Lezione 3: Richiami di statistica

Nella lezione precedente abbiamo assunto che i parametri $\vec{\vartheta}$ fossero tutti noti. Questa lezione discute come imparare $\vec{\vartheta}$ dai dati $\mathcal{D}$. Come delineato nella prima lezione, lo scopo del gioco si riduce a

$$\boxed{\hat{\vartheta} = \underset{\vec{\vartheta}}{\arg\min}\, \mathcal{L}(\vec{\vartheta})}$$

3.1 MAXIMUM LIKELIHOOD ESTIMATION (MLE)

Definiamo la MLE come $\hat{\vartheta}_{mle} \triangleq \underset{\vec{\vartheta}}{\arg\max}\, p(\mathcal{D}/\vec{\vartheta})$. Se i dati sono campioni indipendenti della stessa distribuzione, allora

$$p(\mathcal{D}/\vec{\vartheta}) = \prod_{n=1}^{N} p(\vec{y}_n/\vec{x}_n, \vec{\vartheta})$$

Definiamo la log-likelihood come (negativa $\equiv$ NLL)

$$\ell(\vec{\vartheta}) \triangleq -\log p(\mathcal{D}/\vec{\vartheta}) = -\sum_{n=1}^{N} \log p(\vec{y}_n|\vec{x}_n, \vec{\vartheta})$$

La MLE è data da

$$\hat{\vartheta}_{mle} = \underset{\vec{\vartheta}}{\arg\min} -\sum_{n=1}^{N} \log p(\vec{y}_n/\vec{x}_n, \vec{\vartheta}) \qquad \text{(supervised learning)}$$

$$\hat{\vartheta}_{mle} = \underset{\vec{\vartheta}}{\arg\min} -\sum_{n=1}^{N} \log p(\vec{y}_n/\vec{\vartheta}) \qquad \text{(unsupervised learning)}$$

La MLE può essere giustificata pensandola come approssimazione del "posterior" Bayesiano, dato un "prior" uniforme. Per esempio:

$p(\vec{\vartheta}/\mathcal{D}) = \delta(\vec{\vartheta} - \hat{\vartheta}_{MAP})$ con $\hat{\vartheta}_{MAP}$ il "posterior"

$$\hat{\vartheta}_{MAP} = \underset{\vec{\vartheta}}{\arg\min} -\log p(\vec{\vartheta}/\mathcal{D}) = \underset{\vec{\vartheta}}{\arg\min} -\log p(\mathcal{D}/\vec{\vartheta}) - \log p(\vec{\vartheta}) \overset{p(\vec{\vartheta})=1}{=} \hat{\vartheta}_{MLE}$$

Supponiamo che Y sia una variabile casuale distribuita normalmente $Y \sim \mathcal{N}(\mu, \sigma^2)$ e sia $\mathcal{D}$ un dataset i cui punti sono campionati indipendentemente: $\mathcal{D} = \{y_n : n = 1:N\}$. Allora

$$NLL(\mu, \sigma^2) = -\sum_{n=1}^{N_D} \ln\left[\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}(y_n-\mu)^2\right)\right]$$

$$= \frac{1}{2\sigma^2}\sum_{n=1}^{N_D}(y_n-\mu)^2 + \frac{N_D}{2}\ln(2\pi\sigma^2)$$

$$\frac{\partial}{\partial\mu} NLL(\mu, \sigma^2) = 0 \iff \hat{\mu}_{MLE} = \frac{1}{N_D}\sum_{n=1}^{N_D} y_n = \bar{y}$$

$$\frac{\partial}{\partial\sigma^2} NLL(\mu, \sigma^2) = 0 \iff \hat{\sigma}^2_{MLE} = \frac{1}{N_D}\sum_{n=1}^{N_D}(y_n - \hat{\mu}_{MLE})^2 = \underbrace{\frac{1}{N}\sum_{n=1}^{N} y_n^2}_{s^2} - \bar{y}^2$$

