

Classificazione - 18/11

Intro

Il problema di classificazione è un problema rilevante, vogliamo classificare dei dati per capire se appartengono ad una classe di un tipo o di un'altra. Si addestra il proprio sistema come si faceva nel problema di regressione, usando i dati X e le Y (si tratta comunque di un problema supervisionato), le Y non saranno più valori continui ma saranno discreti o addirittura non valori, quindi variabili categoriche.

Vogliamo capire come classificare le Y corrispondenti a future X .

Notiamo che quando le Y sono discrete, non continue, potremmo usare un approccio "tipo regressione lineare", si parla di modelli lineari generalizzati in statistica.

Ovviamente la classificazione è un qualcosa di più ampio, possiamo avere anche altre classi, qualcosa di categorico, il che ci fa chiedere ma il categorico è la stessa cosa del discreto? Sì, anche se sono approcci diversi.

Il problema della classificazione può essere impostato come un problema di regressione ma nel quale la Y è discreta, questo ci fa capire che dobbiamo solamente adeguare alcune caratteristiche del problema di regressione, questo è l'approccio degli statistici; alternatively si può vedere la Classificazione come un problema di decisione, abbiamo delle X e delle Y e vogliamo associare ogni X ad una classe, la Y quindi sarà discreta: numerica o categorica.

Features e Labels

Le X , che saranno un vettore di variabili Reali generiche, sono chiamate features.

$$X \in \mathbb{R}^d$$

feature

Si parla di vettore delle features o addirittura di matrice delle features se abbiamo più X e ogni X è a più dimensioni.

Ovviamente il nostro scopo è creare un modello che alle X associ le Y , le Y prendono il nome di labels.

Assumiamo che la Y sia monodimensionale ma si potrebbe estendere al caso a più dimensioni.

$$Y \in \{a, b, c\}$$

discreta

label $\{\text{cani, gatti, cavalli}\}$

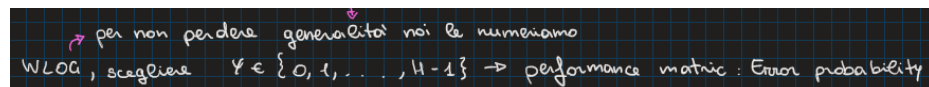
$\{\text{😊, 😞}\}$

Abbiamo detto che la Y deve essere discreta, o categorica, e abbiamo presentato alcuni esempi, si nota che anche se sono tutti valori discreti potrebbero anche essere infiniti valori.

Il problema di classificazione è un problema di decisione.

Pensiamo ad un caso semplice, due classi, vogliamo un modello di previsione discreta per sapere se domani piove o c'è il sole, notiamo che le due classi non c'entrano niente l'una con l'altra, non c'è nessuna relazione d'ordine.

Visto che quindi per Y si tratta di classi al fine di non perdere generalità le numeriamo, quindi abbiamo la prima classe, la seconda, la terza ecc.



WLOG, scegliere $Y \in \{0, 1, \dots, H-1\} \rightarrow$ performance metric: Error probability

WLOG = without loss of generality.

La metrica di misura delle prestazioni

In regressione il modo in cui valutavamo le prestazioni del modello di regressione era usando la distanza della \hat{Y} dalla vera Y (che in una dimensione era la differenza), e sommavamo tutti gli errori ottenuti con le N diverse Y , se Y fosse stato a più dimensioni avremmo potuto usare una norma. Quanto più piccola era la distanza tanto migliore era la stima.

In Classificazione il concetto di distanza perde di significato, qual è la distanza tra sole e pioggia? O tra cane, gatto e cavallo? Visto che il nostro problema è la decisione come metrica della qualità della stima possiamo usare l'errore, il modello ha scelto la classe giusta o sbagliata? Dell'errore è di interesse la probabilità.

Questo ci ricorda il Ricevitore ad Analisi dei Segnali, il Ricevitore riceveva infatti il vettore R del segnale + rumore e doveva decidere di quale segnale si trattava in una costellazione ad M segnali. Si tratta proprio dello stesso ragionamento, era un classificatore che prendeva dei dati in ingresso e decideva a quale regione ascrivere quel punto secondo regole geometriche. Le prestazioni erano misurate, come abbiamo detto per la classificazione, in termini di probabilità d'errore.

