# APPUNTI SULLA STATISTICA BAYESIANA

DARIO TREVISAN

## Indice



## 1. Introduzione

Tradizionalmente la *statistica* si occupa di raccogliere informazioni (perlopiù dati numerici o qualitativi, pensate ai censimenti), elaborarli ed eventualmente discutere la validità di modelli (inferenza). Da questo punto di vista, notiamo che questo approccio è comune a tante discipline scientifiche (Fisica, Chimica, Biologia, ma anche Medicina, Informatica): questo da un lato permette alla statistica di fornire strumenti *generali* per lo studio di problemi particolari, d'altra parte però non deve far cadere in errore che la *statistica* sia una scatola nera che può risolvere in modo automatico ogni problema!

Inoltre ogni settore scientifico tende a sviluppare tecniche proprie per trattare i dati che ha sulla base dei problemi tipici che incontra. Storicamente la Fisica (in particolare l'astronomia) è stata una delle prime discipline in cui "grandi" quantità di dati numerici erano a disposizione e dovevano essere elaborati per produrre previsioni e validare teorie: questo problema è stato affrontato da matematici quali Legendre, Gauss e Laplace che elaborarono alcuni strumenti basandosi sul calcolo delle probabilità.

Successivamente, nella prima metà del ventesimo secolo, le discipline biologiche (in particolare la genetica) fecero sì che altre tecniche fossero elaborate (da studiosi quali Fisher, Pearson e altri) per studiare altri problemi. Tali tecniche, di cui non ci occuperemo, vanno sotto il nome di statistica *frequentista*, perché portano con sé la filosofia che probabilità corrisponda alla *frequenza* di osservazioni ripetute (e indipendenti), un dato "oggettivo" che porta a pensare alla probabilità come ad un attributo *fisico* di una situazione.

In realtà, come abbiamo visto, è utile e possiamo accettare un concetto più ampio di probabilità, associandola al *grado di fiducia* che un soggetto ha sulla validità di un evento. Questa accezione, che era stata messa in discussione

dalla scuola frequentista, ma che prima era ben chiara a matematici come Gauss e Laplace, è poi "rinata" nella seconda metà del ventesimo secolo sulla base di tecniche che vengono dette di statistica *bayesiana*, perché fanno uso del teorema di Bayes.

Vedremo che la statistica bayesiana si fonda su due concetti molto semplici: il primo è quello (che ormai conosciamo bene) di usare il teorema di Bayes per aggiornare il grado di fiducia su una ipotesi (una teoria, o modello) nel momento in cui arriva nuova informazione che vogliamo incorporare; il secondo è quello di prendere decisioni cercando di minimizzare un opportuna nozione di rischio associato.

Anche la statistica bayesiana però non è immune da critiche: la principale è che i calcoli in forma chiusa difficilmente si possono portare a termine in tutti i problemi e quindi si deve ricorrere ad approssimazioni o simulazioni.

## 2. Probabilità a priori e a posteriori

Una volta raccolti dei dati, il problema principale della statistica è di usarli per validare o meno delle ipotesi. Tali ipotesi possono essere delle affermazioni molto semplici (ad esempio: "il numero di nascite sta diminuendo") oppure delle teorie molto complesse (ad esempio: "la teoria della relatività di Einstein"). Il vantaggio (ma che a volte è anche uno svantaggio) della statistica è che non si vuole entrare troppo nel dettaglio di *quali* dati si tratta né di quali ipotesi stiamo considerando. Ovviamente il prezzo è che le affermazioni saranno sempre di tipo probabilistico, ma sapremo quantificare il grado di fiducia/incertezza su quanto affermato.

Supponiamo quindi di avere un certo numero di ipotesi (a volte le chiamiamo teorie o modelli)

$$H_0, H_1, \dots, H_n.$$

che sono a due a due esclusive, e siamo sicuri che almeno una sia vera (nel senso che non ce ne vengono in mente altre che potrebbero essere ragionevolmente vere). Dal punto di vista probabilistico si tratta quindi di un sistema di alternative e quindi possiamo considerare la *variabile aleatoria*

$$\theta \in \{0, 1, \dots, n\}$$

per cui $\{\theta = i\}$ significa che l'ipotesi $H_i$ vale. Questo punto di vista ci permette anche di estendere il problema anche al caso in cui $\theta$ sia una variabile continua $\theta \in \Theta \subseteq \mathbb{R}$ (possiamo ad esempio pensare che il modello riguardi una certa quantità continua, ad esempio una temperatura o una posizione).

Il problema principale della statistica è quindi ridotto a *stimare* $\theta \in \Theta$ sulla base di osservazioni (o altre informazioni che possiamo ottenere). Come abbiamo già anticipato, nella statistica bayesiana, $\theta \in \Theta$ è una *variabile aleatoria*, assunzione del tutto naturale se pensiamo che siamo *incerti* su quale ipotesi valga. Per stimare $\theta$ allora ci basterà capire quale sia la legge di questa variabile aleatoria. Per fare ciò, procediamo in due passaggi:

(1) Stabiliamo, sulla base di tutta l'informazione $\Omega$ che abbiamo *prima* di fare osservazioni, la legge di $\theta$ (detta "probabilità a priori")

$$P(\theta = z|\Omega) \quad (\text{caso } \theta \text{ discreta}),$$

$$\varrho(\theta = z|\Omega) \quad \text{(caso } \theta \text{ continua)};$$

(2) Stabiliamo quale sia la probabilità di ricevere una certa informazione $I$ (ad esempio, mediante osservazioni), sapendo che vale $\{\theta = z\} \cap \Omega$, ossia
$$P(I|\Omega \cap \{\theta = z\})$$
oppure $\varrho(I|\Omega \cap \{\theta = z\})$ nel caso di informazione $I = \{Y = y\}$ per qualche variabile aleatoria continua $Y$.