Per una distribuzione multivariata:

$$\ell(\vec{\mu}, \vec{\Sigma}) = \ln p(\mathcal{D}|\vec{\mu}, \vec{\Sigma}) = \frac{N_D}{2}\ln|\vec{\Lambda}| = -\frac{1}{2}\sum_{n=1}^{N_D}(\vec{y}_n - \vec{\mu})^T \Lambda (\vec{y}_n - \vec{\mu})$$

con $\Lambda = \Sigma^{-1}$ la PRECISION MATRIX. Segue (come sopra)

$$\hat{\mu} = \frac{1}{N_D}\sum_{n=1}^{N_D}\vec{y}_n = \bar{y} \qquad \hat{\Sigma} = \frac{1}{N_D}\sum_{n=1}^{N_D}(\vec{y}_n - \bar{y})(\vec{y}_n - \bar{y})^T$$

empirical mean — empirical covariance matrix

3.2 EMPIRICAL RISK MINIMISATION (ERM)

La MLE può essere generalizzata sostituendo la loss function logaritmica con qualunque altra funzione

$$\mathcal{L}(\vec{\vartheta}) = \frac{1}{N}\sum_{n=1}^{N}\ell(\vec{y}_n, \vec{\vartheta}; \vec{x}_n)$$

Ciò è noto come EMPIRICAL RISK MINIMISATION (ERM) dal momento che è il loss atteso quando l'aspettazione è presa rispetto alla distribuzione empirica. Per esempio, in un problema di classificazione

$$\ell_{01}(\vec{y}_n, \vec{\vartheta}; \vec{x}_n) = \begin{cases} 0 & \text{se } \vec{y}_n = f(\vec{x}_n, \vec{\vartheta}) \\ 1 & \text{se } \vec{y}_n \neq f(\vec{x}_n, \vec{\vartheta}) \end{cases}$$

con $f(\vec{x}, \vec{\vartheta})$ un predittore. L'empirical risk diventa

$$\mathcal{L}(\vec{\vartheta}) = \frac{1}{N} \sum_{n=1}^{N} l_{01}(\vec{y}_n, \vec{\vartheta}; \vec{x}_n)$$

che è il misclassification rate (nel training set). Per problemi binari possiamo riscrivere il misclassification rate nella forma seguente

$\tilde{y} \in \{-1; +1\}$ true label $\quad \hat{y} \in \{-1, +1\} = f(\vec{x}, \vec{\vartheta})$ prediction

$$l_{01}(\tilde{y}, \hat{y}) = \mathbb{1}(\tilde{y} \neq \hat{y}) = \mathbb{1}(\tilde{y}\hat{y} < 0)$$

Il rischio empirico diventa

$$\mathcal{L}(\vec{\vartheta}) = \frac{1}{N} \sum_{n=1}^{N} l_{01}(y_n, \hat{y}_n) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}(\tilde{y}_n, \hat{y}_n < 0)$$

La funzione l_{01} è non-smooth, il che la rende difficile da ottimizzare. In luogo di essa, si può utilizzare una funzione surrogate (definita generalmente come limite superiore convesso. Per esempio

$$l_{ll}(\tilde{y}, \eta) = -\ln p(\tilde{y}/\eta) = \ln(1 + e^{-\tilde{y}\eta}) \qquad p(\tilde{y}/\vec{x}, \vec{\vartheta}) = \sigma(\tilde{y}\eta) = \frac{1}{1+e^{-\tilde{y}\eta}}$$

Altri esempi sono l'hinge loss e l'exp loss.

.3 ALTRI METODI DI MINIMIZZAZIONE

.3.1 METODO DEI MOMENTI

Talora il calcolo di $\nabla_{\vec{\vartheta}} NLL(\vec{\vartheta}) = 0$ è difficile. Il metodo dei momenti consiste nell'eguagliare i momenti teorici della distribuzione ai momenti empirici.

