Classificazione Supervisionata: L'approccio discriminativo - 21/11

Intro

Noi non abbiamo la distribuzione a posteriori che serve a fare il MAP, abbiamo un training set e dobbiamo stimare la posteriori, otterremo un surrogato che sarà parametrico, quindi avremo dei β che sono parametri che descrivono questa funzione.

Vogliamo scegliere i β che danno la migliore approssimazione.

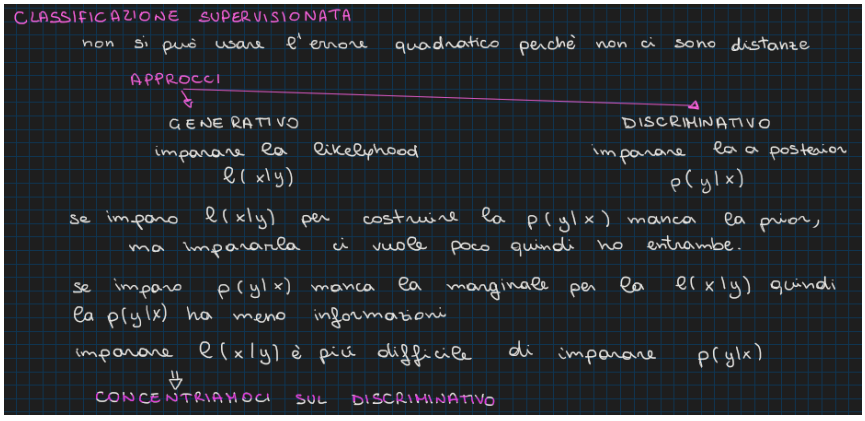
Scegliamo una forma funzionale $p_\beta(Y|X)$, al variare di β si ottengono diverse soluzioni per questa funzione e noi vogliamo scegliere la migliore, per stabilire la migliore serve un criterio, nella regressione si minimizzava l'RSS, ora l'RSS (in generale l'errore quadratico) non ha senso perché non hanno senso le distanze.

Un approccio potrebbe essere che visto che l'MSE si usava per la regressione, mentre in classificazione l'indice di performance è la probabilità d'errore, invece di minimizzare l'MSE quando si usa MAP bisogna minimizzare la probabilità d'errore.

Purtroppo però la minimizzazione della probabilità d'errore è in generale non risolvibile, calcolare la probabilità d'errore empirica da luogo a problemi di ottimizzazione che non si risolvono, quindi si **cambia approccio**.

Gli approcci che vanno per la maggiore sono due:

- approccio **generativo**
  - voglio imparare, dal training set, la likelihood
- approccio **discriminativo**
  - voglio imparare, dal training set, direttamente la funzione che serve a discriminare cioè la posterior



L’approccio discriminativo è più diretto, visto che ci serve la posterior facciamo direttamente quello invece di passare prima dalla likelihood.

Si nota però che l’approccio generativo, una volta imparata anche la prior che è banale, permette di ricavare la posterior. Al contrario per trovare la likelihood dopo aver ottenuto la posterior nell’approccio discriminativo ci serve anche $p(x)$ (la marginale), il che non è fattibile. Quindi l’approccio discriminativo in un certo senso da meno informazioni.

Concettualmente perché è di solito più difficile imparare $l(X|Y)$? Perché la likelihood si ottiene con la congiunta di tutte le X e in problemi ad alta dimensionalità diventa, per ovvie ragioni, molto complicato.

Se le variabili $X_1, X_2, \ldots, X_n$ sono **indipendenti**, allora la distribuzione congiunta può essere scritta come il prodotto delle distribuzioni marginali, cioè:

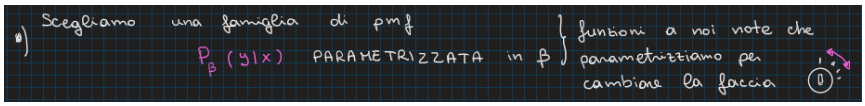
$$P(X_1, X_2, \ldots, X_n) = P(X_1) \cdot P(X_2) \cdot \cdots \cdot P(X_n)$$

In questo caso, il calcolo è semplice e non comporta un'esplosione combinatoria. Tuttavia, **se le variabili non sono indipendenti**, non puoi semplificare così facilmente. In presenza di dipendenze tra le variabili, devi considerare le **distribuzioni condizionate**, e la formula della congiunta diventa:

