Note del corso
Machine Learning per la fisica applicata e fisica delle alte energie

Raviola Alessio

10 novembre 2022

1 Introduzione

> A computer program is said to learn from experience E with respecto to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

L'impostazione del corso è di tipo *probabilistico* (statistical learning). Le quantità non note sono trattate come **variabili aleatorie** (random variables) a cui viene associata una **distribuzione di probabilità** (probability distribution) che descrive il set (pesato) di valori che la variabile può assumere.

Abbiamo tre tipi di machine learning:

- supervised learning;
- unsupervised learning;
- reinforcement learning;

il corso si focalizza sui primi due tipi.

1.1 Supervised learning

Il **compito** T consiste nell'imparare una mappa f dagli input $x \in X$ agli output $y \in Y$. Gli **input** x sono chiamati features (o covariates o predictors) e sono in genere costituiti da un vettore reale con dimensione fissata, ovvero abbiamo $X \equiv \mathbb{R}^D$. Gli **output** sono chiamati label (o target o response).

L'**esperienza** E consiste in un training set $\mathcal{D}$ di N coppie input-output:

$$\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}, \tag{1}$$

dove N è detta sample size.

La **performance** dipenda dal compito T.

1.1.1 Classificazione

Problemi comuni in machine learning sono quelli di **classificazione**. In un problema di questo tipo lo spazio degli output C è un set *non ordinato* di label $y = \{1, 2, \dots, C\}$ dette classes. Quello che chiede il problema è di predire una classe dato un input, problemi di questo tipo sono detti di pattern recognition[1].

Esempio 1.1 (Classificazione specie di iris)**.** In generale in image classification gli input X sono immagini, quindi:

$$X = R^D, \quad D = C \times D_1 \times D_2, \tag{2}$$

[1] Se abbiamo solo due classi, i.e. solo due output, allora il problema si dice di classificazione binaria

Index	sl [cm]	sw [cm]	pl [cm]	pw [cm]	Label
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
⋮	⋮	⋮	⋮	⋮	⋮
50	7.0	3.2	4.7	1.4	Versicolor
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3.0	5.1	1.8	Virginica

Tabella 1: Design matrix del training set per classificazione specie di iris.

ove $C = 3$ sono i canali RGB. E cerchiamo una mappa

$$f : X \longrightarrow Y \tag{3}$$

che ci dica a quale delle classi appartenenti a Y l'immagine appartiene. Per le specie di iris però i botanisti hanno individuato 4 caratteristiche numeriche: lunghezza e larghezza del sepalo e del petalo; dunque abbiamo $X = \mathbb{R}^4$. Supponiamo che il traning set sia una collezione di 150 esempi delle 3 specie, 50 per ognuna. I dati possono essere raccolti in una matrice detta DESIGN MATRIX come TABULAR DATA - come in **??**.

Se abbiamo N elementi nel training set, ognuno con dimensione $D = \dim X + \dim Y$, allora abbiamo:

- BIG DATA se $N \gg D$, ovvero se il numero di elementi è molto superiore alla loro dimensione;
- WIDE DATA se $D \gg N$, ovvero se la dimensione degli elementi è molto superiore al loro numero.

Una buona idea è fare una *esplorazione dei dati* (EXPLOATORY DATA ANALYSIS) per vedere se ci sono dei pattern ovvi, ad esempio tramite grafici. Per grandi basi dati (big data) possiamo procedere mediante DIMENSIONALITY REDUCTION:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \left\{ \begin{array}{ll} p_l < 2.45\,\text{cm} & \text{Setosa} \\ \text{Altrimenti} & \left\{ \begin{array}{ll} p_w < 1.75\,\text{cm} & \text{Versicolor} \\ \text{Altrimenti} & \text{Virginica} \end{array} \right. \end{array} \right\} \text{ DECISION TREE,} \tag{4}$$

ove $\boldsymbol{\theta}$ è detto THRESHOLD PARAMETER. Questo decision tree è visualizzato in **??**. La performance può essere quindi misurata con il MISCALSSIFICATION RATE:

$$\mathcal{L}(\theta) \equiv \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left(y_n \neq f(\boldsymbol{x_n}, \boldsymbol{\theta})\right), \tag{5}$$

dove $\mathbb{I}(e)$ è l'**indicatore binario**

$$\mathbb{I}(e) = \left\{ \begin{array}{l} 1 \text{ se } e \text{ è vero} \\ 0 \text{ se } e \text{ è falso} \end{array} \right. . \tag{6}$$

Nel caso in cui alcuni errori di classificazione siano più dannosi di altri posso definire una loss function $l(y, \hat{y})$ e ridefinire il misclassification rate come l'EMPIRICAL RISK:

$$\mathcal{L}(\theta) \equiv \frac{1}{N} \sum_{n=1}^{N} l\left(y_n, f(\boldsymbol{x_n}, \boldsymbol{\theta})\right). \tag{7}$$

Un modo che abbiamo per definire il TRAINING (o MODEL FITTING) è modificare questo rischio empirico, ovvero trovare $\hat{\theta}$ tale che

$$\mathcal{L}(\hat{\theta}) = \min[\mathcal{L}(\theta)]. \tag{8}$$

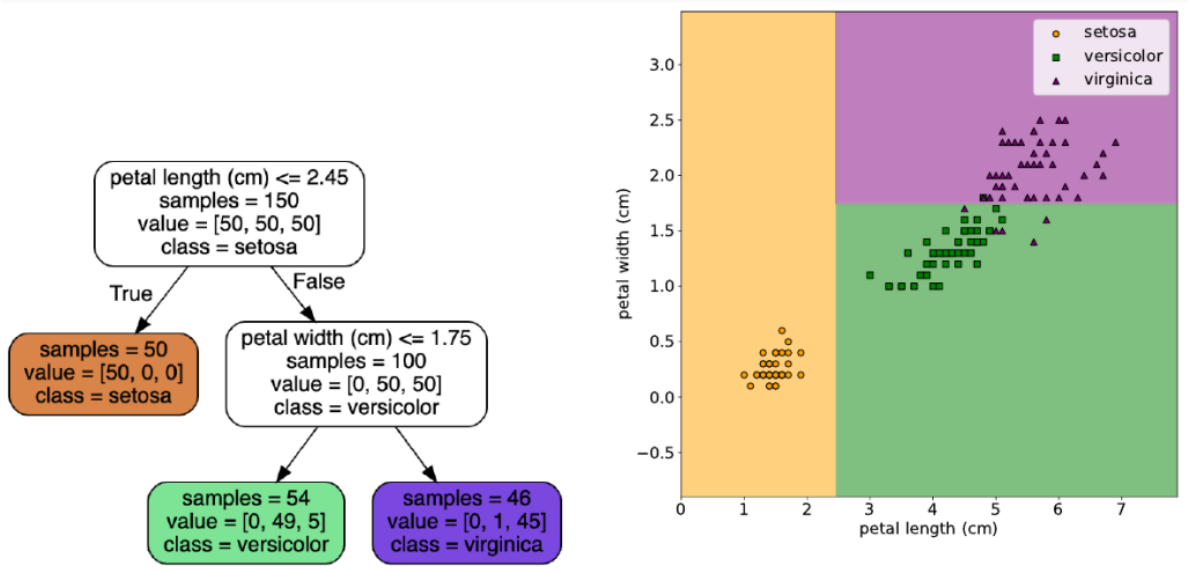

Figura 1: Decision tree per problema di classificazione specie di iris.

2 Richiami di probabilità

Abbiamo diverse definizioni di probabilità.

Definizione Frequentistica La probabilità di un evento è il rapporto tra il numero di casi favorevoli e il numero di casi possibili.

Definizione Soggettiva La probabilità di un evento è il prezzo che un individuo ritiene equo pagare per ricevere 1 se l'evento si verifica e 0 altrimenti.

Definizione Bayesiana La probabilità di un evento è l'*incertezza* con cui l'evento si verifica.

Definizione Assiomatica Kolmogorov nel 1933 costruisce la teoria della probabilità a partire da degli assiomi.

L'incertezza può essere di due tipi:

aleatoria ovvero è una DATA UNCERTANTY;

epistemica ovvero è una MODEL UNCERTANTY;

2.1 Proprietà della probabilità

Chiamiamo $\Pr(A)$ la probabilità dell'evento A, allora abbiamo le seguenti proprietà.

