Note sull'inferenza di parametri

Corso di Data Mining
a.a. 2024-2025

Giorgio Gambosi

Apprendimento supervisionato

Il nostro obiettivo è prevedere il valore sconosciuto di una caratteristica aggiuntiva, chiamata *target*, per un dato elemento $\mathbf{x}$, basandoci sui valori di un insieme di altre caratteristiche, dette **feature**. Questo compito di previsione assume due forme principali:

Regressione Quando il target è un valore reale $t \in \mathbb{R}$

Classificazione Quando il target è un valore discreto, appartenente a un insieme predefinito $t \in \{1, \ldots, K\}$

Per raggiungere questo obiettivo, adottiamo un approccio generale che prevede la definizione di un modello (funzionale o probabilistico) della relazione tra le feature e il target. Questo modello viene ottenuto attraverso un processo di apprendimento da un insieme di esempi, che illustrano la relazione tra l'insieme delle feature e il target. Gli esempi sono raccolti in un **training set** $\mathcal{T} = (\mathbf{X}, \mathbf{t})$, e ciascun esempio è composto da:

- Un vettore delle feature $\mathbf{x}_i = \{x_{i1}, \ldots, x_{im}\}$
- Il corrispondente valore target t_i

Il modello che costruiamo può assumere una delle seguenti due forme:

1. Una funzione $y(\cdot)$ che, per ogni elemento $\mathbf{x}$, restituisce un valore $y(\mathbf{x})$ come stima di t. Questa funzione agisce come un predittore diretto, mappando le feature in ingresso nello spazio del target.
2. Una distribuzione di probabilità che associa a ciascun possibile valore $\overline{y}$ nel dominio del target la sua probabilità corrispondente $p(y = \overline{y}|\mathbf{x})$. Questo approccio probabilistico fornisce una visione più articolata, catturando l'incertezza insita nel compito di previsione.

La scelta tra questi tipi di modelli dipende spesso dai requisiti specifici del problema, dalla natura dei dati e dal livello di interpretabilità desiderato per i risultati. L'approccio basato su funzione offre previsioni dirette e immediate, mentre il modello probabilistico fornisce una rappresentazione più ricca delle incertezze e degli esiti possibili.

In entrambi i casi, il modello funge da ponte tra le feature osservate e il target che vogliamo prevedere, sfruttando i pattern e le relazioni apprese dai dati di addestramento per effettuare previsioni informate su nuovi elementi non ancora visti.

Apprendimento di modelli funzionali

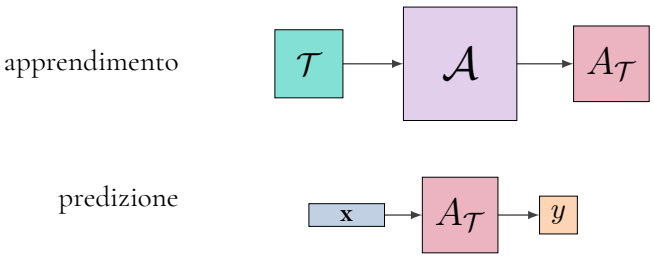
In questo case derivare un predittore implica l'apprendimento di una funzione da una classe predefinita. L'approccio può essere formalizzato come segue:

- Definire una classe di funzioni $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$, dove $\mathcal{X}$ e $\mathcal{Y}$ indicano il dominio delle features e del target rispettivamente
- Utilizzare un algoritmo di apprendimento $\mathcal{A}$ che, avendo in input il training set $\mathcal{T}$, deriva da esso una funzione specifica $h_{\mathcal{T}} \in \mathcal{H}$.

L'idea, qui, è che $\mathcal{A}$ trovi la funzione in $\mathcal{H}$ (in effetti, un algoritmo $A_{\mathcal{T}}$ che implementa questa funzione) che "meglio" predice t da $\mathbf{x}$ quando applicato agli esempi in $\mathcal{T}$, ossia predice al meglio t_i da $\mathbf{x}_i$ per ogni $(\mathbf{x}_i, t_i) \in \mathcal{T}$*.

Per ogni nuovo elemento $\mathbf{x}$, il valore da restituire come (stima del) target corrispondente viene calcolato come $h_{\mathcal{T}}(\mathbf{x})$, ossia applicando $A_{\mathcal{T}}$ all'input $\mathbf{x}$.

Tipicamente, $\mathcal{H}$ è un insieme di funzioni parametriche (indicato come $\mathcal{H}_{\boldsymbol{\theta}}$, dove $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ è l'insieme dei parametri) con la stessa struttura e che differiscono tra loro solo per i valori dei parametri $\boldsymbol{\theta}$. In questo caso, cercare una funzione in $\mathcal{H}_{\boldsymbol{\theta}}$ equivale a cercare un valore per $(\theta_1, \ldots, \theta_m)$.



Un semplice esempio di questo approccio è la regressione lineare, dove il valore predetto per l'elemento $\mathbf{x}$ viene calcolato come la combinazione lineare dei suoi valori di attributo $x_1, x_2, \ldots, x_d$, ciascuno pesato da un opportuno insieme di **pesi** costanti $w_1, w_2, \ldots, w_d$, più un **bias** b.