Avendo fatto questo, la formula di Bayes ci permette di *aggiornare* la "probabilità a priori" e ottenere la "probabilità a posteriori":

$$P(\theta = z|\Omega \cap I) = P(\theta = z|\Omega)\frac{P(I|\Omega \cap \{\theta = z\})}{P(I|\Omega)},$$

nel caso discreto, altrimenti bisogna sostituire $P(\theta = z|\cdot)$ con $\varrho(\theta = |\cdot)$ nel caso $\theta$ continua (e pure $P(I|\cdot)$ con $\varrho(I|\cdot)$ nel caso di informazione $\{Y = y\}$ per una variabile continua $Y$).

Il termine $\frac{P(I|\Omega \cap \{\theta = z\})}{P(I|\Omega)}$ è anche detto *rapporto di verosimiglianza*.

Nonostante il meccanismo sia piuttosto semplice (e l'abbiamo in realtà già applicato molte volte nei problemi incontrati finora), sorgono diverse questioni/critiche, di cui la principale è: come stabilire la probabilità "a priori"? In realtà abbiamo già visto che spesso possiamo attribuire probabilità uniforme alle varie ipotesi (principio di Laplace), però il seguente esempio dovrebbe fare riflettere sulla importanza della scelta della probabilità a priori.

*Esempio* 1. Supponiamo di sapere che vi sono due teorie fisiche in competizione: la relatività di Einstein ($E$) e la relatività di Smith ($S$). Quella di Einstein afferma che nulla viaggia più veloce della luce (velocità $c$), mentre quella di Smith dice che in realtà i neutrini (delle particelle) viaggiano poco più veloce della luce. Alcuni scienziati affermano di aver fatto un esperimento da cui risulterebbe che la velocità dei neutrini è $c + \varepsilon$, dove $\varepsilon > 0$ è un numero molto piccolo, potenzialmente frutto di un errore. Supponiamo di dare inizialmente ($\Omega$ = prima di venire a sapere di questo esperimento)

$$P(E|\Omega) = p, \quad P(S|\Omega) = 1 - p,$$

dove $p \in (0, 1)$ è un numero fissato (in questo esempio ci serve per vedere come scelte di $p$ danno risultati diversi). Questo stabilisce la probabilità a priori. Supponiamo che $I$ sia l'informazione "un esperimento mostra che i neutrini viaggiano a velocità $c + \varepsilon$". Allora $P(I|E \cap \Omega)$ deve essere molto bassa (ad esempio, il risultato dell'esperimento dovrebbe essere un errore numerico), diciamo $10^{-3}$, mentre $P(I|S \cap \Omega)$ dovrebbe essere 1 (perché la teoria di Smith prevede appunto il risultato di questo esperimento). Per la formula di Bayes abbiamo

$$P(E|\Omega \cap I) = \frac{p10^{-3}}{10^{-3}p + 1 \cdot (1 - p)}.$$

Cosa accade se $p = 1/2$, ossia diamo probabilità a priori uniformi ad entrambe le teorie? Ne segue che

$$P(E|\Omega \cap I) \sim 10^{-3},$$

ossia la relatività di Einstein ne esce molto screditata (e invece Smith diventa molto più sicura). D'altra parte, se l'informazione $\Omega$ ci rendeva molto sicuri sulla relatività di Einstein, diciamo ad esempio $p = 1 - 10^{-5}$, ne segue che

$$P(E|\Omega \cap I) = \left(1 - 10^{-5}\right) \frac{10^{-3}}{10^{-3}(1 - 10^{-5}) + 1 \cdot 10^{-5}} \sim 1,$$

ossia questa "prova" non ci smuove molto dalle nostre convinzioni (e siamo quindi convinti che ci sia qualcosa di sbagliato nell'esperimento).

Chiaramente, la procedura di "inferenza bayesiana" si può ripetere nel momento in cui arrivi nuova informazione $I'$: basta considerare come probabilità "a priori" la probabilità a posteriori trovata dopo il primo passaggio, e così via... Il teorema di Bayes permette di aggiornare ogni volta il nostro grado di fiducia tenendo conto di ogni nuova informazione che riusciamo ad ottenere dalla realtà.

## 3. Decisioni e rischio bayesiani

Il secondo concetto fondamentale della statistica bayesiana è quello di decisione e rischio associato. Infatti, immaginiamo di avere aggiornato l'informazione a nostra disposizione (ad esempio sulla base di osservazioni) e di essere giunti alla legge a posteriori(discreta o continua) di $\theta \in \Theta$ rispetto alla informazione $I$. Questo ancora non ci darà la certezza di *quale* sia il "vero" $\theta$ , ma solamente delle probabilità.

Tuttavia nella realtà dobbiamo spesso comportarci come se un qualche $\hat{\theta}$ sia vero (pensate ad un medico che deve decidere se procedere o meno con una cura, oppure un sistema automatico di assistenza alla guida che deve decidere se fermare l'auto in presenza o meno di pericolo). Dovendo fare una "decisione" di un qualche $\hat{\theta}$, possiamo chiederci: a seconda della scelta $\hat{\theta}$, come posso valutare il "rischio" associato al fatto che il vero $\theta$ sia diverso? Introduciamo il seguente concetto.