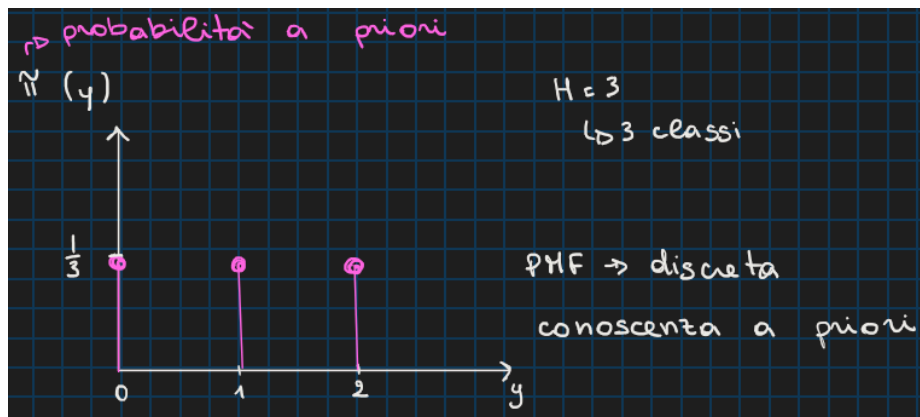
Bayesiano

L'idea della classificazione è che il modello classifica sulla base di una serie storica e semmai possiamo aggiornare il modello con le evidenze, i dati.

Stiamo quindi pensando ad un modello di apprendimento classico, abbiamo una decisione a-prioristica e poi volendo possiamo mettere questo insieme con le evidenze, quindi otteniamo un modello che è "dato i parametri" (una specie di modello generativo).

Mettendo insieme questi due elementi (prior e dati) otteniamo una probabilità a-posterior e quindi in pratica seguiamo un approccio Bayesiano.

Per prima cosa abbiamo la **probabilità a priori**.



Al fine di non perdere generalità le classi della Y sono solo numeri, come precedentemente introdotto, con $H=3$.

Notiamo una PMF omogenea, questa essendo prior è la conoscenza precedente alla raccolta dei dati, senza un'esperienza pregressa (che spesso c'è) non c'è modo di sapere cosa è più probabile e quindi la prior è uniforme.

Questo caso di omogeneità è molto interessante in quanto è molto semplificato dal fatto che siamo in discreto, in continuo una prior piatta sarà più complessa da considerare.

Se si tratta di una pdf a supporto compatto è facile farla piatta ma se invece facciamo inferenza su un qualcosa che è definito positivo, come scelgo l'altezza del plateau? L'integrale deve fare 1 quindi l'area non può essere infinita.

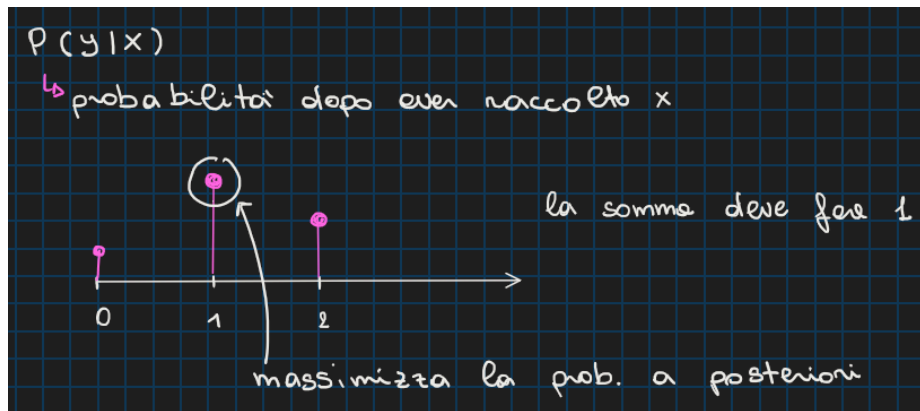
Per questo esistono in Bayesiano delle Prior continue che tengono conto della non informatività, sono tali che talvolta non sono vere e proprie pdf in quanto non integrano ad 1; tuttavia la cosa importante è che quando vengono usate nella regola di Bayes portano ad una posterior che è una pdf vera, per questo per Bayes esistono delle uniformi costanti però infinite, si sfrutta la proprietà importante che la regola di Bayes è definita a meno di una costante, l'importante è che la Posterior possa essere gestita come una probabilità a meno di una costante, non può essere infinita.

