UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Ingegneria dell'Informazione ed Elettrica e Matematica applicata

Corso di Laurea Magistrale in Ingegneria Informatica

APPUNTI DI DATA SCIENCE

DI FRANCESCO PIO CIRILLO

https://github.com/francescopiocirillo

“Sii sempre forte”

😄 Ehi, un attimo prima di iniziare!

Hai appena aperto una raccolta di appunti che ho deciso di condividere **gratuitamente** su GitHub, se ti sono utili fai **una buona azione digitale**:

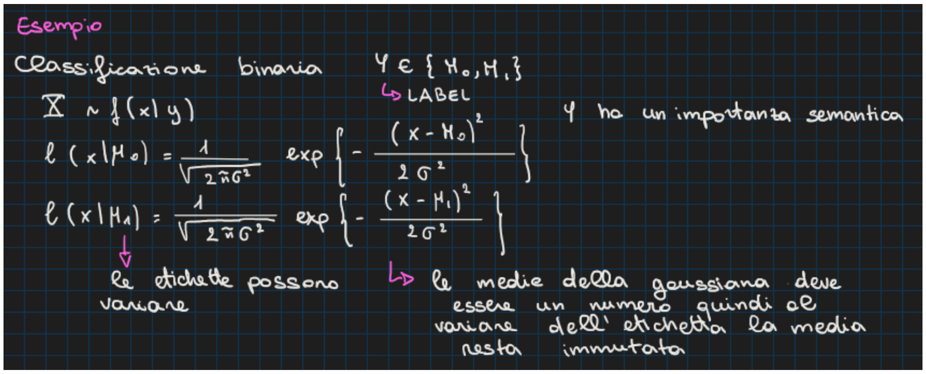
- ⭐ **Lascia una stellina alla repo**: è gratis, indolore e fa super piacere!
- 📤 **Condividila con amici**, compagni di corso, o chiunque possa averne bisogno.

Insomma, se questi appunti ti salvano anche solo una giornata di studio... fammelo sapere con una **stellina**!

Grazie di cuore ❤️

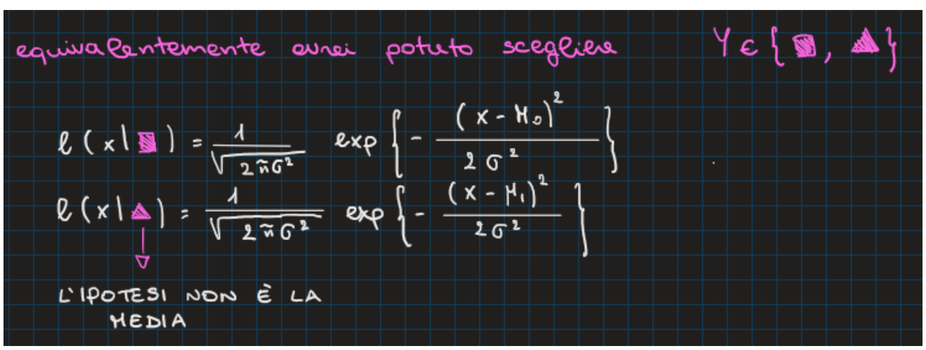
Classificazione: caso binario - 19/11

Esempio: classificazione binaria



Y è una variabile categorica, ha una importanza semantica, ad esempio se è sole o pioggia è ovvio che il significato di Y ha valore, ma non ha un valore che è matematicamente rilevante, differentemente da quanto avveniva in regressione.

Nell'esempio presentato anche se le Y hanno lo stesso nome delle medie delle Gaussiane le due cose non sono correlate e cambiando le etichette di Y assolutamente non varia il valore delle medie delle Gaussiane in quanto le etichette non hanno nessun valore al di là di quello semantico.



Ovviamente una volta acquisita dimestichezza se so che le medie hanno come simboli rappresentativi μ_0 e μ_1 allora invece di introdurre altri simboli uso quelli come etichette delle Y ma è solo una questione di comodo.

Abbiamo visto lo scenario del criterio Bayesiano ottimo: MAP, e il caso con parametro deterministico cioè il criterio N-P.

Ora calcoliamo i due detector per questi due casi.

Caso Bayesiano

Ci manca solo un'informazione, nel caso Bayesiano assumiamo:

$$\mathbb{P}[Y = H_0] = \frac{1}{2}$$

Nell’altro caso invece non abbiamo bisogno di assumere niente

caso Bayesiano $\mathbb{P}[Y = H_0] = \frac{1}{2}$

Calcolare M detector ottimo

Criterio MAP: max prob. a posteriori

Vogliamo calcolare il detector ottimo, seguendo il criterio MAP.