Cioè, la previsione viene calcolata come

$$y = \sum_{i=1}^{d} w_i x_i + b$$

Osserviamo che, in questo caso, l'insieme delle funzioni $\mathcal{H}$ è l'insieme delle funzioni lineari di dimensione $d+1$, parametrizzate dalla coppia rappresentata dal vettore $\mathbf{w} = (w_1, \ldots, w_d)$ dei pesi e dal bias b.

I 2 parametri $\mathbf{w}, b$, e quindi i loro $d+1$ valori , vengono **appresi** da $\mathcal{A}$ a partire dal training set $\mathcal{T}$.

La qualità di un predittore h viene valutata in termini di **rischio**, inteso come costo associato a un evento pesato dalla probabilità dell'evento stesso. Nel contesto considerato qui l'evento è rappresentato dalla necessità di effettuare una predizione relativa ad un particolare elemento $\mathbf{x} \in \mathcal{X}$, avente valore target dato dall'applicazione della funzione (sconosciuta) $f(\mathbf{x})$†. Il rischio relativo è dato da:

- il relativo costo, determinato dal confronto tra il valore predetto $h(\mathbf{x})$ e il valore corretto $f(\mathbf{x})$, quantificato mediante l'applicazione di una **funzione di loss** predefinita $L : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, ed è quindi dato da $L(h(\mathbf{x}), f(\mathbf{x}))$.
- la probabilità $p(\mathbf{x})$ che $\mathbf{x}$ sia l'elemento rispetto al quale effettuare la predizione. L'idea è quindi che esista una distribuzione predefinita di probabilità su $\mathcal{X}$ che assegna ad ogni possibile elemento una probabilità di comparire come elemento rispetto al quale effettuare una predizione. Si noti che la distribuzione p è però sconosciuta.

*Si noti che è necessario specificare cosa significhi "meglio" in questo contesto, ovvero quale misura di qualità della previsione si stia applicando.

†Si sta assumendo che, per ogni elemento $\mathbf{x}$ esiste uno specifico valore del target associato $t = f(\mathbf{x})$. Questa potrebbe essere un'ipotesi eccessivamente semplicistica in molti casi reali, visto che le feature sono solo un modello degli elementi reali, per cui elementi diversi, con valori diversi del target, potrebbero essere coincidenti rispetto alle feature utilizzate.

Il rischio relativo all'elemento $\mathbf{x}$, se la predizione è calcolata utilizzando la funzione h, è quindi

$$\mathcal{R}_{p,f}(\mathbf{x}; h) = L(h(\mathbf{x}), f(\mathbf{x})) \cdot p(\mathbf{x})$$

dove a pedice compaiono le caratteristiche (distribuzione p degli elementi, funzione "corretta" di predizione) sconosciute.

L'errore del predittore h, in quanto tale, indipendentemente dai singoli elementi, può essere definito in termini di costo atteso su tutti gli elementi in $\mathcal{X}$:

$$\mathcal{R}_{p,f}(h) \triangleq \mathop{\mathbb{E}}_{\mathbf{x}\sim p} [\, L(h(\mathbf{x}), f(\mathbf{x}))\,]$$

Nel caso della loss 0-1, cioé del contare quanti elementi sono classificati male in un problema di classificazione binaria, questo corrisponde semplicemente alla probabilità di una previsione errata su un elemento estratto casualmente, cioè:

$$\mathcal{R}_{p,f}(h) = \mathop{\mathbb{E}}_{\mathbf{x}\sim p} [\, \mathbb{1}[h(\mathbf{x}) \neq f(\mathbf{x})]\,] = \mathop{\mathbb{P}}_{\mathbf{x}\sim p} [\, h(\mathbf{x}) \neq f(\mathbf{x})\,]$$

dove $\mathbb{1}[h(\mathbf{x}) \neq f(\mathbf{x})]$ è pari ad 1 se $h(\mathbf{x}) \neq f(\mathbf{x})$ e 0 altrimenti.

Poiché p e f sono sconosciuti, il rischio può essere solo stimato dai dati disponibili (il training set $\mathcal{T}$). Questo porta alla definizione del **rischio empirico** $\overline{\mathcal{R}}_{\mathcal{T}}(h)$, che fornisce una stima del valore atteso della funzione loss come costo medio sul training set:

$$\overline{\mathcal{R}}_{\mathcal{T}}(h) = \frac{1}{|\mathcal{T}|} \sum_{(\mathbf{x},t)\in\mathcal{T}} L(h(\mathbf{x}), t)$$

Nel caso della loss 0-1, questo rappresenta la frazione di elementi di $\mathcal{T}$ che sono classificati erroneamente da h:

$$\overline{\mathcal{R}}_{\mathcal{T}}(h) = \frac{1}{|\mathcal{T}|} \sum_{(\mathbf{x},t)\in\mathcal{T}} \mathbb{1}[h(\mathbf{x}) \neq t] = \frac{|\{(\mathbf{x}, t) \in \mathcal{T} \mid h(\mathbf{x}) \neq t\}|}{|\mathcal{T}|}$$

In questo modo, un problema di apprendimento si riduce a un problema di minimizzazione in uno spazio funzionale $\mathcal{H}$, l'insieme di tutti i possibili predittori h (minimizzazione del rischio empirico).