▼ pdf a supporto compatto

- Il **supporto** di una PDF è l'insieme dei valori in cui la funzione è diversa da zero.
- Un **insieme compatto** in matematica è un insieme chiuso e limitato.

Quindi, una PDF a supporto compatto è una densità di probabilità che è nulla al di fuori di un intervallo finito $[a, b]$.

Dopo aver raccolto i dati avremo una **probabilità a posteriori**.



Ovviamente otteniamo ancora una volta una PMF con 3 classi (0, 1, 2) ma cambia la distribuzione delle probabilità che è stata aggiornata grazie ai dati.

Ottenuta la posteriori la domanda diventa: come scelgo la classe? Quale Y è la più plausibile secondo i dati che ho raccolto?

La risposta intuitiva è anche quella corretta, si sceglie la Y a probabilità più alta, la scelta più plausibile è quella che massimizza la probabilità a posteriori.

Questo criterio di scelta prende il nome di MAP (Maximum A Posterior).

la regola di scegliere il massimo → MAP rule
 a
 x
 i
 m
 u
 m
 o
 s
 t
 e
 r
 i
 o
 r

Il classificatore ottimo

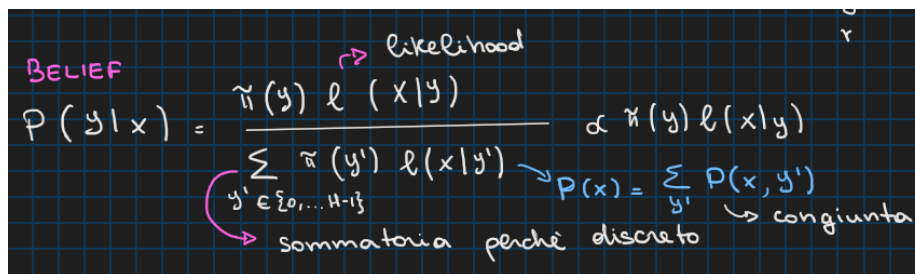
CLASSIFICATORE OTTIMO (minimizza la probabilità d'errore)
 $\hat{Y} = \underset{y \in \{0, 1, \dots, H-1\}}{\operatorname{argmax}} P(y|x)$
 funzione di x scelta dalle ipotesi

Come accennato il Classificatore Ottimo è scelto in base ad un criterio di massimizzazione della posterior, questo coincide anche con una minimizzazione della probabilità d'errore.

Dipendendo dalle X osservate la \hat{Y} è funzione di X .

Richiami sulla Regola di Bayes

Per esplicitare il calcolo usiamo la Regola di Bayes.



$$P(y|x) = \frac{\pi(y) l(x|y)}{\sum_{y' \in \{0, \dots, H-1\}} \pi(y') l(x|y')} \propto \pi(y) l(x|y)$$

$P(x) = \sum_{y'} P(x, y')$ (congiunta)

vengono chiamate y' penso per non creare ambiguità con la y a numeratore, comunque si riferiscono alla stessa y però a denominatore c'è la sommatoria che "gira" su una variabile diversa altrimenti influenzava anche il numeratore

Anche qui come per il modello di regressione la l è il modello generativo delle X se Y è noto. Il modello generativo sarà costruito, ovviamente, in base ai dati. Scriviamo l perché è proprio la likelihood.

Ricordiamo che il denominatore serve solo a normalizzare il numeratore, invece dell'integrale come dimostrammo in passato ora c'è la sommatoria perché siamo in discreto.

Ad essere precisi ciò che c'è a denominatore è la $p(x)$ cioè la distribuzione di probabilità delle X e visto che la X è fissata ($P(Y|X)$ fissa la X) possiamo confermare che a denominatore c'è una costante.