MOMENTI TEORICI: $\mu_k = \mathbb{E}[Y^k]$

MOMENTI EMPIRICI $\hat{\mu}_k = \frac{1}{N} \sum_{n=1}^{N} y_n^k$

$\mu_k = \hat{\mu}_k$ ho un set di k equazioni

Per la gaussiana:

$$\mu_1 = \mu = \bar{y}$$

$$\mu_2 = \sigma^2 + \mu^2 = s^2$$

in questo caso MOM ≡ MLE. Non sempre è così (p. es. distr. uniforme)

3.3.2. ONLINE RECURSIVE ESTIMATION

Se tutto il dataset $\mathcal{D}$ è noto e disponibile prima che il learning venga inizializzato, si dice che si fa "batch learning". In alcuni casi, però, il dataset è disponibile a blocchi.

$\hat{\vartheta}_{t-1}$ è la predizione dato $\mathcal{D}_{1:t-1}$

dobbiamo trovare $\vec{\vartheta}_t = f(\hat{\vec{\vartheta}}_{t-1}, \vec{y}_t)$ con un update ricorsivo

Per una gaussiana univariata

$$\hat{\mu}_t = \frac{1}{t}\sum_{n=1}^{t} \vec{y}_n \longrightarrow \hat{\mu}_t = \frac{1}{t}\sum_{n=1}^{t} \vec{y}_n = \frac{1}{t}\left((t-1)\hat{\mu}_{t-1} + y_t\right)$$

$$= \hat{\mu}_{t-1} + \frac{1}{t}\left(\vec{y}_t - \hat{\mu}_{t-1}\right) \quad \text{MOVING AVERAGE}$$

La moving average può eventualmente essere pesata se la distribuzione cambia cont.

3.4 REGOLARIZZAZIONE

Un problema della MLE e della ERM è che i parametri tenderanno ad essere determinati minimizzando il loss nel training set, ma non necessariamente nei dati futuri. Il problema si corregge con la regolarizzazione

$$\mathcal{L}(\vec{\vartheta}, \lambda) = \left[\frac{1}{N}\sum_{n=1}^{N} \ell(\vec{y}_n, \vec{\vartheta}; \vec{x}_n) + \lambda C(\vec{\vartheta})\right]$$

regularisation parameter

penalty function p. es. $C(\vec{\vartheta}) = -\ln p(\vec{\vartheta})$

Per $\lambda = 1$ e riscalando $p(\vec{\vartheta})$, si ha la NLL

$$NLL(\vec{\vartheta}, \lambda) = -\left[\sum_{n=1}^{N} \ln p(\vec{y}_n / \vec{x}_n, \vec{\vartheta}) + \ln p(\vec{\vartheta})\right] = -\left[\ln p(\mathcal{D}/\vec{\vartheta}) + \ln p(\vec{\vartheta})\right]$$

che implica

$$\hat{\vartheta} = \underset{\vec{\vartheta}}{\operatorname{argmax}} \ln p(\vec{\vartheta}/\mathcal{D}) = \underset{\vec{\vartheta}}{\operatorname{argmax}} \left[\ln p(\mathcal{D}/\vec{\vartheta}) + \ln p(\vec{\vartheta}) - \text{cost}\right]$$

Questa si chiama MAP (MAXIMUM A POSTERIOR) estimation.