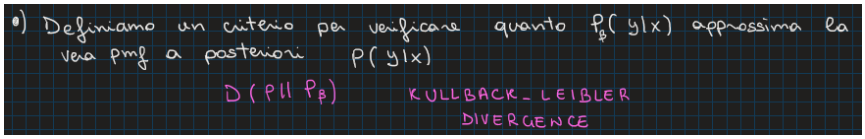
$$P(X_1, X_2, \ldots, X_n) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdot \cdots \cdot P(X_n|X_1, X_2, \ldots, X_{n-1})$$

Questo richiede di calcolare e memorizzare tutte le probabilità condizionate, e quindi il calcolo diventa esponenzialmente più complesso con l'aumentare delle variabili, proprio perché la dipendenza tra le variabili aumenta la quantità di informazioni da considerare.

L’approccio discriminativo



Divergenza di Kullback-Leibler



Impropriamente è anche chiamata “distanza”, impropriamente perché non rispetta le proprietà di una distanza (ad esempio che la distanza tra A e B è uguale alla distanza tra B ed A).

Anche chiamata entropia relativa.

Divergenza di Kullback-Leibler

Date due 2 pmf, p e q, definite sullo stesso alfabeto A:

$$D(p||q) \triangleq \sum_{z \in A} p(z) ln(\frac{p(z)}{q(z)})$$

Questo oggetto è tale che vale 0 solo se le due pmf sono uguali, noi vogliamo trovare un β tale da minimizzare questo oggetto.

Mettendo ln la base è e, l'unità di misura è nat.

Un nat (a volte detto anche nit o nepit) è un'unità logaritmica di informazione o e ntropia, basata su logaritmi naturali e potenze di e, invece che su potenze di 2 e logaritmi in base 2 che definiscono il bit. Il nat è l'unità naturale per l'entropia dell'informazione.

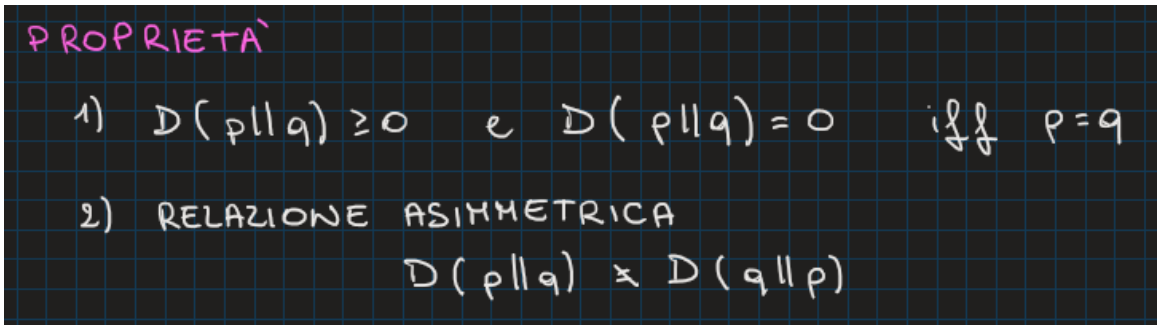
Se cambiassimo base del logaritmo questo corrisponderebbe alla moltiplicazione per un fattore, il significato fisico di usare un logaritmo in una certa base è solo quello di cambiare l'unità di misura di una certa metrica. Siccome queste metriche nascono nell'ambito della teoria dell'informazione per i sistemi di comunicazione, usando un logaritmo in base 2 stiamo misurando in bit. In comunicazioni la scelta è sempre il bit ma i nat sono migliori per noi perché potrebbero uscirci esponenziali visto che trattiamo distribuzioni e la cosa ci faciliterà.

Situazioni particolari

Riscontriamo che la divergenza vale 0 se le due pmf sono uguali. Ma vale 0 se e solo se sono uguali?

Se $p(z) = 0$ ci troviamo in una situazione particolare perché il logaritmo è definito solo per valori positivi, tuttavia per un limite notevole possiamo dire che è come se non lo calcolassimo e visto che è moltiplicato per 0 fa tutto 0: $0ln(0) = 0$.