$$h_{\mathcal{T}} = \mathop{\operatorname{argmin}}_{h\in\mathcal{H}} \overline{\mathcal{R}}_{\mathcal{T}}(h)$$

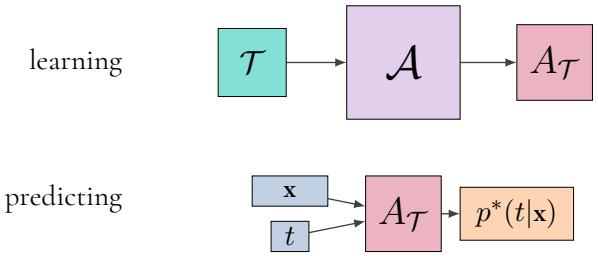
Apprendimento probabilistico

Nell'approccio probabilistico, assumiamo che la relazione tra elementi $\mathbf{x}$ e relativi target t non sia deterministica (mediante una funzione $f(\mathbf{x})$), ma che a un stesso $\mathbf{x}$, che può rappresentare elementi diversi nel mondo reale coincidenti sulle feature utilizzate, possano corrispondere valori target diversi. In questo caso, la relazione tra features e target è espressa mediante una distribuzione di probabilità condizionata $p(t|\mathbf{x})$, chiaramente sconosciuta, che vogliamo apprendere, almeno approssimativamente, a partire dal training set $\mathcal{T}$.

Così come nel caso non probabilistico, in cui vogliamo apprendere una funzione h^* che associa ad ogni elemento $\mathbf{x}$ un corrispondente valore $h^*(\mathbf{x})$ nel dominio target, operiamo definendo un universo di possibili distribuzioni caratterizzate da parametri, che differiscono tra loro per il valore dei parametri stessi, ed effettuiamo la ricerca di p^* all'interno di quell'universo.

Quindi,

1. consideriamo, per ipotesi, una classe di possibili distribuzioni condizionate $\mathcal{P}$ (un **modello probabilistico**) e
2. selezioniamo (inferiamo) la "migliore" distribuzione condizionata $p^* \in \mathcal{P}$ a partire dalla conoscenza disponibile (cioè, il training set), secondo una qualche misura q
3. dato un qualsiasi nuovo elemento $\mathbf{x}$, applicheremo $p^*(t|\mathbf{x})$ per assegnare probabilità a ciascun valore possibile del corrispondente target



Come definire la classe delle possibili distribuzioni condizionate $\mathcal{P}$? Di solito, viene utilizzato l'approccio parametrico: le distribuzioni sono definite da una struttura comune (arbitraria) e da un insieme di parametri.

Esempio: regressione logistica per la classificazione binaria

La probabilità $p(t|\mathbf{x})$, dove $t \in \{0, 1\}$, si assume essere una distribuzione di Bernoulli

$$p(t|\mathbf{x}) = \begin{cases} \pi(\mathbf{x}) & \text{se} \quad t = 1 \\ 1 - \pi(\mathbf{x}) & \text{se} \quad t = 0 \end{cases}$$

che può essere scritta in modo equivalente come

$$p(t|\mathbf{x}) = \pi(\mathbf{x})^t (1 - \pi(\mathbf{x}))^{1-t}$$

come si vede, la probabilità dipende dall'elemento, in quanto funzione $\pi(\mathbf{x})$ di $\mathbf{x}$.

Ipotizziamo ora che $\pi(\mathbf{x})$ derivi da una combinazione lineare delle features di $\mathbf{x}$ attraverso l'applicazione della funzione sigmoide, che proietta il valore reale, ottenuto per mezzo della combinazione lineare, sull'intervallo $[0, 1]$ su cui è definita una probabilità[‡]. Questo fa sì che la distribuzione sia ora parametrica rispetto alla coppia $\mathbf{w}, b$:

$$\pi(\mathbf{x}; \mathbf{w}, b) = \sigma\left(\mathbf{w}^T\mathbf{x} + b\right) = \sigma\left(\sum_{i=1}^{d} w_i x_i + b\right) = \frac{1}{1 + e^{-(\sum_{i=1}^{d} w_i x_i + b)}}$$

Quel che si richiede ora è calcolare i valori dei parametri $\mathbf{w}, b$ che minimizzano il rischio empirico calcolato su $\mathcal{T}$ facendo riferimento alla funzione di costo, ancora da definire.

Inferire una distribuzione ottimale

Qual è una misura $q(p, \mathcal{T})$ della qualità della distribuzione condizionata parametrica $p(t|\mathbf{x}; \boldsymbol{\theta})$, dato il training set $\mathcal{T} = (\mathbf{X}, \mathbf{t})$? Possiamo evidentemente interpretare q come l'inverso della funzione di loss e quindi massimizzarla al variare dei parametri.

- Idea: fare riferimento a quanto un dataset generato campionando casualmente n coppie (dove $n = |\mathbf{X}|$) $(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)$ secondo la distribuzione congiunta $p(\mathbf{x}, t; \boldsymbol{\theta})$, potrebbe essere simile al training set disponibile $\mathcal{T}$.§

[‡]In effetti la funzione sigmoide ha come codominio l'intervallo aperto $(0, 1)$, ma questo cambia poco, in quanto i valori 0 e 1 per una probabilità, soprattutto se stimata come nel caso attuale sono tipicamente non considerati, corrispondendo a situazioni di certezza.