Come scegliere il valore di λ? Un valore (troppo) piccolo significa minimizzare il rischio empirico; un valore (troppo) grande significa essere troppo vicini al prior (overfitting vs underfitting).
Dividiamo il dataset in due classi: training e validation (80/20%).
Si fitta il modello su $\mathcal{D}_{train}$ $\forall$ setting λ e si valuta le performance $\mathcal{D}_{valid}$. Si prende quindi il valore di λ associato alla performance migliore. Definiamo l'empirical risk regolarizzato

$$R_\lambda(\vec{\vartheta}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\vec{x}, \vec{y}) \in \mathcal{D}} \ell(\vec{y}, f(\vec{x}, \vec{\vartheta})) + \lambda C(\vec{\vartheta})$$

$$\hat{\vartheta}_\lambda(\mathcal{D}_{train}) = \arg\min_{\vec{\vartheta}} R_\lambda(\vec{\vartheta}, \mathcal{D}_{train}) \quad \forall \lambda$$

$$R_\lambda^{val} \triangleq R_0(\hat{\vartheta}_\lambda(\mathcal{D}_{train}), \mathcal{D}_{valid})$$ si calcola il VALIDATION RISK

$$\lambda^* = \arg\min_{\lambda \in S} R_\lambda^{val}$$ Dopo aver preso λ^* si rifitta il modello al $\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{val}$

$$\hat{\vartheta}^* = \arg\min_{\vec{\vartheta}} R_{\lambda^*}(\vec{\vartheta}, \mathcal{D})$$

Se il dataset è piccolo, eliminare il 20% dei dati dal training può essere dannoso. Si usa allora il CROSS-VALIDATION in cui il dataset di training è diviso in K folds. Per ogni fold, il modello viene allenato su tutti i folds, tranne il K-esimo, che viene usato come test fold.

$$R_\lambda^{cv} \triangleq \frac{1}{K} \sum_{k=1}^{K} R_0(\hat{\vec{\vartheta}}_\lambda(\mathcal{D}_{-k}), \mathcal{D}_k)$$

.6 STATISTICA BAYESIANA

Abbiamo discusso vari metodi per determinare i parametri dai dati ma non abbiamo detto nulla sulle loro incertezze. Il TB ci aiuta

$$p(\vec{\vartheta}/\mathcal{D}) = \frac{p(\vec{\vartheta})\, p(\mathcal{D}/\vec{\vartheta})}{p(\mathcal{D})} = \frac{p(\vec{\vartheta})\, p(\mathcal{D}/\vec{\vartheta})}{\int p(\vec{\vartheta}')\, p(\mathcal{D}/\vec{\vartheta}')\, d\vec{\vartheta}'}$$

$\mathcal{D} = \{(\vec{x}_n, \vec{y}_n)\}$ SUPERVISED LEARNING

$n = 1, \dots, N$

$\mathcal{D} = \{(\vec{y}_n)\}$ UNSUPERVISED LEARNING

Una volta determinato il posteriore sui parametri, possiamo calcolare il posteriore della distribuzione predittiva marginalizzando su $\vec{\vartheta}$

$$p(\vec{y}/\vec{x}, \mathcal{D}) = \int p(\vec{y}/x, \vec{\vartheta})\, p(\vec{\vartheta}/\mathcal{D})\, d\vec{\vartheta}$$ Bayes model averaging (BMA)

Consideriamo una distribuzione gaussiana di cui sia nota la varianza. Nel caso univariato, la likelihood per μ ha la forma

$$p(\mathcal{D}/\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N_{\mathcal{D}}}(y_n-\mu)^2\right\}$$

Uno può mostrare che il prior coniugato è un'altra gaussiana $\mathcal{N}(\mu/\breve{m}, \breve{\tau}^2)$. Usando il teorema di Bayes, troviamo che il posteriore è

$$p(\mu/\mathcal{D}, \sigma^2) = \mathcal{N}(\mu/\hat{m}, \hat{\tau}^2)$$

$$\hat{\tau}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\breve{\tau}^2}} = \frac{\sigma^2\breve{\tau}^2}{N\breve{\tau}^2 + \sigma^2} \qquad \hat{m} = \hat{\tau}^2\left(\frac{\breve{m}}{\breve{\tau}^2} + \frac{N\bar{y}}{\sigma^2}\right) = \frac{\sigma^2}{N\breve{\tau}^2+\sigma^2}\breve{m} + \frac{N\breve{\tau}^2}{N\breve{\tau}^2+\sigma^2}\bar{y}$$

con $\bar{y} \triangleq \frac{1}{N}\sum_{n=1}^{N} y_n$ è la media empirica.