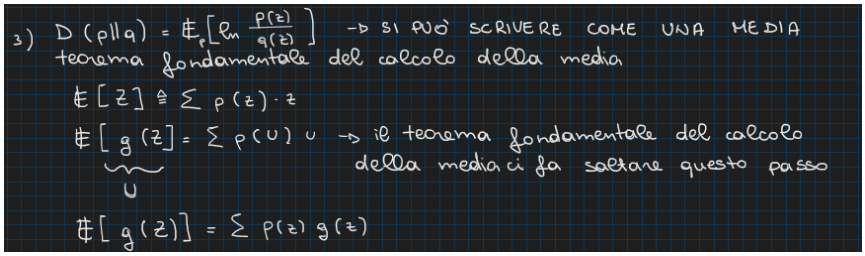
Se $q(z) = 0$ invece $ln(\frac{1}{0}) = +\infty$ e questo ha un significato fisico, se un certo simbolo porta ad avere una certa probabilità in p mentre 0 in q allora le due distribuzioni sono molto lontane. Ogni volta che vedo quel simbolo so che è p ad averlo creato e non q.



Domanda interessante

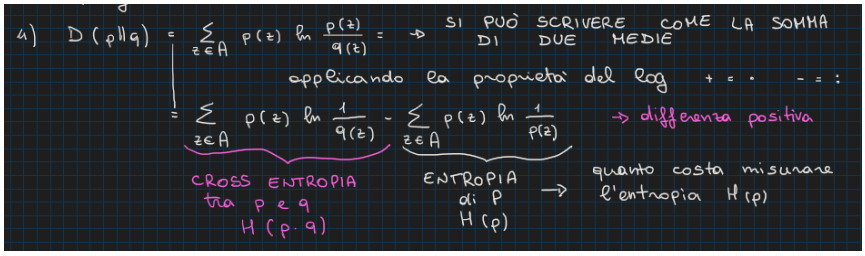
Sapendo che la divergenza è asimmetrica, chi gioca tra p e q il ruolo della "distribuzione vera" e chi il ruolo della "distribuzione approssimata"?

Il fatto che si scrive come sommatoria di p per qualcosa ci ricorda la media.



Si può esprimere come una media

Visto che è quella rispetto alla quale valuto la media la distribuzione vera deve essere p.



Si può decomporre come la differenza di due importanti termini

Una piccola digressione... quando valutiamo l'errore quadratico medio nell'MMSE con il modello noto facciamo la media sulla congiunta vera.

Se qui mettessimo p_β che è quella che vogliamo approssimare ha senso che "mediamo" sulla p vera e non sulla p falsa.

Siamo sicuri che saranno positive sia la parte a sinistra che a destra.

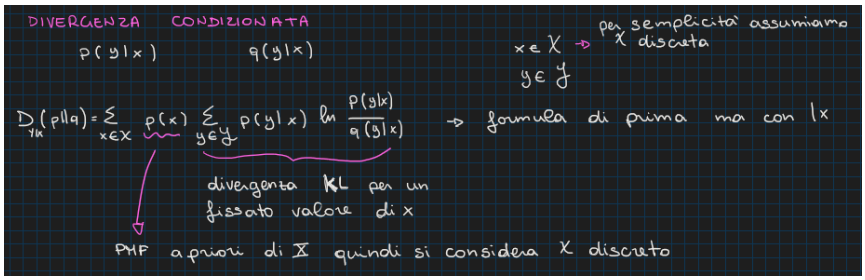
L'entropia di p è una quantità fondamentale, ad esempio nella compressione l'entropia misura il massimo livello di compressione ottenibile senza perdere informazioni.

La cross entropia tra p e q, ricordando che la vera p è p mentre q è l'approssimazione, rappresenta fisicamente quanti bit occorrono se siamo convinti che la vera p sia q e invece è p. Cioè applichiamo l'algoritmo di compressione come se fosse vera q ma invece è vera p.

Concettualmente è ovvio che visto che sto sbagliando scegliendo q mi serviranno più bit, quindi la cross entropia è maggiore dell'entropia e quindi la Divergenza risulta, correttamente, **positiva**. Infatti noi sappiamo che la divergenza KL è sempre non negativa e vale 0 se e solo se le due pmf sono uguali.

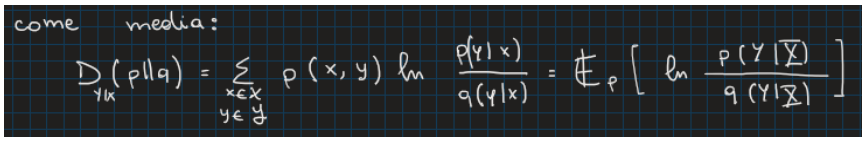
Notiamo che l'entropia di p non dipende da q, quindi una volta fissata la p minimizzare la divergenza significa minimizzare la cross-entropia.