§Le medesime considerazioni che verranno fatte nel seguito potrebbero esserlo anche nel caso in cui il dataset comprendesse soltanto un insieme di elementi $\mathbf{X}$, senza alcun target associato (apprendimento non supervisionato). In tal caso, le considerazioni che qui verranno fatte riguardo alla distribuzione congiunta $p(\mathbf{x}, t; \boldsymbol{\theta})$ da apprendere verrebbero effettuate rispetto all'apprendimento della distribuzione dei soli elementi $p(\mathbf{x}; \boldsymbol{\theta})$.

- In particolare, possiamo considerare la probabilità che il dataset ottenuto in questo modo corrisponda al traing set $\mathcal{T} = (\mathbf{X}, \mathbf{t})$, vale a dire la probabilità che proprio $\mathcal{T}$ sia ottenuto mediante il procedimento seguente, basato sull'identità $p(\mathbf{x}, t; \boldsymbol{\theta}) = p(\mathbf{x}; \boldsymbol{\theta})p(t|\mathbf{x}; \boldsymbol{\theta})$
  - $n = |\mathbf{X}|$ coppie $\mathbf{x}_i, t_i$ sono estratte indipendentemente l'una dall'altra. Per ogni valore i:
    1. $\mathbf{x}_i$ è estratto casualmente su $\mathcal{X}$ secondo la distribuzione $p(\mathbf{x}; \boldsymbol{\theta})$ (per semplicità tipicamente assunta indipendente da $\boldsymbol{\theta}$ e, in particolare, uniforme)
    2. t_i è campionato da $p(t|\mathbf{x}_i; \boldsymbol{\theta})$
  - si noti che, applicando questa procedura, le n coppie $\mathbf{x}_i, t_i$ sono generate sotto l'ipotesi di essere **indipendenti** e **identicamente distribuite**, con probabilità $p(\mathbf{x}, t; \boldsymbol{\theta}) = p(\mathbf{x})p(t|\mathbf{x}; \boldsymbol{\theta})$. Questo comporta che la probabilità dell'intero training set $p(\mathbf{X}, \mathbf{t}; \boldsymbol{\theta})$ può essere espressa come prodotto delle probabilità dei suoi singoli elementi, calcolate con riferimento alla stessa distribuzione

$$p(\mathbf{X}, \mathbf{t}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i, t_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(t_i|\mathbf{x}_i; \boldsymbol{\theta})p(\mathbf{x}_i) \propto \prod_{i=1}^{n} p(t_i|\mathbf{x}_i) = q(p, \mathcal{T}; \boldsymbol{\theta})$$

- Possiamo usare tale probabilità come misura della qualità $q(p, \mathcal{T})$ e cercare la distribuzione $p(t|\mathbf{x}; \boldsymbol{\theta}^*)$ che rende massima $p(\mathbf{X}, \mathbf{t}; \boldsymbol{\theta})$ assumendo che $p(\mathbf{x})$ sia la distribuzione uniforme e $p(t|\mathbf{x}; \boldsymbol{\theta})$ sia la distribuzione parametrica condizionata. Questa probabilità, essendo il tipo di distribuzione condizionata, così come il training set, fissati, è funzione dei parametri $\boldsymbol{\theta}$ e possiamo cercare di trovare il valore $\boldsymbol{\theta}^*$ per cui è massima. Cerchiamo quindi, in definitiva, (all'interno della classe delle distribuzioni parametriche $p^*(t|\mathbf{x}; \boldsymbol{\theta})$ la probabilità condizionata $p^*(t|\mathbf{x}; \boldsymbol{\theta}^*)$ che rende massima $p(\mathbf{X}, \mathbf{t}; \boldsymbol{\theta})$.

Osserviamo che apprendere la distribuzione $p(t|\mathbf{x}; \boldsymbol{\theta}^*)$, che massimizza $q(p, \mathcal{T}; \boldsymbol{\theta})$, corrisponde, nel caso di un predittore probabilistico, ad apprendere la funzione h^* che minimizza il rischio empirico $\overline{\mathcal{R}}_{\mathcal{T}}(h)$ nel caso di un predittore funzionale. In entrambi i casi, l'apprendimento viene eseguito tramite ottimizzazione.

Modelli probabilistici

Un **modello probabilistico** è una collezione di distribuzioni di probabilità con la stessa struttura, definite sul dominio dei dati. Le distribuzioni di probabilità sono istanze del modello probabilistico e sono caratterizzate dai valori assunti da un insieme di **parametri**.

