*

*Costante perché
I dati sono noti*

- In questo contesto è importante la scelta del prior:
 - Deve riflettere l'incertezza a priori sul valore di θ
 - Distribuzione uniforme o gaussiana sul dominio di variazione
 - Elevata entropia del prior

Stima Maximum A Posteriori (MAP)

- La stima MAP non è altro che la scelta di un valore puntuale di θ al posto di un'intera distribuzione

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

- Si sceglie il valore di θ che massimizza il posterior
- Si tratta della somma del MLE e del prior
- La stima MAP riduce la varianza rispetto a quella MLE
- Utile quando abbiamo una conoscenza su θ che non possiamo trovare nei dati