Per massimizzare la posteriori dobbiamo sicuramente calcolarla.

Sappiamo che la posteriori è una PMF perché Y è discreta.

likelihood

$$P(y|x) = \frac{\pi(y)\,\ell(x|y)}{\sum_{y' \in \{0,\dots,M-1\}} \pi(y')\,\ell(x|y')} = \frac{\ell(x|y)}{\ell(x|H_0) + \ell(x|H_1)}$$

ma per massimizzare la prob. a post il denominatore non serve !!!

al den deve esserci la somma del num perchè la PMF $\triangleq 1$

$\pi(y) = \mathbb{P}[Y = y] = \frac{1}{2}$

Usiamo Bayes per il calcolo della Posterior, a denominatore c’è l’evidenza, la distribuzione di X che è però uguale alla sommatoria (integrale in continuo) su tutte le Y del numeratore visto che il denominatore è solo una costante per normalizzare la frazione. Il denominatore tecnicamente non è una costante ma lo è se fissiamo X come si fa appunto nella posterior.

Svolgiamo la sommatoria a denominatore visto che è molto facile nel caso binario.

La prior vale $1/2$ per tutti i valori quindi si semplifica a numeratore e denominatore.

Abbiamo fatto tutti questi conti, ma per calcolare il detector davvero era necessario?

Il criterio MAP dice che dobbiamo massimizzare la posterior ma il denominatore non serve perché la massimizzazione non cambia a meno di una costante.

Noi comunque abbiamo calcolato la posterior perché potrebbe essere un esercizio.

Calcoli sul MAP binario

Nel caso binario è molto semplice procedere dopo aver trovato la posterior con la regola di Bayes:

$$\frac{\tilde{\pi}(\mu_1)\,\ell(x|\mu_1)}{\tilde{\pi}(\mu_0)\,\ell(x|\mu_0)} \underset{\mu_0}{\overset{\mu_1}{\gtrless}} 1 \iff \frac{\ell(x|\mu_1)}{\ell(x|\mu_0)} \underset{\mu_0}{\overset{\mu_1}{\gtrless}} \left[\frac{\tilde{\pi}(\mu_0)}{\tilde{\pi}(\mu_1)}\right] = 1$$

(0)

Ma a quanto equivale il Likelihood Ratio numericamente?

Sono due Gaussiane per le quale la varianza è la stessa quindi il loro rapporto è solo un grande esponenziale.

$$\frac{\ell(x|\mu_1)}{\ell(x|\mu_0)} = \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_0)^2}{2\sigma^2}\right\}$$

$$= \exp\left\{\frac{2x(\mu_1-\mu_0)+\mu_0^2-\mu_1^2}{2\sigma^2}\right\}$$

Quindi possiamo sostituire questo valore all'interno della disuguaglianza trovata in precedenza:

$$\exp\left\{\frac{x(\mu_1-\mu_0)}{\sigma^2} - \frac{\mu_1^2-\mu_0^2}{2\sigma^2}\right\} \gtrless 1$$

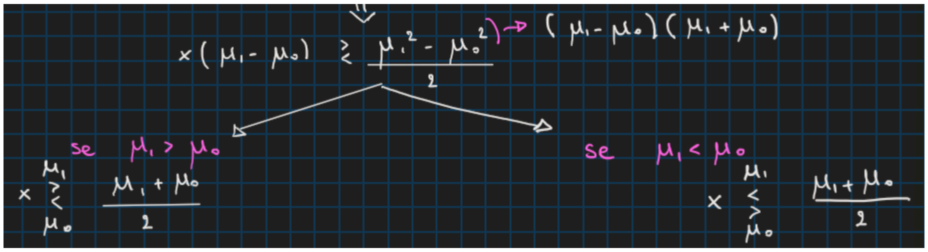
$$\Updownarrow \quad \text{APPLICO } \ln$$