Risulta di interesse notare che la $p(x)$ può anche essere calcolata come marginalizzazione, cioè partendo dalla congiunta $p(x, y')$ e facendone l'integrale (visto che siamo in discreto la sommatoria) rispetto alla variabile che intendiamo eliminare cioè Y' . Quest'osservazione ha senso perché ricordando che la congiunta si può esprimere come condizionata per condizionante allora notiamo che è vera l'uguaglianza:

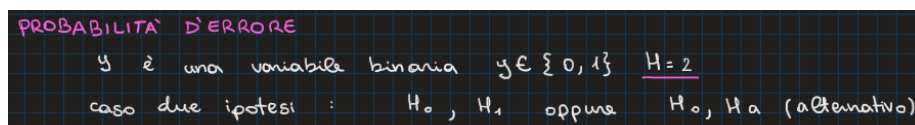
$$\sum_{y' \in \{0, \dots, H-1\}} \pi(y') l(x|y') = \sum_{y' \in \{0, \dots, H-1\}} p(x, y')$$

Le probabilità a posteriori, come $P(Y|X)$ hanno anche un nome in questo contesto: **Belief**, le "plausibilità", le "probabilità" di avere una certa classe Y dopo aver osservato X .

Probabilità d'Errore

Per parlare della probabilità d'errore partiamo da un caso semplice e poi estendiamo a casi più particolari.

Partiamo dal caso binario.



PROBABILITA' D'ERRORE

y è una variabile binaria $y \in \{0, 1\}$ $H=2$

caso due ipotesi: H_0, H_1 oppure H_0, H_a (alternativo)

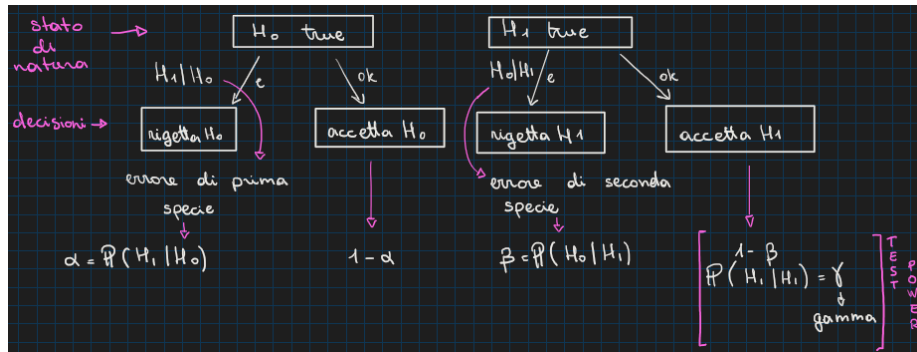
Nel caso binario consideriamo $H=2$ quindi ci sono due classi che Y può rappresentare.

Esistono moltissimi contesti di applicazione della Classificazione binaria, 0 nessun nemico in avvicinamento, 1 nemici in avvicinamento, oppure 0 nessuna malattia rilevata, 1 malattia rilevata.

I test di ipotesi in regressione erano effettivamente delle Classificazioni.

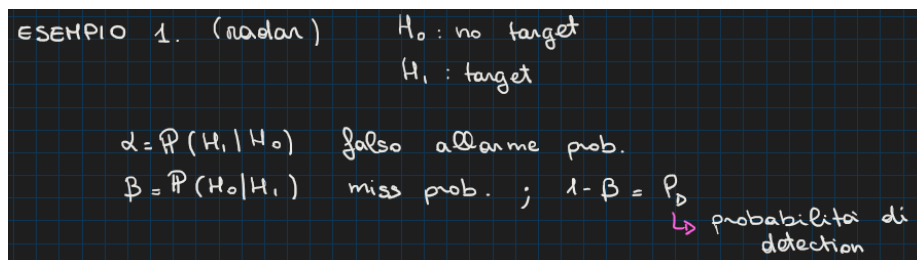
Se abbiamo due ipotesi le chiamiamo H_0 e H_1 oppure H_0 e H_a .

Analizziamo tutte le possibilità.