Proprietà 2.1.1 (Joint probability)**.** Se A e B sono due eventi indipendenti, allora:

$$\Pr(A \wedge B) \equiv \Pr(A, B) = \Pr(A) \cdot \Pr(B). \tag{9}$$

Proprietà 2.1.2 (Union probability)**.** Anche detta regola di unione esclusione. Se A e B sono due eventi indipendenti, allora:

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A \wedge B). \tag{10}$$

Proprietà 2.1.3 (Conditional probability)**.**

$$\Pr(A \mid B) = \frac{\Pr(A, B)}{\Pr(A)}. \tag{11}$$

Se i due eventi sono indipendenti questa si riduce a $\Pr(A \mid B) = \Pr(A)$.

Proprietà 2.1.4 (Conditional independece)**.**

$$\Pr(A, B \mid C) = \Pr(A \mid C) \cdot \Pr(B \mid C) \tag{12}$$

2.2 Random variables

Rappresentiamo con X una variabile di cui non conosciamo il valore e la chiamiamo *variabile casuale* (RANDOM VARIABLE). Il set dei valori che X può assumere è detto *spazio di sampling* (SAMPLING SPACE). Un evento è dunque un set di risultati dato un sampling space definito.

Se la variabile è **discreta** abbiamo un sampling space numerabile e la PMF (PROBABILITY MASS FUNCTION):

$$p(x) \equiv \Pr(X = x). \tag{13}$$

Se invece la variabile è **continua** abbiamo un samplig space non numberabile e la CDF (CUMULATIVE DISTRIBUTION FUNCTION):

$$P(x) \equiv \Pr(X \leq x), \tag{14}$$

da cui possiamo definire la PDF (PROBABILITY DENSITY FUNCTION):

$$p(x) \equiv \frac{d}{dx} P(x), \tag{15}$$

da cui segue:

$$\Pr(a \leq X \leq b) = \int_a^b p(x) = P(b) - P(a) \tag{16}$$

$$\implies \Pr(x \leq X \leq x + dx) \approx p(x). \tag{17}$$

Se la CDF $P(x)$ è monotona crescente allora la sua inversa $P^{-1}(q)$ è detta *quantile*. Il valore $x_q = P^{-1}(q)$ è il valore per cui $\Pr(X \leq x_q) < q$, ovvero il quantile q della distribuzione P.

Se abbiamo due variabili casuali X e Y allora possiamo definire la JOINT DISTRIBUTION:

$$p(x, y) = p(X = x, Y = y) \equiv p(X = x \wedge Y = y). \tag{18}$$

Se le due variabili sono indipendenti e con cardinalità finita possiamo definire la *distribuzione marginale* (MARGINAL DISTRIBUTION) come:

$$p(X = x) = \sum_y p(X = x, Y = y), \tag{19}$$

altrimenti se sono dipendenti la *distribuzione condizionale* (CONDITIONAL DISTRIBUTION) come:

$$\begin{aligned} p(Y = y \mid X = x) &= \frac{p(X = x, Y = y)}{p(X = x)} \\ \implies \quad p(x, y) &= p(y \mid x) \cdot p(x), \end{aligned} \tag{20}$$

da cui segue la **chain rule**:

$$p(\boldsymbol{x}_{1,D}) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_1, x_2) \cdots p(x_D \mid p_{1:D-1}). \tag{21}$$

Due variabili si dicono MARGINALMENTE INDIPENDENTI se

$$X \perp Y \quad \iff \quad p(X, Y) = p(X)\, p(Y), \tag{22}$$

mentre si dicono **condizionalmente indipendenti** se

$$X \perp Y \quad \iff \quad p(X, Y \mid Z) = p(X \mid Z)\, p(Y \mid Z). \tag{23}$$

2.3 Momenti di una distribuzione

Definizione 2.1 (Media)**.** Definiamo MEDIA di una distribuzione come

$$\mathbb{E}[X] \equiv \int_X x\, p(x)\, dx \qquad \left(\mathbb{E}[X_{\text{discreta}}] = \sum_{x \in X} x\, p(x)\right). \tag{24}$$

La media è lineare, ovvero $\mathbb{E}[\sum_{i=1}^n] = \sum_{i=1}^N \mathbb{E}(X_i)$.

Definizione 2.2 (Varianza)**.** Definiamo VARIANZA di una distribuzione come

$$\mathbb{V}[X] \equiv \mathbb{E}\left[(X-\mu)^2\right], \tag{25}$$

che con la definizione di prima possiamo anche scrivere come:

$$\begin{aligned}\mathbb{V}[X] &= \int_X (x-\mu)^2\, p(x)\, dx \\ &= \int x^2\, p(x)\, dx \,+\, \mu^2 \underbrace{\int p(x)\, dx}_{=1} - 2\mu \underbrace{\int x\, p(x)\, dx}_{=\mu} \\ &= \mathbb{E}\left[X^2\right] - \mu^2.\end{aligned} \tag{26}$$

Con questa definizione è facile dimostrare l'identità

$$\mathbb{V}[aX+b] = a^2\, \mathbb{V}[X], \tag{27}$$

inoltre anche la varianza è lineare, ma solo se le variabili sono indipendenti.

Definizione 2.3 (Moda)**.** Definiamo MEDIA di una distribuzione il valore (o i valori) più probabile, ovvero

$$x^\star \text{ t.c. } p(x^\star) = \max p(x), \tag{28}$$

Questi stimatori non danno tutta l'informazione contenuta nella distribuzione.

2.4 Teorema di Bayes

Teorema 2.1 (Teorema di Bayes)**.** *Data una quantità non nota H (hypotesis) e dei dati noti $Y = y$ abbiamo che*

$$p(H = h \mid Y = y) = \frac{p(H=h)\, p(Y = y \mid H = h)}{p(Y = y)}. \tag{29}$$

Il teorema segue dall'identità $p(h \mid y)\, p(y) = p(h)\, p(y \mid h)$.

- $p(h)$ viene detta PRIOR ovvero ciò che conosciamo o assumiamo per H *prima* di fare qualunque misura;
- $p(y \mid h)$ è la *distribuzione osservata* detta LIKELIHOOD, è una funzione di y ad h fisso, ma non è un distribuzione di probabilità;
- $p(y)$ è detta MARGINAL LIKELIHOOD, indipendente da h, funge da costante di normalizzazione

$$p(Y = y) = \sum_{h' \in H} p(Y = y \mid H = h') = \sum_{h' \in H} p(H = h', Y = y);$$

- $p(h \mid y)$ è detta POSTERIOR DISTRIBUTION.

Esempio 2.1 (Monty Hall problem)**.** Posso scegliere una di tre porte. Dietro una delle porte c'è un premio. Inizio con il scegliere la porta 1. Ho che la probabilità degli eventi $H_{1,2,3}$ che il premio sia dietro la porta 1, 2 o 3 è

$$p(H_1) = p(H = 1) = \frac{1}{3}.$$

Ora una delle porte non scelte da me si apre rivelando che dietro quella non c'è il premio, in base a dove si trova il premio la probabilità degli eventi $Y_{2,3}$ che si apra una porta o l'altra è data da $p(Y = y, H = h)$:

$$\begin{aligned} &p(Y = 2, H_1) = \tfrac{1}{2}, \qquad p(Y = 2, H_2) = 0, \qquad p(Y = 2, H_3) = 1, \\ &p(Y = 3, H_1) = \tfrac{1}{2}, \qquad p(Y = 3, H_2) = 1, \qquad p(Y = 3, H_3) = 1. \end{aligned},$$

Si apre la porta 3. Mi viene data la scelta di cambiare porta prima che venga rivelato dove si trova il premio, per massimizzare la probabilità di vittoria posso calcolare le probabilità condizionate con Bayes. Dai calcoli di prima ho che $p(() Y = 3) = 1/6 + 1/3 = 1/3$ dunque:

$$\begin{aligned} p(H_1 \mid Y = 3) &= \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{2}} = \frac{1}{3}, \\ p(H_2 \mid Y = 3) &= \frac{\frac{1}{3} \cdot 1}{\frac{1}{2}} = \frac{2}{3}, \\ p(H_3 \mid Y = 3) &= 0, \end{aligned} \tag{30}$$

dunque mi conviene cambiare porta.