Divergenza Condizionata



Nonostante ci sia il |x quella a destra è ancora una divergenza, p(y|x) e q(y|x) sono pmf e tanto basta.

Il problema però sorge se ad esempio vogliamo usare questo criterio per fare la distanza ad esempio tra due a-posteriori, quella vera e quella approssimata. Se lo facciamo per un certo valore di x è come se lo stessimo facendo solo per una delle features osservate nel training set, è come fare l'errore solo per un certo valore di X. Per questo noi la mediamo su tutte le X, per questo davanti c'è la sommatoria di X.



Possiamo unire condizionata e condizionante per ottenere la congiunta che poi ci permette di nuovo di esprimere la divergenza KL come una media.

La cosa interessante è che mentre quando si fa una KL si media solo rispetto alla pmf vera, ora mediando sulla congiunta, si media su tutte e due le variabili.

> 💡 La divergenza condizionata non è la divergenza per un particolare valore di X ma è la divergenza mediata su tutte le X, è la divergenza delle distribuzioni condizionate ma mediata su tutti i valori della variabile che condiziona.

Con l'MMSE nel quale avevamo l'errore dato X ma poi mediavamo su tutte le X.

Conclusione interessante sulla cross-entropy

Considerando che l'entropia rappresenta l'incertezza su Y, l'incertezza condizionata di Y|X rappresenta l'incertezza su Y dato che sono state osservate delle X.

Si potrebbe pensare, intuitivamente, che con le osservazioni l'entropia cali, tuttavia per alcuni valori fissati di X può succedere il contrario. Infatti ad esempio se c'è un valore di X che è molto fuorviante ha senso che l'incertezza su Y sale.

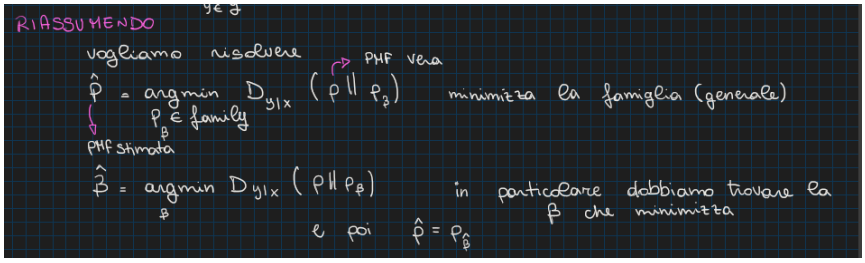
Quello che è vero invece è che se si definisce l'entropia condizionata mediando su tutti i valori di X allora l'entropia deve per forza scendere.

Quantità come l'entropia condizionata e la divergenza condizionata sono definite in questo modo da un punto di vista pratico perché ci serve un numero e non un numero per ogni X diversa. Però da un punto

di vista teorico c'è un significato molto forte, cioè il fatto che la cross entropy definita così garantisce il calo dell'entropia dopo le osservazioni di X come abbiamo detto prima.

Riassumendo

Il nostro obiettivo è trovare $\hat{p}$ che è l'argmin, tra tutte le distribuzioni p_β appartenenti alla nostra family, della Divergenza KL tra p (vera) e p_β.



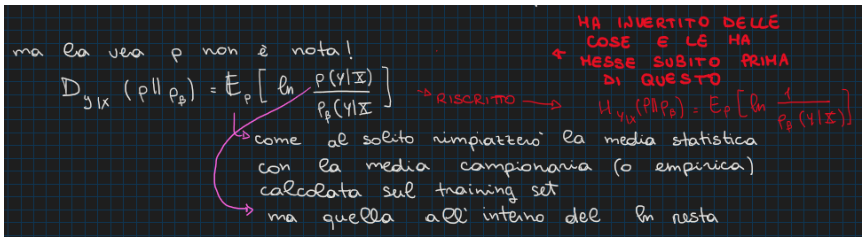
Praticamente questo si fa, visto che la famiglia è parametrizzata in β trovando $\hat{\beta}$ che è l'argmin su tutti i β dell'insieme sul quale è definito β (lui ha detto non lo so) sempre della KL tra p (vera) e p_β.