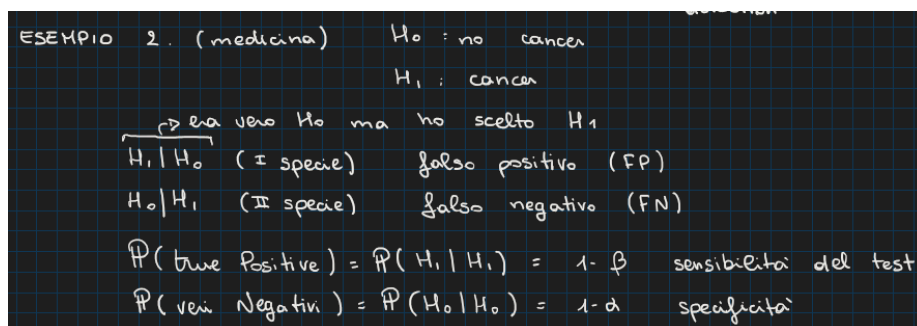
Abbiamo quindi due possibili errori, α (first type error, o risk) e β (second type error, o risk).

H_1 potrebbe rappresentare la rilevazione del nemico e infatti la probabilità di scegliere H_1 quando H_1 è vera è detta **probabilità di detection**, visto che questa probabilità rappresenta la capacità di riconoscere l'ipotesi alternativa viene anche chiamata **Test Power**, talvolta anche chiamata γ .



α è chiamata a seconda del campo applicativo probabilità di falso allarme o probabilità di falso positivo (FP).

β è chiamata a seconda del campo applicativo probabilità di miss (probabilità di mancare l'allarme) o probabilità di falso negativo (FN).



Metriche importanti per la valutazione di un Classificatore sono $1 - \beta$ che è la **sensibilità del test** e $1 - \alpha$ che è la **specificità**.

Ricapitolazione sull'errore.

$$\begin{aligned}y &\in \{0, 1\} \\ P[\text{err} | Y=0] &= P[\hat{Y}=1 | Y=0] \quad \text{FP} \\ P[\text{err} | Y=1] &= P[\hat{Y}=0 | Y=1] \quad \text{FN} \\ P[\text{err}] &= P[\text{err} | Y=1] \pi(1) + \pi(0) P[\text{err} | Y=0]\end{aligned}$$

Sommando i due diversi errori otteniamo la probabilità totale dell'errore.

MAP in caso binario

La MAP nel caso a due ipotesi o fa scegliere 0 o 1 e nello specifico fa scegliere quella a probabilità più elevata, quindi possiamo riscriverla in maniera molto semplice con un rapporto.

$$\frac{P(1|x)}{P(0|x)} \underset{0}{\overset{1}{\geq}} 1 \quad \begin{array}{c} \text{è equivalente a} \\ \uparrow \\ \Leftrightarrow \end{array} \quad \frac{\pi(1) l(x|1)}{\pi(0) l(x|0)} \underset{0}{\overset{1}{\geq}} 1$$

↳ per rappresentare la decisione

Abbiamo confrontato il rapporto tra le due possibili posterior con un valore soglia che in questo caso è 1.

Il simbolo maggiore-minore indica proprio a seconda del caso qual è la scelta che prendiamo, se il rapporto è maggiore di 1 scegliamo 1, viceversa 0.

Ovviamente nulla ci vieta di sostituire le posterior usando la regola di Bayes, notiamo che il denominatore essendo uguale per entrambe le posterior visto che è la stessa costante dati i dati, quindi si elimina.

$$\frac{\pi(1) l(x|1)}{\pi(0) l(x|0)} \underset{0}{\overset{1}{\geq}} 1 \quad \Leftrightarrow \quad \frac{l(x|1)}{l(x|0)} \underset{0}{\overset{1}{\geq}} \frac{\pi(0)}{\pi(1)}$$

likelihood ratio

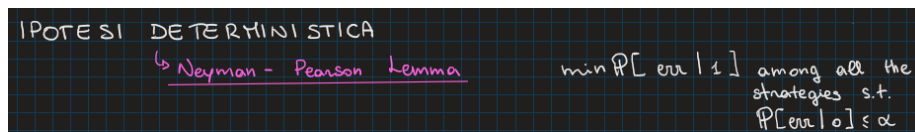
Portiamo il rapporto tra le prior a destra il che ci permette di fare un'osservazione importante, se la prior è uniforme il rapporto $\frac{\pi(0)}{\pi(1)}$ è 1 e quindi l'unica cosa che si deve conoscere per fare la scelta ottima secondo MAP è la Likelihood, la Likelihood maggiore determina la scelta, in pratica il criterio è diventato Massima Verosimiglianza, come succedeva ad Analisi dei Segnali, quando con tutto