2.5 Distribuzione di Gauss

La distribuzione di Gauss (o gaussiana) è una tra le più usate in Machine Learning ed è definita come

$$\begin{aligned} &P(y) \equiv \Pr(Y \leq y), \qquad \Pr(a \leq Y \leq b) = P(b) - (a) \\ &\Phi\left(y; \mu, \sigma^2\right) \equiv \int_{-\infty}^{y} \mathcal{N}\left(\xi/\mu, \sigma^2\right) d\xi = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\xi}{\sqrt{2}}\right)\right], \quad \xi = \frac{y - \mu}{\sigma}, \end{aligned} \tag{31}$$

ove erf è la Gauss error function definita come

$$\operatorname{erf}(u) \equiv \frac{2}{\sqrt{\pi}} \int_0^{\mu} e^{-t^2} dt, \tag{32}$$

mentre $\mathcal{N}$ è la PDF (probability density function) definita come

$$\mathcal{N}\left(y \mid \mu, \sigma^2\right) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right]. \tag{33}$$

Questa distribuzione è popolare per diversi motivi:

- dipende da soli due parametri ed entrambi sono di facile interpretazione;
- per il teorema del limite centrale, nel limite $N \longrightarrow \infty$ ogni distribuzione può essere approssimata da una gaussiana;
- rappresenta la somma di variabili causali indipendenti;
- il numero di assunzioni è minimo.

2.6 Altre distribuzioni notevoli

Abbiamo:

- t-Student

$$\tau\left(y \mid \mu, \sigma^2, \nu\right) \propto \left[1+\frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}} \qquad \begin{cases} \mu & \text{mean} \\ \sigma & \text{mode} \\ \nu & \text{degree of formality} \end{cases}; \tag{34}$$

- Lorentz (Cauchy)

$$C\left(x \mid \mu, \gamma\right) = \frac{1}{\gamma\pi}\left[1+\left(\frac{x-\mu}{\gamma}\right)^2\right]^{-1}; \tag{35}$$

- Laplace

$$L\left(y \mid \mu, b\right) = \frac{1}{2b}\exp\left(-\frac{|y-\mu|}{b}\right) \qquad \begin{cases} \mu & \text{mean} \\ \mu & \text{mode} \\ 2b^2 & \text{variance} \end{cases}; \tag{36}$$

- beta

$$\text{Beta}\left(x \mid a, b\right) = \frac{1}{\beta\left(a,b\right)}x^{a-1}\left(1-x\right)^{b-1} \qquad \begin{cases} \frac{a}{a+b} & \text{mean} \\ \left(a-1\right)\left(a+b-2\right) & \text{mode} \\ \frac{ab}{(a+b)^2(a+b+1)} & \text{variance} \end{cases}; \tag{37}$$

- gamma

$$\text{Ga}\left(x \mid a, b\right) = \frac{b^a}{\Gamma\left(a\right)}x^{a-1}e^{-xb}; \tag{38}$$

- experimental

$$\text{Exp}\left(x \mid \lambda\right) = \text{Ga}\left(x \mid a=1, b=\lambda\right) \tag{39}$$

- chi quadro

$$\chi^2_\nu\left(x\right) = \text{Ga}\left(x \mid a=\frac{\nu}{2}, b=\frac{1}{2}\right) \tag{40}$$

Alcune distribuzioni sono mostrate in **??**.

2.7 Osservazioni

Supponiamo di avere due variabili causali descritte da distribuzioni gaussiane:

$$x_1 \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right), \qquad x_2 \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right), \tag{41}$$

e di voler calcolare la PDF della loro somma $y = x_1 + x_2$, allora:

$$p\left(y\right) = \mathcal{N}\left(x_1 \mid \mu_1, \sigma_1^2\right) \otimes \mathcal{N}\left(x_2 \mid \mu_2, \sigma_2^2\right) = \mathcal{N}\left(y \mid \mu_1+\mu_2, \sigma_1^2+\sigma_2^2\right), \tag{42}$$

ovvero *la convoluzione di due gaussiane è una gaussiana.*

Supponiamo ora che x sia una variabile casuale e $y = f\left(x\right)$ una funzione di essa. Talora è difficile calcolare $p\left(y\right)$ analiticamente. Supponiamo ad esempio che $x \sim \text{Unif}\left(-1, +1\right)$ e $y = x^2$, possiamo approssimare $p\left(y\right)$ con un generatore (uniforme) di numeri casuali e facendone il quadrato per poi prendere la distribuzione empirica

$$p_S\left(y\right) = \frac{1}{N_S}\sum_{x=1}^{N_S}\delta\left(y-y_s\right). \tag{43}$$

Questo è detto metodo monte carlo.

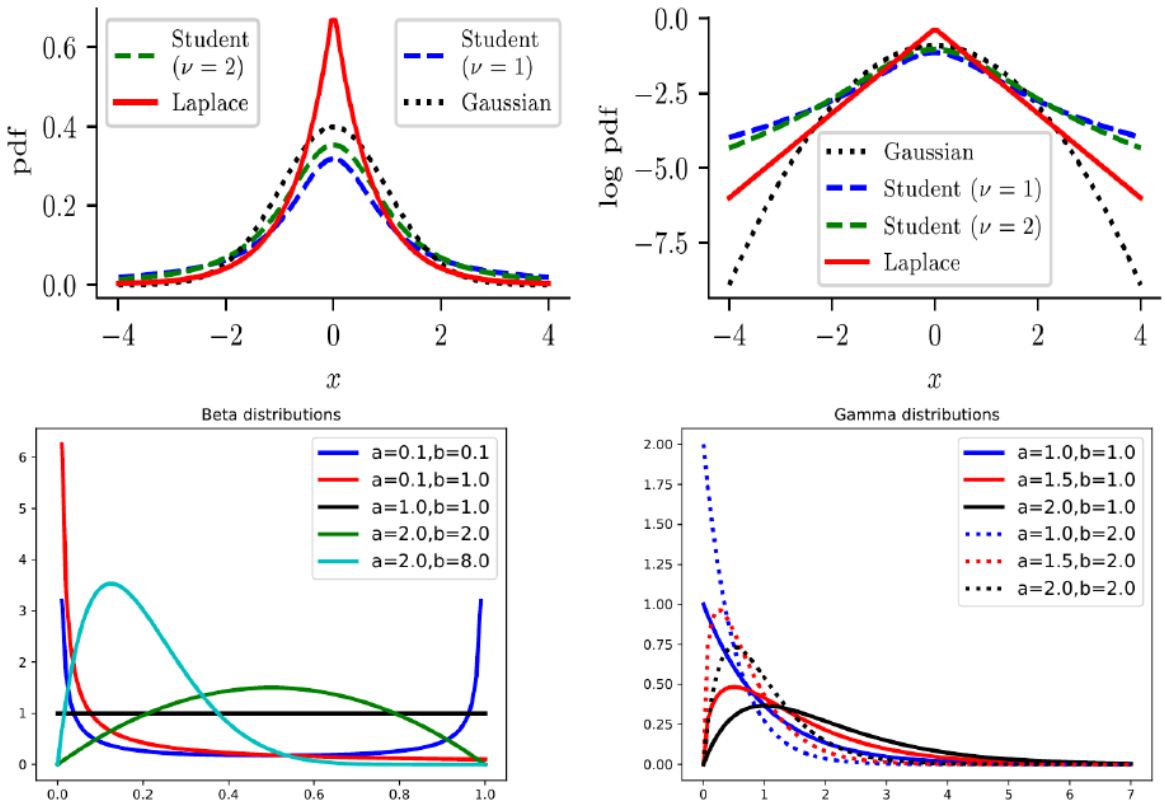

Figura 2: Alcune distribuzioni notevoli.

2.8 Modelli multivariati

Definizione 2.4 (Covarianza)**.** Siano X e Y due variabili. La COVARIANZA è

$$\mathrm{Cov}\left[X,Y\right] \equiv \mathbb{E}\left[\left(X-\mathbb{E}\left(E\right)\right)\left(Y-\mathbb{E}\left(E\right)\right)\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right], \tag{44}$$

ovvero una matrice D-dimensionale, con $D = \dim\left(\boldsymbol{x}\right)$, che contiene le varianze sulla diagonale.