In un modello probabilistico gaussiano bivariato, le distribuzioni sono caratterizzate dai valori assunti da:

1. la media $\boldsymbol{\mu} = (\mu_1, \mu_2)$
2. la matrice di covarianza $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$

dove $\sigma_{12} = \sigma_{21}$

Dato un dataset $\mathcal{T}$ e una distribuzione di probabilità p con parametri $\boldsymbol{\theta}$, l'espressione $p(\mathcal{T}|\boldsymbol{\theta})$ può essere intesa in due modi diversi:

- possiamo considerare $\boldsymbol{\theta}$ come fissato e $\mathcal{T}$ variabile, $p(\mathcal{T}|\boldsymbol{\theta})$; questo è il modo convenzionale di interpretare l'espressione
- possiamo considerare $\boldsymbol{\theta}$ come variabile e $\mathcal{T}$ fissato, $p(\mathcal{T}|\boldsymbol{\theta})$; in questo caso parliamo di **verosimiglianza** di $\boldsymbol{\theta}$ rispetto a $\mathcal{T}$. Possiamo interpretare questo valore come la probabilità che, estraendo a caso un dataset di dimensione $|\mathcal{T}|$ secondo la distribuzione con parametri $\boldsymbol{\theta}$, si ottenga esattamente $\mathcal{T}$. In questo caso possiamo ad esempio cercare il valore di $\boldsymbol{\theta}$ per cui la probabilità $p(\mathcal{T}|\boldsymbol{\theta})$ dei dati osservati è massima ovvero la probabilità del dataset sotto la distribuzione p con parametri $\boldsymbol{\theta}$, ossia il fatto che il dataset sia generato campionando indipendentemente punti da $p(\mathbf{x}, t; \boldsymbol{\theta})$.

Come visto, assumendo che gli elementi in $\mathcal{T}$ siano i.i.d.,

$$p(\mathbf{X}, \mathbf{t}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i, t_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) p(\mathbf{x}_i|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta})$$
$$= p(\mathbf{x}) \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) \qquad \text{assumendo } p(\mathbf{x}) \text{ uniforme}$$

Stima di massima verosimiglianza

Punto di vista frequentista: i parametri sono variabili deterministiche, il cui valore è sconosciuto e deve essere stimato.

Vogliamo, come detto, determinare il valore che massimizza la verosimiglianza

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{X}, \mathbf{t}|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{x}) \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \boldsymbol{\theta})$$

Il logaritmo di questa funzione

$$\log p(\mathcal{T}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i, t_i|\boldsymbol{\theta})$$

detto log-verosimiglianza, è generalmente preferibile da massimizzare, poiché i prodotti vengono trasformati in somme, mentre $\boldsymbol{\theta}^*$ rimane invariato (poiché il logaritmo è una funzione monotona), ovvero

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \log p(\mathcal{T}|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{T}|\boldsymbol{\theta})$$

Il problema di ottimizzazione risultante è quindi

$$\boldsymbol{\theta}^*_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{X}, \mathbf{t}|\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(t_i|\mathbf{x}_i, \boldsymbol{\theta})$$

Una soluzione viene calcolata risolvendo il sistema di equazioni

$$\frac{\partial l(\boldsymbol{\theta}; \mathcal{T})}{\partial \theta_k} = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_k} \log p(t_i|\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{p(t_i|\mathbf{x}_i; \boldsymbol{\theta})} \frac{\partial}{\partial \theta_k} p(t_i|\mathbf{x}_i; \boldsymbol{\theta}) = 0 \qquad k = 1, \ldots, |\boldsymbol{\theta}|$$

più concisamente, ponendo il gradiente uguale a 0

$$\nabla \log p(\mathcal{T}|\boldsymbol{\theta}) = \mathbf{0}$$

Si noti che la condizione di gradiente nullo è solo una condizione necessaria per la massimizzazione della funzione di massima verosimiglianza considerata, poiché in questo caso possiamo solo affermare che il punto corrispondente è un punto stazionario (cioè un massimo, un minimo o un punto di sella). Anche nel caso in cui il punto sia un massimo (che potrebbe essere verificato stimando la seconda derivata o, in generale, l'Hessiano), possiamo concludere che si tratta di un massimo **locale**, mentre siamo interessati al massimo globale.

Queste problematiche vengono tipicamente affrontate sia considerando casi in cui, ad esempio, esiste un solo punto stazionario ed esso è un massimo (e quindi anche il massimo globale), sia applicando strategie di ricerca del massimo più complesse.

Una volta calcolato l'ottimo $\boldsymbol{\theta}^*$, è possibile effettuare predizioni stimando, per ogni nuova osservazione $\mathbf{x}$, la probabilità del suo target, condizionata dal fatto che conosciamo il dataset di esempi $\mathbf{X}, \mathbf{t}$.

Applicando la proprietà $p(x) = \int_y p(x,y)dy = \int p(x|y)p(y)dy$, possiamo scrivere

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}) = \int_{\boldsymbol{\theta}} p(t|\mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{t}) d\boldsymbol{\theta} \approx \int_{\boldsymbol{\theta}} p(t|\mathbf{x}, \boldsymbol{\theta}^*) p(\boldsymbol{\theta}|\mathbf{X}) d\boldsymbol{\theta} = p(\mathbf{x}|\boldsymbol{\theta}^*) \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{t}) d\boldsymbol{\theta} = p(t|\mathbf{x}, \boldsymbol{\theta}^*)$$