$$\frac{x(\mu_1-\mu_0)}{\sigma^2} \gtrless \frac{\mu_1^2-\mu_0^2}{2\sigma^2}$$

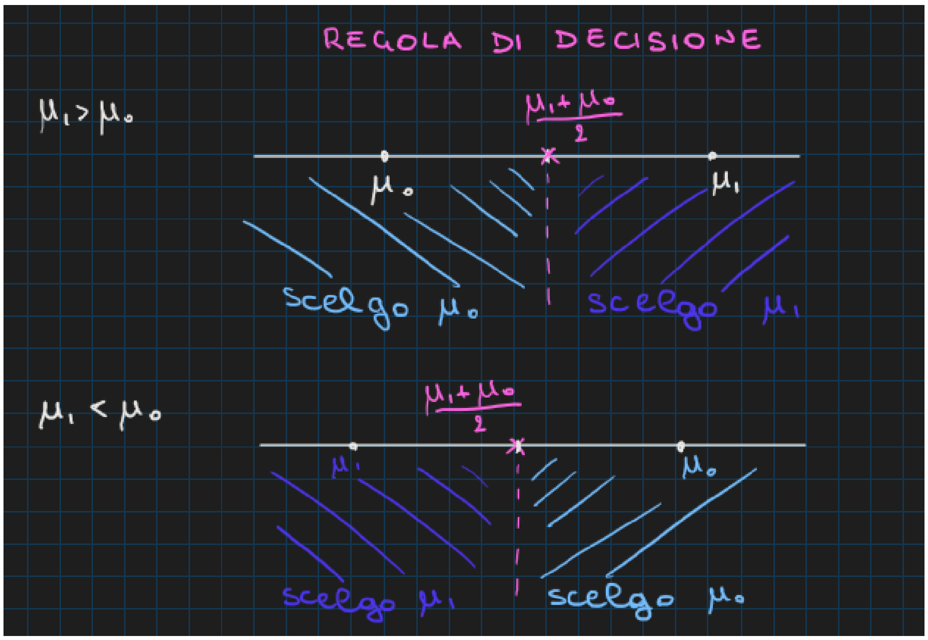
(1)

Visto che σ^2 è positivo e uguale da entrambi i lati possiamo toglierlo senza nessuna accortezza.

Per semplificare $\mu_1 - \mu_0$ occorre sapere se è positivo o negativo quindi dobbiamo distinguere nei due casi.



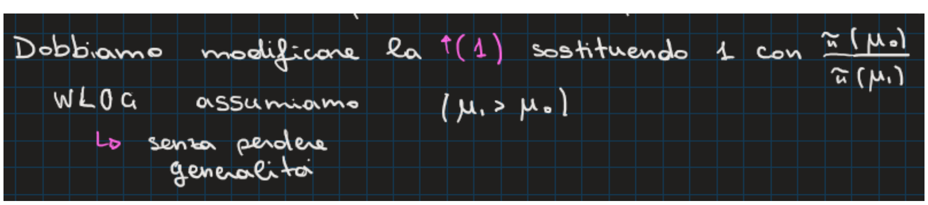
Graficamente i due casi sono rappresentabili come segue:



Il fatto che concettualmente abbia così senso, scelgo sulla base di quanto è vicino al valore che scelto, viene dal fatto che sono distribuzioni Gaussiane e tendenzialmente le distribuzioni Gaussiane portano a risultati intuitivi.

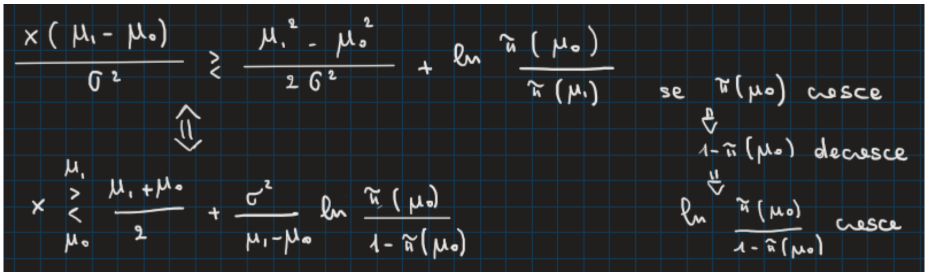
Il fatto che il punto in cui cambia la decisione è al centro è perché sono equiprobabili, la prior sposta la soglia verso la scelta più probabile.

Cosa cambia se la probabilità a priori non è uniforme?