Definizione 2.5. (Correlazione di Pearson) La CORRELAZIONE DI PEARSON tra le due variabili X e Y si definisce come

$$\rho \equiv \mathrm{corr}\left[X,Y\right] = \frac{\mathrm{Cov}\left[X,Y\right]}{\sqrt{\mathbb{V}\left[X\right]\mathbb{V}\left[Y\right]}}. \tag{45}$$

Possiamo scrivere

$$\mathrm{Cov}\left[X\right] = \begin{pmatrix} \mathbb{V}\left[X_1\right] & \mathrm{Cov}\left[X_1,X_2\right] & \cdots & \mathrm{Cov}\left[X_1,X_D\right] \\ \mathrm{Cov}\left[X_2,X_1\right] & \mathbb{V}\left[X_2\right] & \cdots & \mathrm{Cov}\left[X_2,X_D\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left[X_D,X_1\right] & \mathrm{Cov}\left[X_D,X_2\right] & \cdots & \mathbb{V}\left[X_D\right] \end{pmatrix} \tag{46}$$

e

$$\mathrm{corr}\left(x\right) = \begin{pmatrix} 1 & \frac{\mathbb{E}[(x_1-\mu_1)(x_2-\mu_2)]}{\sigma_1\sigma_2} & \cdots & \frac{\mathbb{E}[(x_1-\mu_1)(x_D-\mu_D)]}{\sigma_1\sigma_D} \\ \frac{\mathbb{E}[(x_2-\mu_2)(x_1-\mu_1)]}{\sigma_2\sigma_1} & 1 & \cdots & \frac{\mathbb{E}[(x_2-\mu_2)(x_D-\mu_D)]}{\sigma_2\sigma_D} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathbb{E}[(x_D-\mu_D)(x_1-\mu_1)]}{\sigma_D\sigma_1} & \frac{\mathbb{E}[(x_D-\mu_D)(x_2-\mu_2)]}{\sigma_D\sigma_2} & \cdots & 1 \end{pmatrix}. \tag{47}$$

2.8.1 Osservazioni

- Il fatto che due variabili non siano correlate non vuol dire che siano indipendenti, tuttavia due variabili indipendenti sono necessariamente non correlate;
- correlazione non implica causalità;
- una correlazione che appare simile in diversi set di dati può sparire (o divenire opposta) se i dati sono combinati, questo è noto come *Simpson's paradox*.

2.9 Distribuzione gaussiana multivariata

La definizione che abbiamo dato di distribuzione gaussiana si estende a più variabili. Innanzitutto definiamo la PDF

$$\mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{\mu}, \Sigma\right) \equiv \frac{1}{\left(2\pi\right)^{D/2}\left|\Sigma\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)^T \Sigma^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right\}, \tag{48}$$

dove $\Sigma = \mathrm{Cov}\left[\boldsymbol{y}\right]$.

Supponiamo di avere due variabili aleatorie $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$. Si può definire la distribuzione JOINTLY GAUSSIAN. Con

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \Lambda = \Sigma^{-1}$$

le distribuzioni marginali sono

$$p\left(\boldsymbol{y}_1\right) = \mathcal{N}\left(\boldsymbol{y}_1 \mid \boldsymbol{\mu}_1, \Sigma_{11}\right), \qquad p\left(\boldsymbol{y}_2\right) = \mathcal{N}\left(\boldsymbol{y}_2 \mid \boldsymbol{\mu}_2, \Sigma_{22}\right) \tag{49}$$

e la probabilità condizionata

$$p\left(\boldsymbol{y}_1 \mid \boldsymbol{y}_2\right) = \mathcal{N}\left(\boldsymbol{y}_1 \mid \boldsymbol{\mu}_{1|2}, \Sigma_{1|2}\right), \tag{50}$$

con

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(\boldsymbol{y}_2 - \boldsymbol{\mu_2}\right), \qquad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1},$$

queste sono tutte gaussiane.

3 Richiami di statistica

Nella sezione precedente abbiamo assunto come noti i parametri $\boldsymbol{\theta}$, in questa sezione impariamo come è possibile imparare questi parametri a partire dai dati (training set) $\mathcal{D}$. Come abbiamo visto nell'**??**, il problema si riduce a trovare $\hat{\theta}$ tale che minimizzi il misclassification rate, **??**, ovvero:

$$\hat{\theta} \quad \text{t.c.} \quad \mathcal{L}(\hat{\theta}) = \min\left[\mathcal{L}\left(\theta\right)\right] \equiv \min\left[\frac{1}{N}\sum_{n=1}^{N} l\left(y_n, f\left(\boldsymbol{x_n}, \boldsymbol{\theta}\right)\right)\right].$$

3.1 Maximum likelihood estimation (MLE)

Definizione 3.1 (Maximum likelihood estimation)**.** Definitiamo la MLE, MAXIMUM LIKELIHOOD ESTIMATION, come il valore di θ che massimizza la probabilità condizionata di aver ottenuto il set di dati $\mathcal{D}$, ovvero:

$$\hat{\theta}_{\text{MLE}} \equiv \underset{\boldsymbol{\theta}}{\text{argmax}}\, p\left(\mathcal{D} \mid \boldsymbol{\theta}\right). \tag{51}$$

Se i dati sono campioni indipendenti della stessa distribuzione, allora possiamo anche scrivere

$$p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) = \prod_{n=1}^{N} p\left(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}\right). \tag{52}$$

Definiamo inoltre la NLL, NEGATIVE LOG-LIKELIHOOD, da usare come loss function come:

$$l\left(\boldsymbol{\theta}\right) \equiv -\log p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) = -\sum_{n=1}^{N} \log p\left(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}\right). \tag{53}$$

Con queste definizioni abbiamo che la MLE è data

- per supervised learning da:

$$\hat{\theta}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmin}} -\sum_{n=1}^{N} \log p\left(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}\right); \tag{54}$$

- per unsupervised learning da:

$$\hat{\theta}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmin}} -\sum_{n=1}^{N} \log p\left(\boldsymbol{y}_n \mid \boldsymbol{\theta}\right). \tag{55}$$

Possiamo giustificare l'uso della MLE pensandola come un'approssimazione del *posterior* Bayesiano dato un *prior* uniforme.

Esempio 3.1. Se abbiamo

$$p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) = \delta\left(\boldsymbol{\theta} - \hat{\theta}_{\text{MAP}}\right)$$

con $\hat{\theta}$ il posterior, allora

$$\hat{\theta}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmin}} -\log p\left(\boldsymbol{\theta} \mid \mathcal{D}\right) = \underset{\boldsymbol{\theta}}{\text{argmin}} -\log p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) - \log p\left(\boldsymbol{\theta}\right),$$

ed essendo $p\left(\boldsymbol{\theta}\right) = 1$ abbiamo

$$\hat{\theta}_{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmin}} -\log p\left(\mathcal{D} \mid \boldsymbol{\theta}\right) = \hat{\theta}_{\text{MLE}}. \qquad \square$$

3.1.1 Distribuzione normale

Supponiamo ora di avere una variabile casuale Y distribuita normalmente, i.e. $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$, e sia $\mathcal{D} = \{y_n \text{ t.c. } n = 1, \cdots, N\}$ un dataset con punti campionati indipendentemente, allora:

$$\begin{aligned} \text{NLL}\left(\mu, \sigma^2\right) &= -\sum_{n=1}^{N_D} \log\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_n - \mu\right)^2\right)\right] \\ &= \frac{1}{2\sigma}\sum_{n=1}^{N_D}\left(y_n - \mu\right)^2 + \frac{N_D}{2}\log\left(2\pi\sigma^2\right). \end{aligned} \tag{56}$$

Ora estremizzando otteniamo

$$\frac{\partial}{\partial\mu}\text{NLL}\left(\mu, \sigma^2\right) = 0 \iff \hat{\mu}_{\text{MLE}} = \frac{1}{N_D}\sum_{n=1}^{N_D} y_n = \bar{y} \tag{57}$$

$$\frac{\partial}{\partial\sigma^2}\text{NLL}\left(\mu, \sigma^2\right) = 0 \iff \hat{\sigma}^2_{\text{MLE}} = \frac{1}{N_D}\sum_{n=1}^{N_D}\left(y_n - \hat{\mu}_{\text{MLE}}\right)^2 = \frac{1}{N_D}\sum_{n=1}^{N_D} y_n^2 - \bar{y}^2 = s^2 - \bar{y}^2. \tag{58}$$