**Definizione 2** (Funzione di costo/perdita e richio)**.** Data una funzione $L : \Theta \times \Theta \to \mathbb{R}$, $(\hat{\theta}, z) \mapsto L(\hat{\theta}, z)$ che associa alla scelta $\hat{\theta}$ il costo (o la perdita) se il "vero" parametro è $z$, diciamo che il *rischio* bayesiano associato alla scelta $\hat{\theta}$ è

$$\mathrm{Risk}(\hat{\theta}) = \mathrm{Risk}_L(\hat{\theta}|I) = \mathbb{E}\left[L(\hat{\theta}, \theta)|I\right].$$

Nel caso in cui $\theta$ sia una variabile discreta, abbiamo

$$\mathrm{Risk}_L(\hat{\theta}|I) = \sum_{z \in \Theta} L(\hat{\theta}, z) P(\theta = z|I),$$

mentre nel caso di variabile continua

$$\mathrm{Risk}_L(\hat{\theta}|I) = \int_{\Theta} L(\hat{\theta}, z) \varrho(\theta = z|I) \mathrm{d}z.$$

Una volta stabilito il rischio $\mathrm{Risk}_L$, possiamo cercare di individuare una *decisione bayesiana* $\hat{\theta}$ che *minimizzi* tale rischio, ossia scegliere $\hat{\theta}$ affinché

$$\mathrm{Risk}_L(\hat{\theta}|I) = \min_{u \in \Theta} \mathrm{Risk}_L(u|I).$$

Ovviamente, rischi diversi porteranno a decisioni diverse. Anche nel caso in cui $\Theta$ consista solo di due elementi (due ipotesi), i rischi associati a scelte

diverse possono essere ben diversi, e quindi non sempre conviene scegliere $\hat{\theta}$ tale che $P(\theta = \hat{\theta}|I)$ è massima, anche se questa è un sempre un possibile criterio.

Vediamo alcuni esempi di rischio e calcoliamone le decisioni minimizzanti.

*Esempio* 3 (moda a posteriori (o massima verosimiglianza bayesiana)). Sia $\theta \in \Theta$ una variabile aleatoria discreta e poniamo

$$L(\hat{\theta}, z) = \begin{cases} -1 & \text{se } \theta = z \\ 0 & \text{altrimenti} \end{cases}$$

Calcoliamo

$$\text{Risk}_L(\hat{\theta}|I) = \sum_{z \in \Theta} L(\hat{\theta}, z) P(\theta = z|I) = -P(\theta = \hat{\theta}|I).$$

Una decisione che minimizza il rischio è quindi $\hat{\theta} \in \Theta$ per cui $-P(\theta = \hat{\theta}|I)$ è minimo, o equivalentemente

$$P(\theta = \hat{\theta}|I) \quad \text{è massima.}$$

Nel caso in cui $\theta$ sia una variabile aleatoria continua, si può estendere il criterio cercando invece $\hat{\theta}$ per cui

$$\varrho(\theta = \hat{\theta}|I) \quad \text{sia massima,}$$

anche formalmente questa non rientra nella definizione di rischio data sopra. Nei calcoli spesso conviene passare ai logaritmi e massimizzare le funzioni

$$\hat{\theta} \mapsto \log(P(\theta = \hat{\theta}|I)) \quad \text{o} \quad \hat{\theta} \mapsto \log(\varrho(\theta = \hat{\theta}|I))$$

*Esercizio* 4. Si supponga che la densità di $\theta$ sia della forma

$$\varrho\left(\theta = x|I\right) = c \exp\left(-2x^4 + 8x\right) \quad \text{per } x \in \mathbb{R}.$$

($c$ è una costante che assicura che l'integrale su tutto $\mathbb{R}$ sia 1). Determinare $\hat{\theta}$ per cui la densità è massima.

*Esempio* 5 (rischio quadratico). Sia $\Theta \subseteq \mathbb{R}$ un intervallo e $\theta \in \Theta$ una variabile aleatoria (discreta o continua) e poniamo

$$L(\hat{\theta}, z) = (\hat{\theta} - z)^2.$$

In tal caso la decisione che minimizza il rischio è

$$\hat{\theta} = \mathbb{E}\left[\theta|I\right],$$

e il rischio associato (alla decisione minimizzante) è

$$\text{Risk}(\mathbb{E}\left[\theta|I\right]|I) = \mathbb{E}\left[(\theta - \mathbb{E}\left[\theta|I\right])^2|I\right] = \text{Var}\left(\theta|I\right).$$

Per mostrare queste affermazioni, notiamo che per una decisione generale $\hat{\theta}$ il rischio si decompone nel seguente modo:

$$\begin{aligned} \text{Risk}(\hat{\theta}|I) &= \mathbb{E}\left[(\theta - \hat{\theta})^2|I\right] \\ &= \mathbb{E}\left[(\theta - \mathbb{E}\left[\theta|I\right] + \mathbb{E}\left[\theta|I\right] - \hat{\theta})^2|I\right] \\ &= \mathbb{E}\left[(\theta - \mathbb{E}\left[\theta|I\right])^2 + (\mathbb{E}\left[\theta|I\right] - \hat{\theta})^2 + 2(\theta - \mathbb{E}\left[\theta|I\right])(\mathbb{E}\left[\theta|I\right] - \hat{\theta})|I\right] \\ &= \text{Var}\left(\theta|I\right) + (\mathbb{E}\left[\theta|I\right] - \hat{\theta})^2 \end{aligned}$$

perché

$$\mathbb{E}\left[2(\theta - \mathbb{E}\left[\theta|I\right])(\mathbb{E}\left[\theta|I\right] - \hat{\theta})|I\right] = 2(\mathbb{E}\left[\theta|I\right] - \hat{\theta})\mathbb{E}\left[\theta - \mathbb{E}\left[\theta|I\right]|I\right] = 0.$$

Di conseguenza, si ha

$$\mathrm{Risk}(\hat{\theta}|I) \geq \mathrm{Var}\left(\theta|I\right)$$

e il minimo si ottiene quando $(\mathbb{E}\left[\theta|I\right] - \hat{\theta})^2 = 0$, ossia $\hat{\theta} = \mathbb{E}\left[\theta|I\right]$.