Gaussiano massima verosimiglianza diventava minima distanza e bastava quindi andare nello spazio geometrico, mettere i puntini e tramite le distanze calcolare la probabilità d'errore.

La similitudine con Analisi dei Segnali non è casuale, in telecomunicazioni il ricevitore ottimo si chiama ricevitore MAP, che diventa Massima Verosimiglianza quando i simboli (non per forza due) sono tutti equiprobabili.

Il Rapporto di Verosimiglianza va confrontato con un valore soglia, essendo una soglia fissa che dipende dalle prior che possiamo chiamare γ . Questo ratio torna spesso in statistica in quanto è molto potente per fare test, infatti il likelihood ratio ha la forma ottima di un test di ipotesi.

Criterio storico ottimo per la scelta tra due ipotesi



Per ipotesi deterministica intendiamo che il valore delle H è preciso, la Y può essere uguale oppure no.

▼ Ipotesi deterministica by ChatGPT (leggete il commento)

Nel contesto della classificazione in statistica e machine learning, un'**ipotesi deterministica** si riferisce a una situazione in cui esiste una relazione precisa e senza ambiguità tra i dati osservati (le variabili predittive) e le classi (le etichette o le categorie da prevedere). In altre parole:

- **Ipotesi deterministica:** ogni combinazione di valori delle variabili predittive (X) è associata a una specifica classe (Y) in modo univoco. Non c'è incertezza o variabilità: dato un input X , la classe Y corrispondente è sempre la stessa.

Caratteristiche principali:

1. **Assenza di rumore:** i dati seguono esattamente il modello, senza errori o deviazioni casuali.
 - Ad esempio, se un modello dice che $Y=1$ quando $X_1>0$, allora ogni osservazione che soddisfa $X_1>0$ avrà sempre $Y=1$.
2. **Relazione completamente definita:** non ci sono sovrapposizioni o ambiguità tra le classi. Non esistono due osservazioni con lo stesso X che appartengono a classi diverse.

Contrasto con ipotesi probabilistica

Un'**ipotesi probabilistica**, invece, assume che la relazione tra X e Y sia soggetta a incertezza o variabilità:

- La stessa combinazione di valori di X può essere associata a diverse classi con probabilità diverse.
- Ad esempio, potremmo dire che se $X_1>0$, allora Y ha il 90% di probabilità di essere 1 e il 10% di essere 0.

Il teorema che stiamo per enunciare vale quando un'ipotesi H_0 è fissata e un'altra ipotesi H_1 vale un altro valore diverso dal precedente, si parla di ipotesi semplici in statistica perché c'è un singolo punto

e dobbiamo scegliere tra un valore ed un altro.

Neyman-Pearson Lemma



Questo lemma suggerisce qual è la strategia che minimizza la probabilità d'errore con un vincolo su uno dei due errori, ad esempio fissato α fornisce la strategia ottima per minimizzare β .

Solitamente in statistica essendo la α l'ipotesi nulla era quella che si fissava e si cercava quindi la strategia per minimizzare l'errore di seconda specie in modo da aumentare la Potenza del Test, ad oggi non c'è nessuna preferenza tra le due.

Concludiamo che per il Lemma di Neyman-Pearson il classificatore ottimo è il seguente.

$$\text{Classificatore ottimo (N-P Lemma)} \\ \frac{L(x|1)}{L(x|0)} \underset{0}{\overset{1}{><}} \gamma(\alpha)$$

Likelihood Ratio maggiore-minor, scegliendo 1 e 0 rispettivamente, di una soglia γ che scegliamo in funzione di α , cioè il rischio sull'ipotesi vincolante.