3.1.2 Distribuzione normale multivariata

Per una distribuzione multivariata i risultati sono analoghi. Abbiamo

$$l\left(\boldsymbol{\mu}, \Sigma\right) = \log p\left(\mathcal{D} \mid \boldsymbol{\mu}, \Sigma\right) = \frac{N_D}{2}\log|\Lambda| = -\frac{1}{2}\sum_{n=1}^{N_D}\left(\boldsymbol{y}_n - \boldsymbol{\mu}\right)^T \Lambda\left(\boldsymbol{y}_n - \boldsymbol{\mu}\right) \tag{59}$$

dove $\Lambda = \Sigma^{-1}$ è la PRECISION MATRIX. Dall'estremizzazione segue, come prima, l'empirical mean $\hat{\boldsymbol{\mu}}$ e l'empirical covariance matrix $\hat{\Sigma}$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N_D}\sum_{n=1}^{N_D}\boldsymbol{y_n} = \bar{\boldsymbol{y}} \tag{60}$$

$$\hat{\Sigma} = \frac{1}{N_D}\sum_{n=1}^{N_D}\left(\bar{\boldsymbol{y}}_n - \bar{\boldsymbol{y}}\right)\left(\bar{\boldsymbol{y}}_n - \bar{\boldsymbol{y}}\right)^T \tag{61}$$

3.2 Empirical risk minimization (ERM)

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

$$\mathcal{L}\left(\boldsymbol{\theta}\right) = \frac{1}{N}\sum_{n=1}^{N} l\left(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n\right). \tag{62}$$

Questo processo di cambiare loss function è noto come ERM, EMPIRICAL RISK MINIMIZATION.

Esempio 3.2 (ERM per problema di classificazione)**.** In un problema di classificazione avremo una loss function

$$l_{01}\left(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n\right) = \begin{cases} 1 & \text{se} \quad \boldsymbol{y}_n = f\left(\boldsymbol{x}_n; \boldsymbol{\theta}\right) \\ 0 & \text{se} \quad \boldsymbol{y}_n \neq f\left(\boldsymbol{x}_n; \boldsymbol{\theta}\right) \end{cases} \tag{63}$$

dove $f\left(\boldsymbol{x}, \boldsymbol{\theta}\right)$ è un predittore. L'empirical risk diviene

$$\mathcal{L}\left(\boldsymbol{\theta}\right) = \frac{1}{N}\sum_{n=1}^{N} l_{01}\left(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n\right), \tag{64}$$

che è il misclassification rate nel training set. In problemi binari possiamo riscrivere il misclassification rate usando l'indicatore binario come definito in (**??**). Se $\tilde{y} \in \{-1, +1\}$ è il time label e $\hat{y} \in \{-1, +1\} = f\left(\boldsymbol{x}, \boldsymbol{\theta}\right)$ la prediction, allora:

$$l_{01}\left(\tilde{y}, \hat{y}\right) = \mathbb{I}\left(\tilde{y} \neq \hat{y}\right) = \mathbb{I}\left(\tilde{y}\hat{y} < 0\right), \tag{65}$$

con cui il rischio empirico si scrive

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} l_{01}(\tilde{y}_n, \hat{y}_n) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(\tilde{y}_n \hat{y}_n < 0). \tag{66}$$

La funzione l_{01} è però problematica in quanto non-smooth e quindi difficile da ottimizzare. In questi casi si può utilizzare una funzione surrogata, definita generalmente come limite superiore convesso, ad esempio

$$l_{ll}(\tilde{y}, \eta) = -\log p(\tilde{y} \mid \eta) = \log\left(1 + e^{-\tilde{y}\eta}\right), \qquad p(\tilde{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) = \sigma(\hat{y}\eta) = \frac{1}{1 + e^{-\hat{y}\eta}} \tag{67}$$

3.3 Altri metodi di minimizzazione

3.3.1 Metodo dei momenti

Il calcolo di $\nabla_{\boldsymbol{\theta}} \mathrm{NLL}(\boldsymbol{\theta}) = 0$, ovvero la ricerca dei punti estremi, può essere difficile. Il metodo dei momenti consiste nell'eguagliare i momenti teorici della distribuzione ai momenti empirici, ovvero:

$$\left.\begin{array}{ll} \text{Momenti teorici:} & \mu_k = \mathbb{E}\left[Y^k\right] \\ \text{Momenti empirici:} & \hat{\mu}_k = \frac{1}{N}\sum_{n=1}^{N} y_n^k \end{array}\right\} \mu_k = \hat{\mu}_k \quad \Longrightarrow \quad \text{ho } k \text{ equazioni.} \tag{68}$$

Ad esempio per la gaussiana abbiamo:

$$\begin{aligned} \mu_1 &= \mu = \bar{y}, \\ \mu_2 &= \sigma^2 + \mu^2 = s^2. \end{aligned}$$

In questo caso abbiamo che MOM = MLE, ma questo non è vero in generale (ad esempio non è vero per la distribuzione uniforme).

3.3.2 Online recursive estimation

Se tutto il dataset $\mathcal{D}$ è noto e disponibile prima che inizi il processo di learning si dice che si fa BATCH LEARNING. In alcuni casi però il dataset è disponibile in blocchi e si dice che si fa ONLINE LEARNING. Si ha che $\hat{\boldsymbol{\theta}}_{t-1}$ è la predizione dato il blocco $\mathcal{D}_{1:t-1}$, quello che vogliamo fare è trovare $\boldsymbol{\theta}_t = f(\hat{\boldsymbol{\theta}}_{t-1}, \boldsymbol{y}_t)$ con un update ricorsivo.

Definizione 3.2 (Moving average)**.** Per una gaussiana multivariata si può definire la MEDIA MOBILE come:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_t &= \frac{1}{t}\sum_{n=1}^{t} \boldsymbol{y}_n \\ &= \frac{1}{t}\left[(t-1)\,\hat{\boldsymbol{\mu}}_{t-1} + \boldsymbol{y}_t\right] = \hat{\boldsymbol{\mu}}_{t-1} + \frac{1}{t}\left(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_{t-1}\right). \end{aligned} \tag{69}$$

Nel caso in cui la distribuzione cambi in modo continuo la moving average può anche essere pesata.

3.4 Regolarizzazione

Un problema della MLE (maximum likelihood estimation) e della ERM (empirical risk minimization) è che i parametri tendono ad essere determinati minimizzando il loss nel training set, ma non necessariamente lo minimizzano per i dati futuri. Per ridurre l'entità del problema si può procedere con la regolarizzazione:

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = \Big[\underbrace{\frac{1}{N}\sum_{n=1}^{N} l(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n)}_{\mathcal{L}(\boldsymbol{\theta})} + \lambda C(\boldsymbol{\theta})\Big], \tag{70}$$

dove λ è detto REGULARIZATION PARAMETER e $C(\boldsymbol{\theta})$ è la PENALITY FUNCTION, ad esempio si può avere:

$$C(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}).$$

Per $\lambda = 1$ e riscalando $p(\boldsymbol{\theta})$ si ha la NLL:

$$\text{NLL}(\boldsymbol{\theta}, \lambda) = -\left[\sum_{n=1}^{N} \log p(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right] = -\left[\log p(\mathcal{D} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right], \tag{71}$$

che implica:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \log p(\boldsymbol{\theta} \mid \mathcal{D}) = \underset{\boldsymbol{\theta}}{\text{argmax}} \left[\log p(\mathcal{D} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \text{const}\right]. \tag{72}$$

Questa stima si chiama MAP (MAXIMUM A POSTERIOR) estimation.

3.4.1 Come scegliere λ

Un valore (troppo) piccolo per λ equivale a minimizzare il rischio empirico (OVERFITTING); un valore (troppo) grande equivale a essere troppo vicini al prior (UNDERFITTING). Una soluzione è dividere il training set in due classi: una di training ($\mathcal{D}_{\text{train}}$, 80%) ed una di validation ($\mathcal{D}_{\text{val}}$, 20%). Si va quindi a fittare il modello su $\mathcal{D}_{\text{train}}$ per ogni setting λ e poi a valutare la performance del modello ottenuto su $\mathcal{D}_{\text{val}}$. Si prende dunque il valore di λ per cui si è ottenuta la performance migliore.