Definendo $\kappa = 1/\sigma^2$ e $\breve{\lambda} = 1/\breve{\tau}^2$ ottengo

$$\hat{\lambda} = \breve{\lambda} + N\kappa$$

$$\hat{m} = \frac{N\kappa\bar{y} + \breve{\lambda}\breve{m}}{\hat{\lambda}} = \frac{N\kappa}{N\kappa+\breve{\lambda}}\bar{y} + \frac{\breve{\lambda}}{N\kappa+\breve{\lambda}}\breve{m}$$

La precisione del "posterior" è la precisione del "prior" più N unità di misura di precisione κ. La media del "posterior" è una combinazione convessa della media empirica $\bar{y}$ e della media del prior $\breve{m}$. Consideriamo ora il posteriore dopo aver visto un singolo data point y (quindi $N=1$). Allora

$$\hat{m} = \underbrace{\frac{\breve{\lambda}}{\hat{\lambda}}\breve{m} + \frac{K}{\hat{\lambda}}\bar{y}}_{\text{convex combination of prior and data}} = \underbrace{\breve{m} + \frac{K}{\hat{\lambda}}(\bar{y}-\breve{m})}_{\text{prior mean adjusted to data}} = \underbrace{\bar{y} - \frac{\breve{\lambda}}{\hat{\lambda}}(\bar{y}-\breve{m})}_{\text{data adjusted to the prior mean}}$$

POSTERIOR MEAN

$$se(\mu) = \sqrt{\mathbb{V}[\mu|\mathcal{D}]}$$

STANDARD ERROR

Se uso un "uninformative prior" per μ ponendo $\breve{\lambda}=0$, allora $\hat{m}=\bar{y}$.
Supponendo di approssimare $\sigma^2 \sim s^2 \triangleq \frac{1}{N}\sum_{n=1}^{N}(y_n-\bar{y})^2$ segue che
$\hat{\lambda} = NK = N/s^2 \Rightarrow se(\mu) = \frac{1}{\sqrt{\hat{\lambda}}} = \frac{s}{\sqrt{N}}$.

L'incertezza su μ si riduce ad un rateo di $1/\sqrt{N}$.

Osservazioni:

1) Quando non abbiamo informazioni sul prior, è desiderabile usare un prior "uninformative". Per esempio un flat prior $p(\mu)=1$.
2) Qualunque modello Bayesiano richiede che venga specificato un prior $p(\vec{\vartheta})$ per i parametri. I parametri del prior vengono chiamati iperparametri (e sono denotati con Φ). Se non sono noti, possiamo metterci sopra un prior (multi-level model $\Phi \to \vartheta \to \mathcal{D}$)

La joint distribution è $p(\vec{\Phi}, \vec{\vartheta}, \mathcal{D}) = p(\Phi)\,p(\vec{\vartheta}|\Phi)\,p(\mathcal{D}|\vec{\vartheta})$.

METHOD	DEFINITION
MLE	$\hat{\vartheta} = \operatorname{argmax}_{\vartheta}\, p(\mathcal{D}\|\vec{\vartheta})$
MAP	$\hat{\vartheta} = \operatorname{argmax}_{\vartheta}\, p(\mathcal{D}\|\vec{\vartheta})\,p(\vartheta\|\Phi)$
FULL BAYES	$p(\vec{\vartheta}, \Phi\|\mathcal{D}) \propto p(\mathcal{D}\|\vec{\vartheta})\,p(\vec{\vartheta}\|\Phi)\,p(\Phi)$