Collezione $\mathbf{X}$ di n eventi binari, modellati tramite una distribuzione di Bernoulli con parametro sconosciuto ϕ

$$p(x|\phi) = \phi^x (1-\phi)^{1-x}$$

Verosimiglianza:

$$p(\mathbf{X}|\phi) = \prod_{i=1}^{n} \phi^{x_i} (1-\phi)^{1-x_i}$$

Log-verosimiglianza:

$$\begin{aligned} \log p(\mathbf{X}|\phi) &= \sum_{i=1}^{n} \left(x_i \log \phi + (1-x_i) \log(1-\phi) \right) \\ &= \log \phi \sum_{i=1}^{n} x_i + \log(1-\phi) \sum_{i=1}^{n} (1-x_i) = n_1 \log \phi + n_0 \log(1-\phi) \end{aligned}$$

dove n_0 (n_1) è il numero di eventi $x \in \mathbf{X}$ uguali a 0 (1)

$$\frac{\partial}{\partial \phi} \log p(\mathbf{X}|\phi) = \frac{n_1}{\phi} - \frac{n_0}{1-\phi} = 0 \qquad \text{e quindi} \qquad \phi^* = \frac{n_1}{n_0 + n_1} = \frac{n_1}{n}$$

Regressione lineare: collezione $\mathbf{X}, \mathbf{t}$ di coppie valore-target, modellata come $p(\mathbf{x}, t) = p(\mathbf{x}) p(t|\mathbf{x}; \mathbf{w}, b)$, con $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$, assumendo

- $p(\mathbf{x})$ uniforme
- $p(t|\mathbf{x}; \mathbf{w}, b)$ gaussiana con varianza costante e media $\mu(\mathbf{x}) = \sum_{k=1}^{d} x_i w_i + b$: quindi $p(t|\mathbf{x}; \mathbf{w}, b, \sigma) = \mathcal{N}(\mu(\mathbf{x}), \sigma^2)$

Verosimiglianza:

$$p(\mathbf{t}|\mathbf{X}; \mathbf{w}, b, \sigma) = \prod_{i=1}^{n} p(t_i|\mathbf{x}_i; \mathbf{w}, b, \sigma) = \prod_{i=1}^{n} \mathcal{N}(\mu(\mathbf{x}_i), \sigma^2)$$

Log-verosimiglianza:

$$\begin{aligned} \log p(\mathbf{t}|\mathbf{X}; \mathbf{w}, b, \sigma) &= \sum_{i=1}^{n} \log p(t_i|\mathbf{x}_i; \mathbf{w}, b, \sigma) = \sum_{i=1}^{n} \log \left(\sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\mu(\mathbf{x}_i) - t_i)^2} \right) \\ &= \sum_{i=1}^{n} \left(-\frac{1}{2\sigma^2} \left(\mu(\mathbf{x}_i) - t_i \right)^2 - \log \sigma - \frac{1}{2} \log(2\pi) \right) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mu(\mathbf{x}_i) - t_i)^2 - n \log \sigma - \frac{n}{2} \log(2\pi) \end{aligned}$$

$$\frac{\partial}{\partial w_k}\log p(\mathbf{t}|\mathbf{X},\mathbf{w},b,\sigma) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mu(\mathbf{x}_i)-t_i)x_{ik} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\sum_{k=1}^{d}x_{ik}w_i+b-t_i\right)x_{ik} \qquad k=1,\dots,d$$

$$\frac{\partial}{\partial b}\log p(\mathbf{t}|\mathbf{X},\mathbf{w},b,\sigma) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mu(\mathbf{x}_i)-t_i) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\sum_{k=1}^{d}x_{ik}w_i+b-t_i\right)$$

La stima ML per $\mathbf{w}, b$ (coefficenti della regressione lineare) è ottenuta come soluzione del sistema lineare $(d+1, d+1)$

$$\sum_{i=1}^{n}\left(\sum_{k=1}^{d}x_{ik}w_i+b-t_i\right)x_{ik} = 0 \qquad k=1,\dots,d$$

$$\sum_{i=1}^{n}\left(\sum_{k=1}^{d}x_{ik}w_i+b-t_i\right) = 0$$

in cui, si ricorda, le variabili sono $w_0,\dots,w_d,b$ e le costanti sono x_{ik} $(i=1,\dots,n,\ k=1,\dots,d)$ e t_i $(i=1,\dots,n)$
Lo stesso approccio può essere, volendo, applicato anche per stimare il valore migliore di σ, ottenendo

$$\sigma^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{d}x_{ik}w_i+b-t_i\right)^2}$$

Stima di massima verosimiglianza per logistic regression

Assumiamo che i target degli elementi possano essere modellati condizionatamente (rispetto ai coefficienti del modello) tramite una distribuzione di Bernoulli. Ovvero, assumiamo

$$p(t_i|\mathbf{x}_i;\mathbf{w},b) = \begin{cases} p_i & \text{se } t_i = 1 \\ 1-p_i & \text{se } t_i = 0 \end{cases}$$

e quindi

$$p(t_i|\mathbf{x}_i;\mathbf{w},b) = p_i^{t_i}(1-p_i)^{1-t_i}$$

La logistic regression stima $p_i = p(C_1|\mathbf{x}_i)$ come $p_i = \sigma(y_i)$, dove $y_i = \sum_{k=1}^{d}x_{ik}w_i + b$, applicando quindi la funzione sigmoide al risultato della combinazione lineare delle features dell'elemento.¶