*Esempio* 6 (mediana). Sia $\Theta \subseteq \mathbb{R}$ un intervallo e $\theta \in \Theta \subseteq \mathbb{R}$ una variabile aleatoria continua e poniamo

$$L(\hat{\theta}, z) = |\hat{\theta} - z|.$$

In questo caso una decisione minimizzante è $\hat{\theta} = \mu$ data da una *mediana* per la variabile aleatoria $\theta$, che è definita dalla relazione

$$P(\theta \leq \mu|I) = P(\theta > \mu|I) = \frac{1}{2}.$$

Attenzione: in generale $\mu$ non è unica (dare un esempio)!

Per mostrare questa proprietà, notiamo intanto che vale sempre la disuguaglianza, per ogni $x$, $h \in \mathbb{R}$,

$$|x + h| \geq |x| + \chi_{\{x>0\}}h - \chi_{\{x\leq 0\}}h$$

(geometricamente esprime il fatto che la funzione $x \mapsto |x|$ è convessa, ossia sempre al di sopra della retta tangente al grafico ). Poniamo $x = \theta - \mu$ e $h = \mu - \hat{\theta}$, dove $\hat{\theta} \in \Theta$ è una qualsiasi decisione. Pertanto otteniamo la disuguaglianza tra variabili aleatorie

$$|\theta - \hat{\theta}| \geq |\theta - \mu| + \chi_{\{\theta>\mu\}}(\mu - \hat{\theta}) - \chi_{\{\theta\leq\mu\}}(\mu - \hat{\theta}).$$

Passando ai valori attesi si trova

$$\mathbb{E}\left[|\theta - \hat{\theta}||I\right] \geq \mathbb{E}\left[|\theta - \mu||I\right],$$

perché

$$\mathbb{E}\left[\chi_{\{\theta>\mu\}}(\mu - \hat{\theta}) - \chi_{\{\theta\leq\mu\}}(\mu - \hat{\theta})|I\right] = \left(P(\theta > \mu|I) - P(\theta \leq \mu|I)\right)(\mu - \hat{\theta}) = 0.$$

Nel caso in cui $\theta$ sia discreto, si trova che una decisione minimizzante è sempre una mediana $\mu$, la cui definizione però cambia: deve valere

$$P(\theta \leq \mu|I) \geq \frac{1}{2} \quad \text{e} \quad P(\theta \geq \mu|I) \geq \frac{1}{2}.$$

*Esercizio* 7. Calcolare tutte le mediane $\mu$ nel caso di $\theta$ avente

(1) legge uniforme su $\{1, 2, 3, 4, 5, 6\}$,
(2) legge $\mathrm{Bin}(n, \frac{1}{2})$,
(3) legge $\mathrm{Geom}(\frac{1}{2})$,
(4) legge esponenziale $\mathcal{E}(\lambda)$,
(5) legge $\mathcal{N}(m, \sigma^2)$.

## 4. Test bayesiani

Nel caso in cui $\Theta = \{0, 1\}$, corrispondente a due sole ipotesi, tradizionalmente esse sono dette ipotesi nulla $H_0 = \{\theta = 0\}$ e alternativa $H_1 = \{\theta = 1\}$. Trovate le probabilità "a posteriori"

$$P(H_0|I) \quad \text{e} \quad P(H_1|I) = 1 - P(H_0|I),$$

possiamo chiederci come si specializza il formalismo del rischio bayesiano. In effetti, i possibili errori sono due,

I decidere di comportarsi come se $H_0$ sia falsa quando in realtà è vera (e quindi "rifiutiamo" $H_0$ e la nostra decisione è $H_1$);
II decidere di comportarsi come se $H_0$ sia vera quando in realtà è falsa (e quindi la decisione è $H_0$);

e vanno sotto i nomi rispettivi di errore di prima e seconda specie. A questi errori associamo due costi/perdite diverse (mentre diamo costo zero se facciamo la decisione giusta)

$$\text{costo errore prima specie} := L(H_1, H_0) \quad \text{costo errore seconda specie} := L(H_0, H_1).$$

e di conseguenza i rischi associati sono

$$\text{Risk}_L\,(H_0|I) = L(H_0, H_1)P(H_1|I), \quad \text{Risk}_L\,(H_1|I) = L(H_1, H_0)P(H_0|I).$$

Quindi la decisione bayesiana che minimizza tale rischio sarà $H_0$ se

$$\text{Risk}_L\,(H_0|I) < \text{Risk}_L\,(H_1|I)\,,$$

mentre $H_1$ nel caso opposto.

*Esempio* 8. Supponiamo che una serie di osservazioni mostri che un nuovo farmaco è efficace ($E = H_0$) con probabilità $P(H_0|I) = 7/10$, mentre non lo è con probabilità $P(H_1|I) = 3/10$. Il costo di produrre il farmaco quando questo si rivelasse in realtà inefficace è 100, mentre la perdita che l'azienda farmaceutica ne avrebbe se non lo producesse quando in realtà fosse efficace (ad esempio, se un'azienda competitiva si mettesse a produrlo) è 10, quindi

$$\text{Risk}\,(H_0|I) = 100\frac{3}{10} = 30,$$

mentre

$$\text{Risk}\,(H_1|I) = 10\frac{7}{10} = 7.$$

A questo punto, la decisione che minimizza il rischio è di assumere che il farmaco non è efficace.