Definizione 3.3 (Regularized empirical risk)**.** Definiamo l'EMPIRICAL RISK REGOLARIZZATO come:

$$R_\lambda(\boldsymbol{\theta}, \mathcal{D}) \equiv \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} l(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{\theta})) + \lambda C(\theta) \tag{73}$$

Il processo che seguiamo è quindi il seguente:

1. costruiamo il modello: $\hat{\boldsymbol{\theta}}_\lambda(\mathcal{D}_{\text{train}}) = \text{argmin}_{\boldsymbol{\theta}} R_\lambda(\boldsymbol{\theta}, \mathcal{D}_{\text{train}})\ \forall\lambda$;
2. calcoliamo il validation risk: $R_\lambda^{\text{val}} = R_0\left(\hat{\boldsymbol{\theta}}_\lambda(\mathcal{D}_{\text{train}}), \mathcal{D}_{\text{val}}\right)$;
3. prendiamo il λ che ha dato la miglior performance: $\lambda^\star = \text{argmin}_{\lambda\in S} R_\lambda^{\text{val}}$;
4. si ri-fitta il modello su $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}$ usando $\lambda = \lambda^\star$;
5. si ottengono i parametri $\hat{\boldsymbol{\theta}}^\star = \text{argmin}_{\boldsymbol{\theta}} R_{\lambda^\star}(\boldsymbol{\theta}, \mathcal{D})$.

Se il dataset è piccolo, eliminare il 20% dei dati può essere dannoso e si procede quindi con la CROSS VALIDATION. Questa consiste nel dividere il dataset in K FOLD e per ognuno allenare il modello su tutti i fold meno il k-esimo, il quale viene usato come test fold. Si definisce l'empirical risk regolarizzato per cross validation:

$$R_\lambda^{\text{CV}} \equiv \frac{1}{K} \sum_{k=1}^{K} R_0\left(\hat{\boldsymbol{\theta}}_\lambda(\mathcal{D} - \mathcal{D}_k), \mathcal{D}_k\right) \tag{74}$$

3.5 Statistica Bayesiana

Abbiamo ora visto diversi metodi per determinare i parametri a partire dei dati, ma dobbiamo ancora parlare della loro incertezza. Il **??** (**??**) ci aiuta in questo, infatti:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\boldsymbol{\theta})\, p(\mathcal{D} \mid \boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\boldsymbol{\theta})\, p(\mathcal{D} \mid \boldsymbol{\theta})}{\int p(\boldsymbol{\theta}')\, p(\mathcal{D} \mid \boldsymbol{\theta}')\, d\boldsymbol{\theta}'} \tag{75}$$

dove $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ per supervised learning e $\mathcal{D} = \{(\boldsymbol{y}_n)\}$ per unsupervised learning, $n = 1, \ldots, N$.

Una voltache il posterior sui paramentri è stato determinato, possiamo calcolare il posterior della distribuzione predittiva marginalizzando su $\boldsymbol{\theta}$:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \theta)\, p(\boldsymbol{\theta} \mid \mathcal{D})\, d\theta, \tag{76}$$

questo è noto come BMA (Bayes model averaging).

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

$$p(\mathcal{D} \mid \mu) \propto \exp\left\{-\frac{1}{2\sigma}\sum_{n=1}^{N_D}(y_n-\mu)^2\right\}. \tag{77}$$

Si può mostrate che il prior coniugato è un'altra gaussiana $\mathcal{N}\left(\mu \mid \check{m}, \check{\tau}^2\right)$ e usando il **??** (**??**) troviamo che il posterior è

$$\begin{aligned} &p\left(\mu \mid \mathcal{D}, \sigma^2\right) = \mathcal{N}\left(\mu \mid \hat{m}, \hat{\tau}^2\right) \\ &\text{dove} \begin{cases} \hat{\tau}^2 &= \frac{1}{\frac{N}{\sigma^2}+\frac{1}{\check{\tau}^2}} = \frac{\sigma^2\check{\tau}^2}{N\check{\tau}^2+\sigma^2} \\ \hat{m} &= \hat{\tau}^2\left(\frac{\check{m}}{\check{\tau}^2}+\frac{N\bar{y}}{\sigma^2}\right) = \frac{\sigma^2}{N\check{\tau}^2+\sigma^2}\check{m}+\frac{N\check{\tau}^2}{N\check{\tau}^2+\sigma^2}\bar{y} \end{cases} \end{aligned} \tag{78}$$

e con $\bar{y} \equiv \frac{1}{N}\sum_{n=1}^{N} y_n$ la media empirica. Definendo ora $\kappa = 1/\sigma^2$ e $\check{\lambda} = 1/\check{\tau}^2$ posso scrivere:

$$\begin{aligned} \hat{\lambda} &= \check{\lambda} + N\kappa \\ \hat{m} &= \frac{N\kappa\bar{y}+\check{\lambda}\check{m}}{\hat{\lambda}} = \frac{N\kappa}{N\kappa+\check{\lambda}}\bar{y} + \frac{\check{\lambda}}{N\kappa+\check{\lambda}}\check{m}. \end{aligned} \tag{79}$$

La precisione del posterior $\hat{\lambda}$ è dunque la precisione del prior $\check{\lambda}$ più N unità di misura di precisione κ. La media del posterior ($\hat{m}$) è invece una combinazione convessa della media empirica $\bar{y}$ e della media del prior $\check{m}$. Consideriamo ora il posterior dopo aver visto un singolo dato y (i.e. abbiamo $N = 1$), si ha che la posteriori mean può essere data in tre modi:

$$\hat{m} = \underbrace{\frac{\check{\lambda}}{\hat{\lambda}}\check{m} + \frac{\kappa}{\hat{\lambda}}\bar{y}}_{\substack{\text{convex combination}\\ \text{of prior and data}}} = \underbrace{\check{m} + \frac{\kappa}{\hat{\lambda}(\bar{y}-\check{m})}}_{\substack{\text{prior mean adjusted}\\ \text{to data}}} = \underbrace{\bar{y} = \frac{\check{\lambda}}{\hat{\lambda}}(\bar{y}-\check{m})}_{\substack{\text{data adjusted to}\\ \text{the prior mean}}} \tag{80}$$

ed abbiamo lo standard error:

$$\operatorname{se}(\mu) = \sqrt{\mathbb{V}\left[\mu\left[\mathcal{D}\right]\right]}. \tag{81}$$

Se uso un *uninformative prior* per μ ponendo $\check{\lambda} = 0$, allora $\hat{m} = \bar{y}$. Supponendo di approssimare $\sigma^2 \sim s^2 \equiv \frac{1}{N}\sum_{n=1}^{N}(y_n-\bar{y})^2$ segue che

$$\hat{\lambda} = N\kappa = \frac{N}{s^2} \quad \Longrightarrow \quad \operatorname{se}(\mu) = \frac{1}{\sqrt{\lambda}} = \frac{s}{\sqrt{N}}. \tag{82}$$

Ovvero abbiamo che l'incertezza decresce ad un rateo di $1/\sqrt{N}$.

Osservazione 3.5.1. Quando non abbiamo informazioni sul prior è desiderabile usare un prior uninformative. Ad esempio un flat prior, i.e. $p(\mu) = 1$.

Osservazione 3.5.2. Qualunque modello Bayesiano richiede di specificare un prior $p(\theta)$ per i parametri. I parametri del prior sono detti **iperparametri** e sono denotati con Φ. Se non sono noti possiamo mettrci sopra un prior (multi-level model $\Phi \to \theta\mathcal{D}$) e definire la joint distribution $p(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathcal{D}) = p(\boldsymbol{\Phi})\, p(\boldsymbol{\theta} \mid \boldsymbol{\Phi})\, p(\mathcal{D} \mid \theta)$.

3.6 Intervalli credibili e di confidenza

Una distribuzione posteriore è un oggetto multidimensionale difficile da visualizzare e da trattere. Può essere quindi utile calcolare degli stimatori puntuali, come la media e moda posteriore, per determinare un intervallo di credibilità che quantifichi l'incertezza associata alle stime.

Method	Definition
MLE	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p\left(\mathcal{D} \mid \theta\right)$
MAP	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, p\left(\mathcal{D} \mid \theta\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{\Phi}\right)$
Full Bayes	$p\left(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathcal{D}\right) \propto p\left(\boldsymbol{\Phi}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{\Phi}\right) p\left(\mathcal{D} \mid \theta\right)$

Tabella 2: Metodi visti per ottenere i parametri.