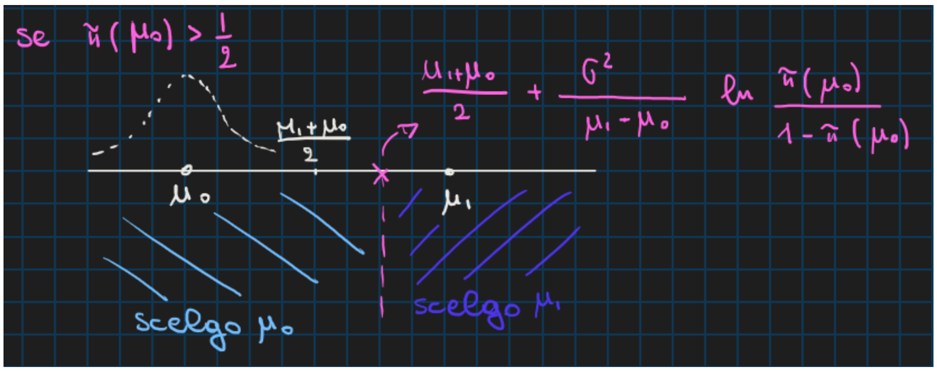
Assumiamo che sia $\mu_1 > \mu_0$.

Dalla (1) otteniamo:



Dalle considerazioni a destra capiamo che se $\pi(0)$ cresce allora la sua area di decisione cresce.

Graficamente:



In altre parole usiamo l'informazione a priori per dare meno importanza al dato, anche se la X è al di sopra della metà e si potrebbe pensare di scegliere μ_1 la prior mi dice che fino alla soglia devo scegliere μ_0.

Ora parliamo di varianza.

Notiamo che una volta fissati $\pi(0)$ e $\pi(1)$ al crescere della varianza i dati sono più rumorosi e sporchi, se i dati sono più sporchi assumiamo che sia più probabile che si siano spostati più del dovuto.

Quindi se ci hanno detto a priori che c'è un motivo per non essere simmetrico, cioè è molto più probabile μ_0, diamo credito a questa informazione e allarghiamo la regione di decisione, ma se non ci è stato detto per simmetria ci mettiamo in mezzo.

Il fatto che se cresce la varianza ha più importanza la prior rispetto ai dati lo vediamo nel fatto che la varianza moltiplica il logaritmo del rapporto delle prior.

Se le due medie sono più vicine la differenza tra le due è più piccola e quindi il termine

$$\frac{\sigma^2}{\mu_1 - \mu_0}$$

è più grande e quindi, come per quando cresce la varianza, diamo meno peso ai dati e più peso alla prior, che ha senso perché se sono vicine le distribuzioni è più probabile che i punti abbiano superato la soglia.

Detector NP

NP mette a confronto il likelihood ratio con una soglia arbitraria γ che quindi è come se andasse a sostituire il rapporto delle prior nella (0).

$$\frac{\ell(x|\mu_1)}{\ell(x|\mu_0)} \underset{\mu_0}{\overset{\mu_1}{\gtrless}} \gamma$$

soglia arbitraria

Il detector ottimo prende il nome di Likelihood Ratio test.

Ripetendo i calcoli otteniamo

$$\exp\left\{ \frac{x(\mu_1-\mu_0)}{\sigma^2} - \frac{\mu_1^2-\mu_0^2}{2\sigma^2} \right\} \underset{\mu_0}{\overset{\mu_1}{\gtrless}} \gamma$$

$$\Updownarrow \ln$$

$$\frac{x(\mu_1-\mu_0)}{\sigma^2} \gtrless \ln\gamma + \frac{\mu_1^2-\mu_0^2}{2\sigma^2}$$

Ora assumiamo $\mu_1 > \mu_0$

per $\mu_1 > \mu_0$

$$x \underset{\mu_0}{\overset{\mu_1}{\gtrless}} \frac{\sigma^2}{\mu_1-\mu_0}\ln\gamma + \frac{\mu_1+\mu_0}{2} \triangleq \tau \in (-\infty, +\infty) \text{ perché } \gamma \in (0, \infty)$$

Dobbiamo capire un fatto fondamentale.

La soglia con la quale confrontiamo il likelihood ratio che significato ha?

La soglia del criterio MAP era una cosa immutabile.

Con il criterio di NP la soglia si può spostare sulla base dell'errore di primo tipo che vogliamo fissare.

La soglia γ è un numero da 0 a ∞, il logaritmo di un numero positivo va da $-\infty$ a $+\infty$, è un numero arbitrario, fissati tutti i parametri resta sempre possibile spostare tutto il lato destro della disequazione tra $-\infty$ a $+\infty$ e questo vuol dire che effettivamente dei parametri μ_0, μ_1 e σ^2 non ci importa, di conseguenza possiamo prendere tutto il lato destro della disequazione e chiamarlo τ e fissare direttamente quello.

Vogliamo fissare τ in modo da ottenere una specifica probabilità di falso allarme/falso positivo.