*Esercizio* 9. Supponiamo che una serie di osservazioni mostri che un nuovo farmaco è efficace ($E = H_0$) con probabilità $P(H_0|I) = 7/10$, mentre non lo è con probabilità $P(H_1|I) = 3/10$. Introduciamo le seguenti funzioni di costo:

$$L(H_0, H_1) = 200, \quad L(H_0, H_0) = -200, L(H_1, H_0) = 200, L(H_1, H_1) = 50.$$

Quale decisione minimizza il rischio?

*Esercizio* 10. Un sito web di previsioni meteo deve decidere se pubblicare una previsione di bel tempo o cattivo tempo per la settimana successiva. Il modello fisico-matematico prevede bel tempo con probabilità 60%, mentre cattivo tempo con probabilità 40%. È noto d'altra parte che una previsione

mancata di bel tempo comporta un costo (in termini di reputazione) di 10 (in una qualche unità di misura), mentre una previsione mancata di cattivo tempo comporta un costo (in termini di reputazione) di 30, mentre una previsione corretta di bel tempo comporta un costo (guadagno) $-5$, e infine una previsione corretta di cattivo tempo comporta un costo (guadagno) $-20$. Quale decisione minimizza il rischio definito da tale costo?

## 5. Regioni di confidenza bayesiane

Immaginiamo ora di avere aggiornato l'informazione a nostra disposizione (ad esempio sulla base di osservazioni) e di essere giunti alla legge a posteriori(discreta o continua) di $\theta \in \Theta$ rispetto alla informazione $I$. Invece di prendere una decisione, possiamo cercare di stabilire una regione dei parametri $C \subseteq \Theta$ in cui, con probabilità alta, $\theta \in C$. Diamo la seguente definizione:

**Definizione 11.** Dato $\alpha \in (0,1)$, diciamo che $C \subseteq \Theta$ è una regione di confidenza bayesiana per $\theta$ al livello $(1-\alpha)100\%$ se vale

$$P(\theta \in C|I) \geq (1-\alpha).$$

A volte si chiede $= (1-\alpha)$ invece di $\geq$. Ovviamente siamo interessati a trovare una regione di confidenza più *piccola* possibile. Per fare questo, nel caso in cui $\theta$ abbia densità $\varrho(\theta = \cdot|I)$, possiamo cercare il più piccolo $d \in [0,\infty)$ affinché la regione

$$C_d := \{z \in \Theta : \varrho(\theta = z|I) \geq d\}$$

abbia probabilità $\geq (1-\alpha)$.

*Esercizio* 12. Supponendo che $\theta$ abbia legge $\mathcal{E}(\lambda)$, per ogni $\alpha \in (0,1)$ determinare la più piccola regione di confidenza bayesiana per $\theta$ al livello $(1-\alpha)$.

## 6. Statistica Bayesiana con variabili gaussiane

Un caso notevole in cui le densità a posteriori si determinano analiticamente è quello in cui $\theta \in \Theta = \mathbb{R}$ è una variabile gaussiana $\mathcal{N}(\theta_0, s_0^2)$ e, sapendo $\{\theta = z\}$, l'informazione acquisita $I$ è ottenuta dall'osservazione di variabili aleatorie $X_1, \dots X_n$ che sono a loro volta gaussiane, tutte *indipendenti* tra loro, tutte con legge $\mathcal{N}(z, \sigma^2)$, dove $\sigma^2 > 0$ è una costante (ossia un numero positivo noto). Poniamo quindi

$$I = \{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \ldots \cap \{X_n = x_n\}.$$

Per determinare la legge di $\theta$ a posteriori, introduciamo la notazione

$$I_k := \{X_1 = x_1\} \cap \ldots \cap \{X_k = x_k\}$$

e applichiamo ripetutamente la formula di Bayes (caso continuo/continuo), trovando

$$\begin{aligned}
\varrho(\theta = z|I_n) &= \varrho(\theta = z|I_{n-1} \cap \{X_n = x_n\}) \\
&= \varrho(X_n = x_n|I_{n-1} \cap \{\theta = z\}) \frac{\varrho(\theta = z|I_{n-1})}{\varrho(X_n = x_n|I_{n-1})} \\
&= \frac{\varrho(X_n = x_n|\theta = z)}{\varrho(X_n = x_n|I_{n-1})} \varrho(\theta = z|I_{n-1}) \quad \text{per indipendenza delle } X_i, \text{ sapendo } \{\theta = z\}, \\
&= \frac{\varrho(X_n = x_n|\theta = z)}{\varrho(X_n = x_n|I_{n-1})} \frac{\varrho(X_{n-1} = x_{n-1}|\theta = z)}{\varrho(X_{n-1} = x_{n-1}|I_{n-2})} \varrho(\theta = z|I_{n-2}) \\
&= \prod_{i=1}^{n} \frac{\varrho(X_i = x_i|\theta = z)}{\varrho(X_i = x_i|I_{i-1})} \varrho(\theta = z|I_0)
\end{aligned}$$

Osserviamo ora che, ai fini di calcolare la densità di $\theta$ nel punto $z$, i denominatori $\varrho(X_i = x_i|I_{i-1})$ sono delle *costanti* (non dipendono da $z$) e quindi abbiamo trovato che la densità a posteriori è

$$\begin{aligned}
\varrho(\theta = z|I_n) &= c \prod_{i=1}^{n} \varrho(X_i = x_i|\theta = z) \varrho(\theta = z|I_0) \\
&= c' \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\frac{(x_i - z)^2}{\sigma^2}\right) \exp\left(-\frac{1}{2}\frac{(z - \theta_0)^2}{s_0^2}\right) \\
&= c' \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - z)^2 + \frac{(z - \theta_0)^2}{s_0^2}\right)\right)
\end{aligned}$$