Definizione 3.4 (Intervallo di credibilità)**.** L'INTERVALLO DI CREDIBILITÀ $100\left(1-\alpha\right)\%$ è la regione $C = (l, u)$ che contiene $1-\alpha$ della probabilità posteriore, ovvero:

$$C_\alpha\left(\mathcal{D}\right) = (l, u) \text{ t.c. } P\left(l \leq \theta \leq u \mid \mathcal{D}\right) = 1 - \alpha \tag{83}$$

dove l si dice *lower* e u *upper*. Inoltre se il posterior ha una PDF nota, allora $l = F^{-1}\left(\alpha/2\right)$ e $u = F^{-1}\left(1-\alpha/2\right)$, ove F è la CDF del posterior.

Definizione 3.5 (Intervallo di confidenza)**.** Si dice INTERVALLO DI CONFIDENZA $100\left(1-\alpha\right)\%$, per la stima di un parametro θ, la regione $I(\tilde{\mathcal{D}}) = (l(\tilde{\mathcal{D}}), u(\tilde{\mathcal{D}}))$ ottenuto da un data set $\mathcal{D}$ tale che

$$\Pr\left(\theta \in I\left(\tilde{\mathcal{D}} \mid \mathcal{D} \sim \theta\right)\right) = 1 - \alpha. \tag{84}$$

Osservazione 3.6.1. L'intervallo di credibilità è un concetto Bayesiano mentre l'intervallo di confidenza è un concetto frequentista. CL (frequentista): θ è una costante fissa non nota e i dati sono aleatori. CI (Bayesiano): i dati sono fissi perchè noti, θ è invece ignota. Un CL al 95% non significa che il parametro stia verosimilmente dentro l'intervallo il 95% delle volta dati i dati osservati.

Esempio 3.3. Supponiamo di generare due interi $\mathcal{D} = (y_1, y_2)$ da:

$$p\left(y \mid \theta\right) = \begin{cases} 0.5 & \text{se} \quad y = \theta \\ 0.5 & \text{se} \quad y = \theta + 1 \\ 0 & \text{se} \quad \text{altrimenti} \end{cases}.$$

Se $\theta = 39$ possiamo ottenere i seguenti risultati:

$$\left(39, 39\right), \left(39, 40\right), \left(40, 39\right), \left(40, 40\right).$$

Sia $m = \min\left(y_1, y_2\right)$ e l'intervallo $[l(\mathcal{D}), U(\mathcal{D})] = [m, m]$. Segue che i possibili intervalli sono:

$$[39, 39], [39, 39], [39, 39], [40, 40]$$

e dunque l'intervallo definito risulta in un CI 75% per $\theta = 39$. Se però osserviamo $\mathcal{D} = (39, 40)$ allora $p\left(\theta = 39 \mid \mathcal{D}\right)$, ma CI è solo 75%. Il CI fallisce per esperimenti non ripetibili.

3.7 Bias-variance tradeoff

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

Definizione 3.6 (Bias)**.** Il BIAS di uno stimatore $\hat{\theta}$ è definito come:

$$\text{bias}\left(\theta(\cdot)\right) \equiv \mathbb{E}\left[\hat{\theta}\left(\mathcal{D}\right)\right] - \theta^\star \tag{85}$$

dove $\theta^\star$ è il valore vero. Se il bias è nulla, lo stimatore viene detto UNBIASED.

Per una gaussiana abbiamo:

$$\text{bias}\left(\hat{\mu}\right) = \mathbb{E}\left[\bar{x}\right] - \mu = \mathbb{E}\left[\frac{1}{N_D}\sum_{n=1}^{N_D} x_n\right] - \mu = \frac{N_D\mu}{N_D} - \mu = 0 \quad \Longrightarrow \quad \text{Unbiased}, \tag{86}$$

$$\text{bias}\left(\hat{\sigma}^2\right) = \mathbb{E}\left[s^2\right] - \sigma^2 = \frac{N_D - 1}{N_D}\sigma^2 - \sigma^2 \neq 0 \quad \Longrightarrow \quad \text{Biased}. \tag{87}$$

Possiamo però definire una versione unbiased della varianza come:

$$\sigma^2_{\text{unb}} = \frac{1}{N_D - 1} \sum_{n=1}^{N_D} (x_n - \bar{x})^2 = \frac{N_D}{N_D - 1} \sigma^2_{\text{MLE}} \tag{88}$$

Definizione 3.7 (Varianza di uno stimatore)**.** La VARIANZA DI UNO STIMATORE è

$$\mathbb{V}(\hat{\theta}) \equiv \mathbb{E}(\hat{\theta}^2) - (\mathbb{E}[\hat{\theta}])^2 \tag{89}$$

Idealmente vogliamo avere una varianza minima per tutti gli stimatori. Il seguente teorema fornisce un limite minimo alla varianza.

Teorema 3.1 (Disuguaglianza di Cramér-Rao)**.** *Per uno stimatore unbiased $\theta^\star$ si ha che:*

$$\mathbb{V}\left[\hat{\theta}\right] \geq \frac{1}{NF(\theta^\star)}, \tag{90}$$

ove $F(\theta^\star)$ è la Fisher information.

Si può dimostrare che la MLE raggiunge il bond imposto da Cramér-Rao. Sia $\hat{\theta} = \hat{\theta}(\mathcal{D})$ l'estimato e $\bar{\theta} = \mathbb{E}\left[\hat{\theta}\right]$ il valore di aspettazione corrsispondente (tutti da $p(\mathcal{D} \mid \theta^\star)$), allora:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\theta} - \theta^*\right)^2\right] &= \mathbb{E}\left[\left[(\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta^\star)\right]^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \bar{\theta}\right)^2\right] + 2\mathbb{E}\left[\left(\hat{\theta} - \bar{\theta}\right)^2\right] + 2\left(\bar{\theta} - \theta^\star\right)\mathbb{E}\left[\hat{\theta} - \theta^\star\right] + \bar{\theta} - \theta^{\star 2} \\
&= \mathbb{E}\left[\left(\hat{\theta} - \bar{\theta}\right)^2\right] + \left(\bar{\theta} - \theta^\star\right)^2 = \mathbb{V}\left[\hat{\theta}\right] + \text{bias}^2\left(\hat{\theta}\right)
\end{aligned} \tag{91}$$

Esempio 3.4. Supponiamo di stimare la media di una distribuzionne gaussiana a partire dai dati $\boldsymbol{x} = (x_1, \ldots, x_n)$ assumendo che questi siano stati campionati da $x_n \sim \mathcal{N}\left(\theta^\star = 1, \sigma^2\right)$. Uno stimatore ovvio è dato dalla MLE che ha bias nullo e varianza $\mathbb{V}[\bar{x}/\theta^\star] = \sigma^2/N$. Possiamo dare uno stimatore MAP sotto un prior gaussiano, in tal caso:

$$\tilde{x} = \frac{N}{N + \kappa_0}\bar{x} + \frac{\kappa_0}{N + \kappa_0}\theta_0 = w\bar{x} + (1 - w)\,\theta_0$$

con prior $\mathcal{N}\left(\theta_0, \sigma^2/\kappa_0\right)$. Allora abbiamo:

$$\begin{aligned}
&\mathbb{E}_{\text{MAP}}\left[\tilde{x}\right] - \theta^\star = w\theta^\star + (1 - w)\theta_0 - \theta^\star = (1 - w)\left(\theta_0 - \theta^\star\right), \\
&\mathbb{V}_{\text{MAP}}\left[\tilde{x}\right] = w^2\frac{\sigma^2}{N} < \mathbb{V}_{\text{MLE}.}\left[\hat{x}\right],
\end{aligned}$$

ovvero è uno stimatore biased essendo $0 < w < 1$.

4 Ottimizzazione

Abbiamo viesto nelle precedenti lezioni che il machine learning ha come obbiettivo quello di stimare dei parametri $\boldsymbol{\theta}$ (o degli iperparametri Φ) di un modello, ovvero determinare i valori di best-fit e relativi intervalli di credibilità. Questo si realizza minimizzando una funzione scalare, la less function

$$\mathcal{L} : \Theta \to \mathbb{R},$$

dove $\theta \in \Theta$, per trovare:

$$\theta^* = \underset{\boldsymbol{\theta}}{\text{argmin}}\,\mathcal{L}(\boldsymbol{\theta}) \quad \text{o meglio} \quad \hat{\theta} = \underset{\boldsymbol{\theta}}{\text{argmin}}\,\mathcal{L}[\boldsymbol{\theta}] \quad \text{per } \theta^\star \sim \hat{\theta}.$$

Assumiamo per semplicità che $\Theta \subseteq \mathbb{R}^D$, con D il numero delle variabili ottimizzate, in questo modo abbiamo che l'ottimizzazione è **continua**.