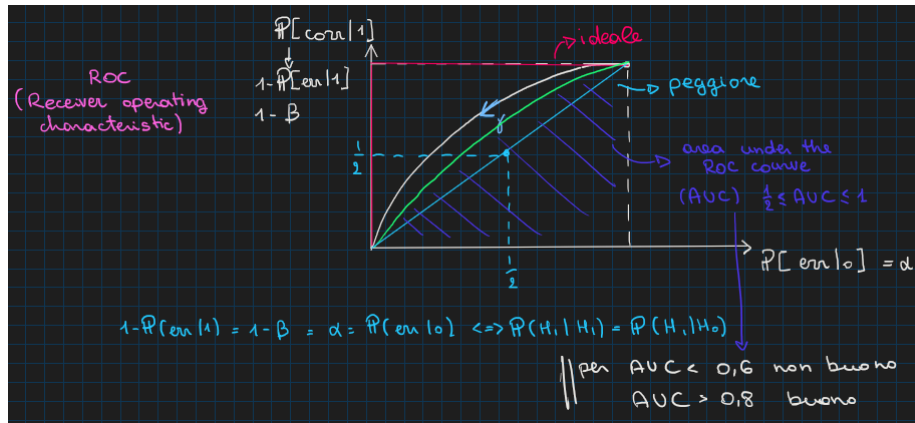
Il Lemma di N-P dice che affinché si possa ottenere la struttura ottima bisogna rispettare due condizioni:

- il test va fatto con il rapporto di verosimiglianza;
- la soglia, cioè la regione critica, di rifiuto, deve essere decisa sulla base dell' α che si sta fissando.

ROC - Receiver Operating Characteristic

Al variare della soglia α si avranno diversi valori con i quali confrontare il ratio per fare il test.

Se si rappresenta il test utilizzando sulle X la probabilità d'errore data l'ipotesi 0 e sulle Y $1 - \beta$ (storicamente si metteva β) cioè la potenza del test.



Notiamo una forma curva, essendo probabilità su entrambi gli assi abbiamo gli intervalli $[0, 1]$.

Queste curve saranno parametrizzate sulla base della soglia γ , al crescere di gamma il punto si muove da in alto a destra al punto con probabilità d'errore 0 (in basso a sinistra).

La **curva ideale sarebbe la rossa**, che ha sempre la perfetta capacità di predizione.

Questa curva prende il nome di ROC, caratteristica operativa del ricevitore, rappresenta il riferimento con il quale bisognerebbe confrontare diverse strategie.

Più il test è buono ed efficiente più la curva si avvicina a quella ideale.

Raccogliendo più dati il test migliora e la ROC si sposta verso il punto in alto a sinistra.

Il punto peggiore per la ROC non è in basso a destra, massimo errore, minima capacità predittiva, perché basta ribaltare la decisione ogni volta.

Il punto peggiore quindi è la bisettrice del quadrante (azzurra), perché porta la massima incertezza, il punto $(1/2, 1/2)$ nello specifico è proprio il lancio della moneta, vedere i dati non serve a nulla.

Leggere l'equazione azzurra.

$$P[H_1|H_1] = P[H_1|H_0]$$

Con la linea azzurra indipendentemente da H_1 o H_0 la probabilità di scegliere H_1 è uguale, si sceglie indipendentemente dai dati raccolti e questa è la situazione peggiore.

La ROC essendo grafica si può tracciare in molti modi, ad esempio per simulazione, in Matlab, in R o in altri modi.

Quali numeri si potrebbero ottenere dalle curve per decidere qual è la migliore?



L'area sottesa alla curva, l'area massima è l'area 1×1 , la peggiore è $1/2$. L'area sottesa alla curva ROC è considerata un parametro molto importante, prende il nome di Area Under the Curve o AUC.

$$\frac{1}{2} \leq AUC \leq 1$$

Quindi possiamo usare varie strategie e confrontare le AUC di diverse ROC ottenute.

Grossolanamente si può dire che AUC maggiore di 0.8 è sicuramente un valore interessante mentre minore di 0.6 non è un buon risultato, anche se comunque sono numeri presi a latere ed è invece più rilevante confrontarli tra loro.