Riconosciamo in questa ultima espressione una densità gaussiana (infatti all'esponente abbiamo un polinomio di secondo grado nella variabile $z$). Con qualche calcolo in più ci si riconduce ad una espressione del tipo

$$\varrho(\theta = z|I_n) = c'' \exp\left(-\frac{1}{2}\frac{(z - \theta_n)^2}{s_n^2}\right),$$

quindi a posteriori $\theta$ è $\mathcal{N}(\theta_n, s_n^2)$, dove, posto $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$, i parametri sono dati da

$$\theta_n := \frac{\bar{x} n \sigma^{-2} + \theta_0 \sigma_0^{-2}}{n\sigma^{-2} + \sigma_0^{-2}} \quad \text{e} \quad s_n^2 := \left(n\sigma^{-2} + s_0^{-2}\right)^{-1}.$$

Osserviamo due limiti interessanti. Il primo caso è il limite per $n \to +\infty$: supponendo che $\bar{x}$ converga a qualche valore limite (cosa compatibile con la legge dei grandi numeri), si trova proprio che $\theta_n$ converge a tale valore mentre $s_n$ converge a zero. All'aumentare del numero di osservazioni il grado di fiducia dovrebbe convergere verso una "certezza" sul valor medio del limite.

*Esercizio* 13. Ripetere il ragionamento supponendo stavolta che, sapendo $\{\theta = z\}$ le variabili $X_i$ sono gaussiane con legge $\mathcal{N}(z, \sigma_i^2)$ (le $\sigma_i$ sono costanti, eventualmente diverse). Mostrare che la legge a posteriori di $\theta$ è

$$\varrho(\theta = z|I_n) = c \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{(x_i - z)^2}{\sigma_i^2} + \frac{(z - \theta_0)^2}{s_0^2}\right)\right)$$

e quindi risulta gaussiana $\mathcal{N}(\theta_n, s_n^2)$, dove

$$\frac{\sum_{i=1}^n x_i \sigma_i^{-2} + \theta_0 \sigma_0^{-2}}{\sum_{i=1}^n \sigma_i^{-2} + \sigma_0^{-2}} \quad \text{e} \quad s_n^2 := \left( \sum_{i=1}^n {\sigma_i}^{-2} + s_0^{-2} \right)^{-1}.$$

Se interpretiamo $\sigma_i$ come un parametro di "affidabilità" della osservazione $X_i$ (minore $\sigma_i$, maggiore l'osservazione è affidabile) ne segue che $\theta_n$ è una media delle osservazioni, inclusiva dell'informazione iniziale, dove ciascuna osservazione è "pesata" secondo la propria affidabilità.

*Esempio* 14 (regioni di confidenza gaussiane). Supponiamo che la legge di $\theta \in \Theta = \mathbb{R}$ sia $\mathcal{N}(0,1)$. Fissato $\alpha$, cerchiamo una regione di confidenza bayesiana per $\theta$ al livello $(1-\alpha)$ "più piccola" possibile. Usiamo quindi l'osservazione sopra e cerchiamo una regione del tipo $C_d$ per un qualche $d \in [0, \infty)$. Una semplice osservazione del grafico ci permette di ottenere che $C_d$ è un intervallo del tipo $C_d = [-t, t]$ per qualche $t \geq 0$. Pertanto basterà trovare il più piccolo $t$ affinché valga

$$P(\theta \in [-t, t] | I) \geq (1 - \alpha).$$

In modo equivalente, richiediamo che $P(\theta < -t \quad \text{o} \quad \theta > t | I) \leq \alpha$. Siccome $P(\theta < -t | I) = P(\theta > t | I)$, troviamo la condizione

$$P(\theta < -t | I) \leq \frac{\alpha}{2}.$$

Per trovare $t$ più piccolo possibile, poiché la funzione di ripartizione è crescente, basterà imporre la condizione

$$P(\theta < -t | I) = \frac{\alpha}{2}.$$

Per risolvere questa equazione per $-t$, poiché non esistono "formule" per la funzione di ripartizione gaussiana (né per la funzione inversa, il *quantile gaussiano*) dobbiamo ricorrere ad un calcolatore o alle "tavole della funzione di ripartizione gaussiana". Se vogliamo usare le tavole, poiché queste riportano solamente gli argomenti non-negativi (e quindi i valori $\geq \frac{1}{2}$) per ricondursi al caso di un argomento negativo, visto che cerchiamo $-t \leq 0$, basta usare l'identità

$$P(\theta \leq -t | I) = 1 - P(\theta > -t | I) 1 - P(\theta \leq t | I),$$

da cui l'equazione da risolvere diventa

$$P(\theta \leq t | I) = 1 - \frac{\alpha}{2}.$$

*Esercizio* 15. Calcolare una regione di confidenza bayesiana per $\theta$ al livello $(1-\alpha)$ che sia "più piccola" possibile, sapendo che $\theta$ è $\mathcal{N}(m, \sigma^2)$. *(Suggerimento: scrivere $\theta = \sigma Y + m$ e calcolare prima una regione per $Y$).*

## 7. Metodo dei minimi quadrati

Anche se in linea di principio il metodo bayesiano è applicabile in tutte le situazioni, nella pratica incontra diverse difficoltà:

(1) non è sempre chiaro come stabilire la probabilità rispetto all'informazione *a priori*, e quindi questo è soggetto a scelte spesso criticabili,

(2) a parte per pochi casi (come ad esempio per le gaussiane) i calcoli delle probabilità a posteriori non sono affrontabili in forma analitica, e lo sono diventati numericamente solo recentemente con maggiore potenza dei calcolatori,
(3) spesso si è più interessati ad una *decisione* $\hat{\theta}$ piuttosto che a tutta la probabilità a posteriori, e quindi calcolare tale probabilità (analiticamente o numericamente) non è efficiente.