Osservazione 4.0.1. Lo spazio dei parametri è tale che può esistere un solo minimo globale o più minimi globali, oltre ad una serie di minimi locali. Il punto di minimo è computazionalmente *sempre in termini locali*, i.e.:

$$\exists \delta > 0,\ \forall \boldsymbol{\theta} \in \Theta \quad \text{t.c.} \quad ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|| < \delta,\ \mathcal{L}(\hat{\boldsymbol{\theta}}) < \mathcal{L}(\boldsymbol{\theta}).$$

un minimo locale può essere circondato da minimi simili (flat direction). Un algoritmo che converge ad un punto stazionario è detto GLOBALLY CONVERGENT.

Osservazione 4.0.2. Per loss function che sono due volte differenziabili possiamo considerare il vettore gradiente $g(\boldsymbol{\theta}) = \nabla \mathcal{L}(\boldsymbol{\theta})$ e la matrice Hessiana $H(\boldsymbol{\theta}) = \nabla^2 \mathcal{L}(\boldsymbol{\theta})$. Per un punto $\hat{\boldsymbol{\theta}} \in \mathbb{R}^D$, indicando $\hat{g} = g(\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}}$ e $\hat{H} = H(\boldsymbol{\theta})|_{\hat{\boldsymbol{\theta}}}$, abbiamo:

condizione necessaria: se $\hat{\boldsymbol{\theta}}$ è minimo locale, allora $\hat{g} = \mathbf{0}$ e $\hat{H}$ è definita semi-positiva;

condizione sufficiente: se $\hat{g} = \mathbf{0}$ e H è definita positiva, allora $\hat{\boldsymbol{\theta}}$ è un minimo locale.

Osservazione 4.0.3. A volte possono esserci vincoli sulla parametrizzazione (CONSTRAINED PARAMETRIZATION). I vincoli sono in genere categorizzati in equazioni e disequazioni, ad esempio:

$$h_k(\boldsymbol{\theta}) = 0 \text{ per } k \in \mathcal{E},$$
$$g_j(\boldsymbol{\theta}) \leq 0 \text{ per } j \in \mathcal{I}.$$

Allora il set di vincoli è:

$$\mathcal{C} = \{\boldsymbol{\theta} \text{ t.c. } h_k(\boldsymbol{\theta}) = 0,\ k \in \mathcal{E} \wedge g_j(\boldsymbol{\theta}) \leq 0,\ j \in \mathcal{I}\}$$

ed il parametro è dato da:

$$\hat{\theta} = \underset{\boldsymbol{\theta} \in \mathcal{C}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\theta}).$$

Quando non si hanno vincoli e quindi $\boldsymbol{\theta} \in \mathbb{R}^D$ si parla di UNCONSTRAINED PARAMETRIZATION.

Osservazione 4.0.4. Può essere utile determinare se la loss function è convessa. Ricordiamo quindi che se una funzione $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ è doppio differenziabile, allora questa è convessa se e solo se l'Hessiana è definita semi-positiva in ogni punto. f si dice invece strettamente convessa se l'Hessiana è definita positiva.

Osservazione 4.0.5. Nel caso in cui la loss function non sia continua (in un punto) allora si usa definire i **sottogradienti**. $\boldsymbol{g} \in \mathbb{R}^N$ è un sottogradiente di f in $\boldsymbol{x} \in \text{Dom}(f)$ se:

$$\forall \boldsymbol{\xi} \in \text{Dom}(f),\ f(\boldsymbol{\xi}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{\xi} - \boldsymbol{x}).$$

4.1 Metodi del primo ordine

Questi metodi si basano sulle derivate prime delle loss-function, ovvero guardano a quali sono le direzioni negative nello spazio dei parametri. Un punto di partenza deve essere specificato e viene chiamato $\boldsymbol{\theta}_0$. Per ogni iterazione t i parametri vengono aggiornati come:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \boldsymbol{d}_t, \tag{92}$$

ove η_t è detta STEP SIZE O LEARNING RATE, che in sequenza $\{\eta_t\}$ identifica il LEARNING RATE SCHEDULE. Mentre $\boldsymbol{d}_t$ è la DESCENT DIRECTION. Come descent direction si può usare la STEEPEST DESCENT:

$$\boldsymbol{g}_t = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_t} \quad \longrightarrow \quad \boldsymbol{d}_t = -\boldsymbol{g}_t. \tag{93}$$

Il gradiente $\boldsymbol{g}_t$ punta nella direzionedi massima variazione di $\mathcal{L}(\boldsymbol{\theta}_t)$ quindi questa scelta sembra ovvia.

L'algoritmo viene iterato fino a quando non si raggiunge un punto stazionario. Formalmente richiediamo che esista $\eta_{\max} > 0$ tale che $\mathcal{L}(\boldsymbol{\theta} + \eta \boldsymbol{d}) < \mathcal{L}(\boldsymbol{\theta})\ \forall \eta : 0 < \eta < \eta_{\max}$.

La scelta più semplice è $\eta = \text{const.}$, tuttavia se η è troppo grande il metodo potrebbe non convergere, mentre se è troppo piccolo la convergenza potrebbe essere troppo lenta. In genere è meglio scegliere η in modo adattivo, così da avere la massima riduzione della loss function lungo la direzione scelta:

$$\eta_t = \underset{\eta>0}{\operatorname{argmin}}\, \Phi_t(\eta) = \underset{\eta>0}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\theta}_t + \eta d\boldsymbol{t}). \tag{94}$$

Questo approccio è detto LINE SEARCH poichè cerchiamo lungo la direzione identificata da $d\boldsymbol{t}$.

Vogliamo inoltre usare algoritmi che convergono rapidamente. Nel caso la loss function sia convessa si può mostrare che la gradient descent converge con tasso lineare:

$$|\mathcal{L}(\boldsymbol{\theta}_{t+1})| \leq \mu\, |\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\hat{\boldsymbol{\theta}})|, \tag{95}$$

con $0 < \mu < 1$ il tasso di convergenza.

Il metodo di gradient descent può però essere molto insufficiente qualora ci si trovi in direzioni piatte. Una tecnica nota come METODO DEI MOMENTI o della palla pesata permette di risolvere questo problema. Questa prescrive di muoversi velocemente lungo direzioni piatte, in base alle iterazioni precedenti, e lentamente ove il gradiente cambia rapidamente. Si scrivono le quantità

$$\boldsymbol{m}_t = \beta \boldsymbol{m}_{t-1} + \boldsymbol{g}_{t-1} \quad \text{e} \quad \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t \boldsymbol{m}_t, \tag{96}$$

dove $\boldsymbol{m}_t$ è la *quantità di moto* e $0 < \beta < 1$, tipicamente $\beta \sim 0.9$. Sostituendo $\boldsymbol{m}_{t-1}$, poi $\boldsymbol{m}_{t-2}$ e così via, otteniamo una definizione per la quantità di moto come serie:

$$\boldsymbol{m}_t = \beta \boldsymbol{m}_{t-1} + \boldsymbol{g}_{t-1} = \beta^2 \boldsymbol{m}_{t-2} + \beta \boldsymbol{g}_{t-2} + \boldsymbol{g}_{t-1} = \sum_{\tau=0}^{t-1} \beta^\tau \boldsymbol{g}_{t-\tau-1}. \tag{97}$$

Se tutti i gradienti nel passato sono costanti, allota:

- $\boldsymbol{m}_t = \boldsymbol{g} \sum_{\tau=0}^{t-1} \beta^\tau$ e lo scaling factor è una serie geometrica;
- $\lim_{t\to\infty} \sum_{\tau=0}^{t-1} \beta^\tau = \frac{1}{1-\beta}$ e quindi se $\beta \sim 0.9$ moltiplichiamo il gradiente per 10.

4.2 Metodi del second'ordine

Il classico metodo alle derivate seconde è il METODO DI NEWTON. Questo consiste nell'aggiornare i parametri come:

(98)