La verosimiglianza dei target del training set $\mathbf{t}$, dato $\mathbf{X}$ è

$$p(\mathbf{t}|\mathbf{X};\mathbf{w},b) = \prod_{i=1}^{n}p(t_i|\mathbf{x}_i;\mathbf{w},b) = \prod_{i=1}^{n}p_i^{t_i}(1-p_i)^{1-t_i}$$

e la log-verosimiglianza è

$$\begin{aligned}\log p(t|\mathbf{x};\mathbf{w},b) &= \sum_{i=1}^{n}(t_i\log p_i + (1-t_i)\log(1-p_i)) \\ &= \sum_{i=1}^{n}(t_i\log\sigma(y_i) + (1-t_i)\log(1-\sigma(y_i))) \\ &= \sum_{i=1}^{n}\left(t_i\log\sigma\left(\sum_{k=1}^{d}x_{ik}w_i+b\right) + (1-t_i)\log\left(1-\sigma\left(\sum_{k=1}^{d}x_{ik}w_i+b\right)\right)\right)\end{aligned}$$

¶Si può dimostrare che utilizzare la sigmoide implica la proprietà che la combinazione lineare y_i è pari al logaritmo del rapporto tra le probabilità (stimate) di appartenenza alle due classi, cioè che $y_i = \log\frac{p_i}{1-p_i}$

Utilizzando le proprietà generali delle derivate (in particolare relativamente alla derivazione di funzioni composte) e quella relativa alla funzione sigmoide, per la quale $\frac{d\sigma(x)}{dx} = \sigma(x)(1-\sigma(x))$ è possibile mostrare che

$$\frac{\partial}{\partial w_j} \log p(\mathbf{t}|\mathbf{X}; \mathbf{w}, b) = \sum_{i=1}^{n} (t_i - p_i) x_{ij}$$

e

$$\frac{\partial}{\partial b} \log p(\mathbf{t}|\mathbf{X}; \mathbf{w}, b) = \sum_{i=1}^{n} (t_i - p_i)$$

Questo sistema è non lineare, data la presenza della funzione σ, per cui non è risolubile analiticamente come quello risultante per la regressione lineare.

Per massimizzare la log verosimiglianza, possiamo allora applicare un algoritmo di ascesa (non discesa, in quanto vogliamo massimizzare) del gradiente, dove ad ogni iterazione viene eseguito il seguente aggiornamento della stima attuale:

$$\begin{aligned} w_k^{(j+1)} &= w_k^{(j)} + \eta \sum_{i=1}^{n} (t_i - p^{(j)}) x_{ik} \\ &= w_k^{(j)} + \eta \sum_{i=1}^{n} \left(t_i - \sigma \left(\sum_{r=1}^{d} w_r^{(j)} x_{ir} + b \right) \right) x_{ik} \end{aligned}$$

per $k = 1, \ldots, d$ e

$$\begin{aligned} b^{(j+1)} &= b^{(j)} + \eta \sum_{i=1}^{n} (t_i - p^{(j)}) \\ &= b^{(j)} + \eta \sum_{i=1}^{n} \left(t_i - \sigma \left(\sum_{r=1}^{d} w_r^{(j)} x_{ir} + b \right) \right) \end{aligned}$$

Stima di massimo a posteriori

L'inferenza tramite massimo a posteriori (MAP) è simile alla stima di massima verosimiglianza (ML), ma $\boldsymbol{\theta}$ è ora considerato come una variabile casuale (approccio bayesiano), a cui quindi è associata una distribuzione di probabilità $p(\boldsymbol{\theta})$. In considerazione della conoscenza del training set, la distribuzione di interesse è in effetti la **distribuzione a posteriori** $p(\boldsymbol{\theta}|\mathcal{T})$, a posteriori, per l'appunto della conoscenza di $\mathcal{T} = \{\mathbf{X}, \mathbf{t}\}$. Il valore stimato per $\boldsymbol{\theta}$ potrebbe a questo punto essere il valore per il quale tale distribuzione è massima, per cui

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\mathcal{T})$$

questo valore prende il nome di MAP (**maximum a posteriori**).

Uno strumento utile per determinare la distribuzione a posteriori è la **regola di Bayes**

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

che nel contesto considerato diventa

$$p(\boldsymbol{\theta}|\mathcal{T}) = \frac{p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{T})}$$

Le distribuzioni presenti nell'espressione precedente sono indicate nel modo seguente:

1. $p(\boldsymbol{\theta}|\mathcal{T})$ è la distribuzione a posteriori di $\boldsymbol{\theta}$ conoscendo il valore di $\mathcal{T}$
2. $p(\boldsymbol{\theta})$ è la distribuzione a priori di $\boldsymbol{\theta}$ (a priori rispetto alla conoscenza del valore di $\mathcal{T}$)
3. $p(\mathcal{T}|\boldsymbol{\theta})$ è la verosimiglianza, vale a dire la probabilità di $\mathcal{T}$ conoscendo il valore di $\boldsymbol{\theta}$
4. $p(\mathcal{T})$ che - si noti - è indipendente da $\boldsymbol{\theta}$, è detto evidenza, ed è la probabilità del dataset, assumendo che il valore di $\boldsymbol{\theta}$ non sia noto. Questo valore è dato dal valore atteso dell'evidenza al variare di $\boldsymbol{\theta}$, assumendo che $\boldsymbol{\theta}$ sia distribuito secondo $p(\boldsymbol{\theta})$, vale a dire $p(\mathcal{T}) = \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta})}[\,p(\mathcal{T}|\boldsymbol{\theta})\,] = \int p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$