3.7. INTERVALLI CREDIBILI / INTERVALLI DI CONFIDENZA

Una distribuzione posteriore è un oggetto multidimensionale difficile da visualizzare e da trattare. Può essere utile quindi calcolare stimatori puntuali (come il posterior mean e mode) e calcolare un intervallo di credibilità che quantifica l'incertezza associata a quelle stime. L'intervallo di credibilità $100(1-\alpha)\%$ è la regione $C=(l,u)$ che contiene $1-\alpha$ della probabilità posteriore

$$C_\alpha(\mathcal{D}) = (l,u) : P(l \leq \vartheta \leq u / \mathcal{D}) = 1-\alpha$$

l : lower
u : upper

Se il "posterior" ha una PDF nota, allora $l = F^{-1}(\alpha/2)$ e $u = F^{-1}(1-\alpha/2)$ dove F è la CDF del "posterior". Esempio: gaussiana.

Osservazione

L'intervallo di credibilità è un concetto Bayesiano. L'intervallo di confidenza è un concetto frequentista. Si definisce intervallo di confidenza $100(1-\alpha)\%$, per la stima di un parametro ϑ, l'intervallo $I(\tilde{\mathcal{D}}) = (l(\tilde{\mathcal{D}}), u(\tilde{\mathcal{D}}))$ ottenuto da un dataset $\tilde{\mathcal{D}}$ tale che

$$\Pr(\vartheta \in I(\tilde{\mathcal{D}}) / \tilde{\mathcal{D}} \sim \vartheta) = 1-\alpha$$

Un CL al 95% per un parametro ϑ non significa che il parametro stia verosimilmente dentro l'intervallo il 95% delle volte dati i dati osservati. CL (frequentista): ϑ è una costante fissa non nota e i dati sono aleatori; CI (Bayesiano): i dati sono fissi perché noti, mentre ϑ è ignota.

Esempio

Supponiamo di generare due interi $\mathcal{D} = (y_1, y_2)$ da

$$p(y/\vartheta) = \begin{cases} 0,5 & \text{se } y=\vartheta \\ 0,5 & \text{se } y=\vartheta+1 \\ 0 & \text{altrimenti} \end{cases}$$

Se $\vartheta = 39$, ci aspettiamo i seguenti risultati:

$$(39,39),\ (39,40),\ (40,39),\ (40,40)$$

Sia $m = \min(y_1, y_2)$ e l'intervallo $[\ell(\mathcal{D}), u(\mathcal{D})] = [m, m]$. Segue

$$[39,39],\ [39,39],\ [39,39],\ [40,40] \Rightarrow 75\%\ CI \text{ for } \vartheta = 39.$$

Ma se osserviamo $\mathcal{D} = (39,40)$, allora $p(\theta = 39 / \mathcal{D}) = 1$ (ma CI è solo 75%)

Il CI fallisce per esperimenti NON ripetibili.

3.8 BIAS-VARIANCE TRADEOFF

Sia $\hat{\vartheta}$ lo stimatore statistico e $\hat{\vartheta}(\mathcal{D})$ l'estimando. Nel formalismo frequentista, i dati sono variabili casuali, campionati da una distribuzione $p^*(\mathcal{D})$ che induce una distribuzione sull'estimando $p^*(\hat{\vartheta}(\mathcal{D}))$

Il BIAS di uno stimatore è definito come

$$\text{bias}(\hat{\vartheta}(\cdot)) \triangleq \mathbb{E}\left[\hat{\vartheta}(\mathcal{D})\right] - \vartheta^* \quad (\vartheta^* \text{ è il valore vero})$$