Per questi motivi tante approssimazioni/semplificazioni del metodo sono state introdotte per diverse esigenze, a volte anche in modo del tutto indipendente dal contesto bayesiano. Uno di questi è il *metodo di minimi quadrati*, inizialmente introdotto dai matematici Laplace, Legendre e Gauss, per lo studio dei dati delle osservazioni astronomiche. Questo metodo poi si è generalizzato (col nome di *regressione*, di cui il caso che vedremo è detto lineare) a tantissimi altri ambiti, e in informatica costituisce un elemento fondamentale del "machine learning".

La situazione generale è la seguente. Il parametro $\theta$ è una funzione, e la nuova informazione è un certo numero di osservazioni $(x_i, y_i)$, ossia $y_i = \theta(x_i)$. Sulla base di queste si vuole "stimare" $\theta$, ossia decidere una determinata $\hat{\theta}$ da usare eventualemente per previsioni, avendo osservato

*Esempio* 16. Nel caso dell'astronomia, $\theta$ potrebbe essere la legge del moto di un pianeta (che è incerta perché dipende da alcune quantità non misurabili direttamente) e $y_i$ la posizione osservata all'istante $x_i$.

*Esempio* 17. Nel caso dell'astronomia, $\theta$ potrebbe essere la legge del moto di un pianeta (che è incerta perché dipende da alcune quantità, ad es. certe costanti fisiche, non misurabili direttamente) e $x_i$ la posizione osservata all'istante $y_i$.

*Esempio* 18. Nel caso del "machine learning", la funzione $\theta$ potrebbe essere la funzione che, dato come argomento (input) un'immagine in cui è presente un carattere scritto a mano, restituisce il codice del carattere (output). In problemi di questo tipo solitamente si hanno a disposizione tanti esempi di input/output da cui "imparare" $\theta$.

Una differenza importante rispetto agli esempi che abbiamo visto nelle sezioni precedenti è che $\theta \in \Theta$ appartiene ad un insieme (quello delle funzioni) potenzialmente molto grande. Nell'esempio precedente, se l'immagine è bianco/nero e consiste di $n$ pixel, e l'output è una delle 26 lettere dell'alfabeto inglese, le possibili $\theta$ sono $26^{2^n}$. L'idea è quindi di ridurre le possibilità per $\theta$ introducendo ipotesi sulla sua struttura. Noi studieremo il caso cosiddetto *lineare*, ma negli ultimi anni strutture più complesse hanno portato a miglioramenti notevoli in molti problemi di machine learning.

Prima di fornire la teoria generale (della regressione lineare), iniziamo con un paio di esempi più semplici.

*Esempio* 19 ($\theta$ costante). Si suppone che la funzione $\theta(x) = y$ sia una *costante*. Se questa ipotesi (drastica) fosse vera e $\theta(x) = z$ per ogni $x$, allora si dovrebbero osservare $y_i = z$ tutte uguali (indipendentemente dalle $x_i$). In realtà, se pensiamo a questa come una prima approssimazione della vera $\theta$, possiamo introdurre, per ogni osservazione $(x_i, y_i)$, l'errore (se fosse vera

l'ipotesi $\theta = z$), detto in questo ambito *residuo*,

$$e_i := y_i - \theta(x_i) = y_i - z.$$

Il metodo dei minimi quadrati consiste allora nello scegliere $\hat{\theta} = z$ che *minimizza l'errore quadratico* delle $n$ osservazioni, definito come

$$Q(z) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - z)^2.$$

Per determinare $\hat{\theta}$ in questo esempio, possiamo procedere in due modi (equivalenti). Il primo è di derivare rispetto a $z$ la funzione $Q(z)$ e porre $Q'(z) = 0$. Il secondo consiste nell'osservare che

$$\frac{1}{n} Q(z) = \sum_{i=1}^{n} (y_i - z)^2 \frac{1}{n} = \mathbb{E}\left[(Y - z)^2 | Y \text{ è uniforme su } \{y_i, \dots, y_n\}\right]$$

e quindi la "decisione" $z = \hat{\theta}$ che minimizza è

$$\hat{\theta} := \mathbb{E}\left[Y | Y \text{ è uniforme su } \{y_i, \dots, y_n\}\right] = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

*Esempio* 20 ($\theta$ lineare). Si suppone che la funzione $\theta(x) = y$ sia una funzione *lineare*, ossia $\theta(x) = ax$ per qualche parametro $a \in \mathbb{R}$ da stimare. Come prima, se fosse vera l'ipotesi $\theta(x) = ax$, basterebbe prendere una qualunque coppia $(x, y)$ con $x \neq 0$ e calcolare $a = y/x$, ma questo non accade mai in realtà. Quindi, introduciamo per ogni osservazione $(x_i, y_i)$, il residuo

$$e_i := y_i - \theta(x_i) = y_i - a x_i,$$

e determiniamo $\hat{\theta}$ (ossia $\hat{a}$) che minimizza l'errore quadratico

$$Q(a) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a x_i)^2.$$

Differenziando e ponendo $Q'(\hat{a}) = 0$, troviamo la condizione

$$\sum_{i=1}^{n} (y_i - \hat{a} x_i) x_i = 0,$$

da cui

$$\hat{a} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \sum_{i=1}^{n} \frac{y_i}{x_i} p_i$$

dove abbiamo introdotto $p_i := x_i^2 / \sum_{i=1}^{n} x_i^2$. Notiamo che le $p_i$ possono essere pensate come delle probabilità (sono non negative e sommano ad 1) e $\hat{a}$ diventa un "valore atteso" di $Y/X$ .