In un certo senso questo è esattamente quello che abbiamo fatto in regressione, quello che è cambiato è la "funzione di costo", che prima era l'MSE (RSS quando ci rendiamo conto che non abbiamo i modelli) ora è la Divergenza.

Perché questa cosa non è fattibile però? NON ABBIAMO LA PMF VERA!

Esattamente lo stesso ragionamento per il quale non ci siamo direttamente calcolati la posterior quando abbiamo fatto MMSE, non conosciamo il modello per calcolare la media.

Compreso il problema proviamo a guardare meglio questa D per cercare di capire se possiamo rimpiazzare le parti che non conosciamo.



La cross-entropia calcolata con la media prende il nome di ROC cross-entropia, quella con il training set e la sommatoria delle n prende il nome di cross-entropia empirica

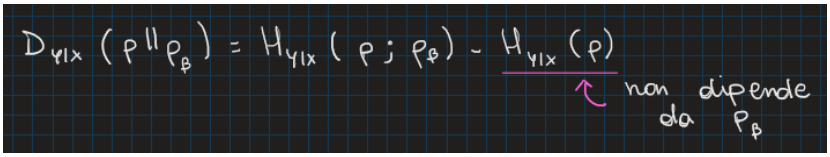
Notiamo che la p vera compare due volte, a livello della media e all'interno del logaritmo.

Per calcolare la media visto che abbiamo un Training Set usiamo la media campionaria (o empirica).

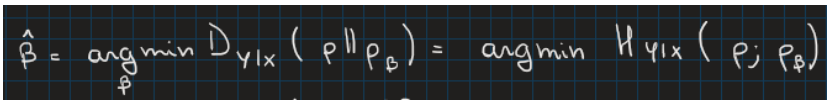
Come risolviamo invece la presenza della pmf vera nel logaritmo?

Notiamo che noi dobbiamo fare l'argmin su β e il numeratore non dipende da β!

Possiamo riscrivere la Divergenza KL come differenza tra cross-entropia condizionata ed entropia condizionata (le definiamo dopo, abbiamo definito quelle non condizionate in precedenza), l'entropia non dipende da β, quindi la minimizzazione è essenzialmente sulla cross entropia!

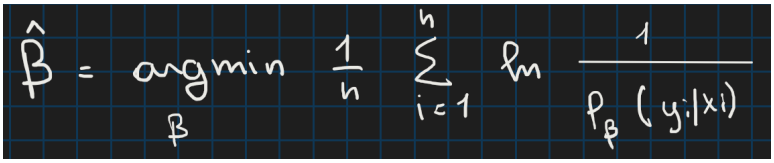


Ora una domanda, è vera la seguente eguaglianza?

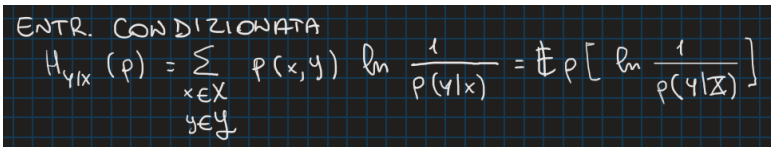


La posizione del minimo, che è quello che ci interessa, non cambia visto che abbiamo tolto una costante, ovviamente il valore del minimo è diverso ma non è importante.

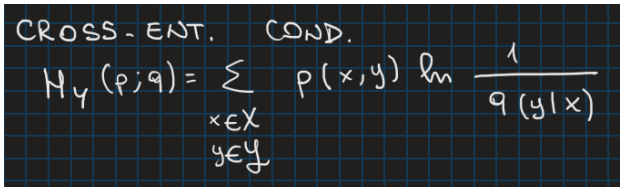
Giungiamo ad una formula che permette di risolvere il nostro problema senza conoscere il modello.



Entropia Condizionata



Cross-entropia Condizionata



Negli episodi successivi

Ora l'ultimo passaggio che ci manca per implementare la Classificazione Supervisionata è dare una faccia a p_β, sulla carta sono libero, p_β è una famiglia parametrica quindi non avrebbe senso dire "questa è una bella faccia e questa è una brutta faccia" MA dobbiamo considerare alcuni casi standard.

Il modello più utilizzato si chiama Regressione Logistica, termine molto fuorviante perché ci fa pensare che dobbiamo ricavare una Classificazione adattando una Regressione ma non è così. Noi il metodo lo usiamo perché funziona ed è popolare ma non lo studieremo dai testi, lo faremo uscire con 3 formule da quello che abbiamo fatto oggi.