Se il BIAS è nullo, lo stimatore viene detto UNBIASED. Per una gaussiana

$$\text{bias}(\hat{\mu}) = \mathbb{E}[\bar{x}] - \mu = \mathbb{E}\left[\frac{1}{N_D}\sum_{n=1}^{N_D} x_n\right] - \mu = \frac{N_D \mu}{N_D} - \mu = 0 \quad \text{UNBIASED}$$

$$\text{bias}(\hat{\sigma}^2) = \mathbb{E}[s^2] - \sigma^2 = \frac{N_D - 1}{N_D}\sigma^2 - \sigma^2 \neq 0 \quad \text{BIASED}$$

$$\Rightarrow \hat{\sigma}^2 \rightarrow \hat{\sigma}^2_{unb} = \frac{1}{N_D - 1}\sum_{n=1}^{N_D}(x_n - \bar{x})^2 = \frac{N_D}{N_D - 1}\sigma^2_{MLE}$$

La VARIANZA di uno stimatore è

$$\mathbb{V}[\hat{\vartheta}] \triangleq \mathbb{E}[\hat{\vartheta}^2] - (\mathbb{E}[\hat{\vartheta}])^2$$

Idealmente vogliamo che $\mathbb{V}$ sia minima $\forall$ stimatore. Il teorema di Cramer-Rao fornisce un limite inferiore alla varianza per uno stimatore UNBIASED ϑ^*:

$$\mathbb{V}[\hat{\vartheta}] \geq \frac{1}{NF(\vartheta^*)}$$ dove $F(\vartheta^*)$ è la Fisher information matrix

Si può dimostrare che la MLE raggiunge il bound di Cramer-Rao.

Lo scopo dell'inferenza statistica è minimizzare l'errore quadratico medio (MSE). Sia $\hat{\vartheta} = \hat{\vartheta}(\mathcal{D})$ l'estimate e $\bar{\vartheta} = \mathbb{E}[\hat{\vartheta}]$ il valore di aspettazione corrispondente (tutti da $p(\mathcal{D}/\vartheta^*)$). Allora

$$\mathbb{E}\left[(\hat{\vartheta}-\vartheta^*)^2\right] = \mathbb{E}\left[\left[(\hat{\vartheta}-\bar{\vartheta})+(\bar{\vartheta}-\vartheta^*)\right]^2\right]$$

$$= \mathbb{E}\left[(\hat{\vartheta}-\bar{\vartheta})^2\right] + 2(\bar{\vartheta}-\vartheta^*)\mathbb{E}\left[\hat{\vartheta}-\bar{\vartheta}\right] + (\bar{\vartheta}-\vartheta^*)^2$$

$$= \mathbb{E}\left[(\hat{\vartheta}-\bar{\vartheta})^2\right] + (\bar{\vartheta}-\vartheta^*)^2 = \mathbb{V}[\hat{\vartheta}] + \text{bias}^2(\hat{\vartheta})$$

cioè MSE = variance + bias^2. BIAS-VARIANCE TRADEOFF

Esempio

Supponiamo di stimare la media di una distribuzione gaussiana da $\vec{x} = (x_1, \dots, x_N)$ assumendo che i dati siano campionati da $x_n \sim \mathcal{N}(\vartheta^* = 1, \sigma^2)$. Uno stimatore ovvio è dato dalla MLE che ha bias 0 e varianza $\mathbb{V}[\bar{x}/\vartheta^*] = \frac{\sigma^2}{N}$. Ma possiamo dare uno stimatore MAP sotto un prior gaussiano. In tal caso

$$\tilde{x} = \frac{N}{N+K_0}\bar{x} + \frac{K_0}{N+K_0}\vartheta_0 = w\bar{x} + (1-w)\vartheta_0$$ con un prior $\mathcal{N}\left(\vartheta_0, \frac{\sigma^2}{K_0}\right)$

$$\mathbb{E}_{MAP}[\tilde{x}] - \vartheta^* = w\vartheta^* + (1-w)\vartheta_0 - \vartheta^* = (1-w)(\vartheta_0 - \vartheta^*)$$

$$\mathbb{V}_{MAP}[\tilde{x}] = w^2\frac{\sigma^2}{N} < \mathbb{V}_{MLE}[\tilde{x}]$$

BIASED

$0 < w < 1$