Si noti che

$$\boldsymbol{\theta}^* = \mathop{\text{argmax}}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{T}) = \mathop{\text{argmax}}_{\boldsymbol{\theta}} p(\mathcal{T}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \mathop{\text{argmax}}_{\boldsymbol{\theta}} \left(\log p(\mathcal{T}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right)$$

che porta a

$$\boldsymbol{\theta}^* = \mathop{\text{argmax}}_{\boldsymbol{\theta}} \left(\sum_{i=1}^{n} \log p(t_i|\mathbf{x}_i, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \right)$$

Collezione $\mathbf{X}$ di n eventi binari, modellata come una distribuzione di Bernoulli con parametro sconosciuto ϕ. La conoscenza iniziale di ϕ è modellata come una distribuzione Beta:

$$p(\phi|\alpha,\beta) = \text{Beta}(\phi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$

Log-verosimiglianza

$$\log p(\mathbf{X}|\phi) = \sum_{i=1}^{n} \left(x_i \log\phi + (1-x_i)\log(1-\phi)\right) = n_1 \log\phi + n_0 \log(1-\phi)$$

$$\frac{\partial}{\partial\phi}\left(\log p(\mathbf{X}|\phi) + \log \text{Beta}(\phi|\alpha,\beta)\right) = \frac{n_1}{\phi} - \frac{n_0}{1-\phi} + \frac{\alpha-1}{\phi} - \frac{\beta-1}{1-\phi} = 0 \qquad \Longrightarrow$$

$$\phi^* = \frac{N_1+\alpha-1}{n_0+n_1+\alpha+\beta-2} = \frac{n_1+\alpha-1}{n+\alpha+\beta-2}$$

La funzione

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

è un'estensione del fattoriale al campo dei numeri reali: infatti, per qualsiasi intero x,

$$\Gamma(x) = (x-1)!$$

Derivazione della distribuzione a posteriori

Una stima migliore può essere ottenuta derivando la distribuzione a posteriori di $\boldsymbol{\theta}$ dato il dataset $\mathbf{X}, \mathbf{t}$. Un valore $\boldsymbol{\theta}^*$ può a quel punto essere calcolato come il valore atteso del parametro rispetto alla distribuzione a posteriori $p(\boldsymbol{\theta}|\mathcal{T})$,

$$\boldsymbol{\theta}^* = E_{p(\boldsymbol{\theta}|\mathcal{T})}[\boldsymbol{\theta}] = \int_{\boldsymbol{\theta}} \boldsymbol{\theta} \cdot p(\boldsymbol{\theta}|\mathcal{T}) d\boldsymbol{\theta}$$

Collezione X di n eventi binari, modellati come una distribuzione di Bernoulli con parametro sconosciuto ϕ. La conoscenza iniziale di ϕ è modellata come una distribuzione Beta:

$$p(\phi|\alpha,\beta) = \mathsf{Beta}(\phi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$

Distribuzione a posteriori

$$\begin{aligned} p(\phi|\mathrm{X},\alpha,\beta) &= \frac{\prod_{i=1}^{N}\phi^{x_i}(1-\phi)^{1-x_i}p(\phi|\alpha,\beta)}{p(\mathrm{X})} \\ &= \frac{\phi^{N_1}(1-\phi)^{N_0}\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}p(\mathrm{X})} = \frac{\phi^{N_1+\alpha-1}(1-\phi)^{N_0+\beta-1}}{Z} \end{aligned}$$

Pertanto,

$$p(\phi|\mathrm{X},\alpha,\beta) = \mathsf{Beta}(\phi|\alpha+N_1,\beta+N_0)$$

In linea di principio, la conoscenza di $p(\boldsymbol{\theta}|\mathcal{T})$ permette di ottenere direttamente la distribuzione predittiva

$$p(t|\mathrm{x},\mathcal{T}) = \int p(t,\boldsymbol{\theta}|\mathrm{x},\mathcal{T})d\boldsymbol{\theta} = \int p(t|\boldsymbol{\theta},\mathrm{x},\mathcal{T})p(\boldsymbol{\theta}|\mathrm{x},\mathcal{T})d\boldsymbol{\theta} = \int p(t|\mathrm{x},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{T})d\boldsymbol{\theta}$$

che deriva dalla proprietà $p(x|z) = \int_y p(x,y|z)dy$ e dall'ipotesi che, dato l'elemento $\mathbf{x}$ e i parametri $\boldsymbol{\theta}$, il valore stimato per t sia indipendente dal training set $\mathcal{T}$ (in quanto l'informazione portata da esso è stata rappresentata, nella fase di apprendimento, nei valori di θ) e che il valore dei parametri del modello dipende solo dal training set (e chiaramente non dagli elementi rispetto ai quali andrà successivamente effettuata la predizione).