Enunciamo una versione generale del *metodo dei minimi quadrati*. Supponiamo di avere una funzione $\theta(x) = y$ (che dipende da qualche parametro non noto) e di osservare $(x_i, y_i)$. Si introducono i residui

$$e_i := y_i - \theta(x_i)$$

e si determina $\hat{\theta}$ in modo tale da minimizzare l'errore quadratico

$$Q(\theta) := \sum_{i=1}^{n} e_i^2 = (y_i - \theta(x_i))^2 .$$

Un caso molto importante in cui i calcoli sono espliciti e semplici è quello *lineare*.

**Definizione 21** (regressione lineare)**.** Data $\theta : \mathbb{R}^k \to \mathbb{R}$, si dice che il modello $\theta(x) = \theta(x^{(1)}, \dots, x^{(k)}) = y$ è lineare nei $k$ parametri $(\theta_1, \theta_2, \dots, \theta_k) \in \mathbb{R}^k$ se è della forma

$$\theta(x) = x \cdot \theta = \sum_{j=1}^{k} x^{(j)} \theta_j .$$

In tal caso i residui sono dati da

$$e_i := y_i - \sum_{j=1}^{k} x_i^{(j)} \theta_j .$$

Per ottenere in modo agevole $\hat{\theta}$, introduciamo la matrice $X := (x_i^{(j)})_{i=1\dots,n}^{j=1,\dots,k} \in \mathbb{R}^{n\times k}$, i vettori $Y := (y_i)_{i=1}^n \in \mathbb{R}^n$, e $v := (\theta_j)_{j=1}^k \in \mathbb{R}^k$. Si ha allora

$$Q(v) := |Y - Xv|^2 \quad \text{dove } |\cdot|^2 \text{ indica la lunghezza del vettore.}$$

Si trova che $\hat{v}$ che minimizza $Q$ è dato dal vettore

$$\hat{v} := (X^t X)^{-1} X^t Y,$$

dove $X^t$ indica la matrice trasposta di $X$ (e si suppone che la matrice quadrata $X^t X$ sia invertibile).

*Esercizio* 22. Verificare che la formula sopra permette di ottenere i due esempi ($\theta$ costante, $\theta$ lineare) dati sopra.

*Osservazione* 23 (funzioni non lineari). Anche se a prima vista sembrerebbe che il modello lineare possa solamente rappresentare funzioni lineari, in realtà notiamo che la linearità è richiesta in *nei parametri* da cui dipende $\theta$. In pratica è possibile trattare anche polinomi nella variabile $x$, con il seguente stratagemma: supponiamo ad esempio che $\theta$ sia un polinomio di secondo grado,

$$\theta(x) = a_0 + a_1 x + a_2 x^2,$$

dove $a_0$, $a_1$, $a_2 \in \mathbb{R}$ devono essere stimati sulla base di osservazioni $(x_i, y_i)$. Allora possiamo introdurre un modello lineare ponendo

$$\theta^{lin}(x^{(0)}, x^{(1)}, x^{(2)}) := a_0 x^{(0)} + a_1 x^{(1)} + a_2 x^{(2)} \quad \text{per } (x^{(0)}, x^{(1)}, x^{(2)} \in \mathbb{R}^3.$$

In questo modo si ottiene $\theta(x) = \theta^{lin}(1, x, x^2)$. La matrice $X$ di osservazioni è quindi formata dalle $n$ righe $(1, x_i, x_i^2)$, per $i \in \{1, \dots, n\}$.

*Esempio* 24 ($\theta$ affine). Consideriamo il caso $\theta(x) = a_0 + a_1 x$. Si trova allora

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}$$

$$\hat{a}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^{n} \frac{y_i - \bar{y}}{x_i - \bar{x}} p_i,$$

dove $p_i := (x_i - \bar{x}) / \left(\sum_{i=1}^n (x_i - \bar{x})^2\right)$ e $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$, $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$.

*Osservazione* 25. C'è un collegamento tra il metodo dei minimi quadrati e il metodo bayesiano? In effetti, si potrebbe dimostrare in generale che il metodo dei minimi quadrati nel caso della *regressione lineare* può essere intepretato introducendo opportune variabili aleatorie gaussiane e in un opportuno limite per l'informazione iniziale. Non avendo gli strumenti per dimostrare l'equivalenza in generale, osserviamolo nel caso particolare di una funzione $\hat{\theta}(x) = \theta$ costante. Abbiamo visto nella sezione precedente che, se supponiamo $\theta$ una v.a. $\mathcal{N}(\theta_0, s_0^2)$ e sapendo $\theta = z$ le variabili $Y$ sono indipendenti tutte $\mathcal{N}(z, \sigma^2)$, allora a posteriori (avendo osservato $Y = y_i$) la variabile $\theta$ è gaussiana, con densità

$$\varrho(\theta = z) = c \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - z)^2 + \frac{(z-\theta_0)^2}{s_0^2}\right)\right).$$

Riconosciamo all'esponenziale il termine $Q(z) = \sum_{i=1}^{n}(y_i - z)^2$ che minimizziamo nel metodo dei minimi quadrati. Nel limite $s_0 \to \infty$, otteniamo

$$\varrho(\theta = z) = exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - z)^2\right)\right)$$

e quindi minimizzare l'errore quadratico è come massimizzare la densità a posteriori! Questa analogia si potrebbe portare avanti nel metodo generale. D'altra parte, ci suggerisce anche altre funzioni da minimizzare. Per esempio, se sappiamo che $\theta$ dovrebbe essere vicina a un valore $\theta_0$, basterà aggiungere all'errore quadratico un termine del tipo $(\theta - \theta_0)^2 s_0^{-2}$, e minimizzare

$$Q(\theta) + (\theta - \theta_0)^2 s_0^{-2}.$$

*E-mail address*, D. Trevisan: dario.trevisan@unipi.it