Detector ottimo secondo criterio NP

$$x \underset{\mu_0}{\overset{\mu_1}{\gtrless}} \tau$$

Fissiamo ora τ in modo da ottenere una PROB. FALSI POSITIVI

$$P[\text{scegli } \mu_1 \mid \text{è vera } \mu_0] = \alpha$$

Per un problema di classificazione binario i numeri che caratterizzano la performance sono (un ragazzo dice “la ROC” e Matta dice di non considerare la AUC) le due probabilità d’errore (falso positivo e falso negativo), se è Bayesiano ne basta 1, la probabilità d’errore totale. Se si confrontano due sistemi per la probabilità di Falsi Positivi bisogna fissare i Falsi Negativi e viceversa, se ci sono più indici bisogna fissarli tutti tranne uno.

$\mathbb{P}[\text{scegli } \mu_1 \mid \text{è vera } \mu_0] = \mathbb{P}[\bar{X} > \tau \mid \text{è vera } \mu_0]$

$= \mathbb{P}[\text{Gauss}(\mu_0, \sigma^2) > \tau] =$

$\stackrel{!}{=} \mathbb{P}\left[\frac{\text{Gauss}(\mu_0, \sigma^2) - \mu_0}{\sigma} > \frac{\tau - \mu_0}{\sigma}\right] = Q\left(\frac{\tau - \mu_0}{\sigma}\right)$

I SISTEMI SI VALUTANO A PARITÀ DI FP O DI FN ERRORE PESANTE!

La funzione **Q** (o **qfunc** in MATLAB) è una funzione matematicamente definita che rappresenta la **funzione complementare della distribuzione cumulativa normale standard**. È largamente utilizzata in campi come le telecomunicazioni, la teoria dei segnali e il calcolo delle probabilità, specialmente per analizzare errori e rumore.

Per portarci all’utilizzo della Q function dobbiamo standardizzare.

imponiamo $Q\left(\frac{\tau - \mu_0}{\sigma}\right) = \alpha$ e otteniamo la soglia in funzione di α

$$\frac{\tau - \mu_0}{\sigma} = Q^{-1}(\alpha)$$

$$\tau = \mu_0 + \sigma Q^{-1}(\alpha)$$ VALORE DI SOGLIA

inversamente proporzionali

Calcoliamo la ROC

sull'asse delle ordinate c'è $\mathbb{P}[\text{scegli } \mu_1 \mid \text{è uscita } \mu_1] = \mathbb{P}[\bar{X} > \tau \mid \text{è vera } \mu_1]$

$\mathbb{P}[\text{Gauss}(\mu_1, \sigma^2) > \tau] = Q\left(\frac{\tau - \mu_1}{\sigma}\right)$

sostituendo τ

$$Q\left(\frac{\mu_0 - \mu_1}{\sigma} + Q^{-1}(\alpha)\right)$$ ROC

se $\sigma \to \infty$ la ROC diventa la retta del caso peggiore

Q^{-1} è la funzione inversa di $Q(X)$. $Q(X)$ è una funzione che restituisce la probabilità di eccedere X e quindi al crescere di X è sempre più improbabile eccedere X e la $Q(X)$ tende a zero.

Se α è una probabilità e tende a 0, Q^{-1} che è il valore che dato alla Q permetterebbe di ottenere una probabilità piccola tende ad ∞.

Ha senso perché se la probabilità di superare la soglia è bassa la soglia deve essere alta.

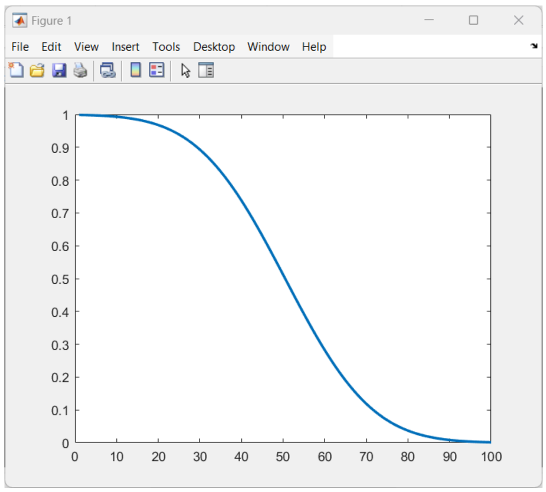
Quindi abbiamo chiarito che al crescere di α Q^{-1} decresce, la parte sommata a Q^{-1} è fatta da tutte quantità positive e non ci crea nessun disturbo nel come dobbiamo scrivere la ROC, quindi possiamo procedere.

Matlab

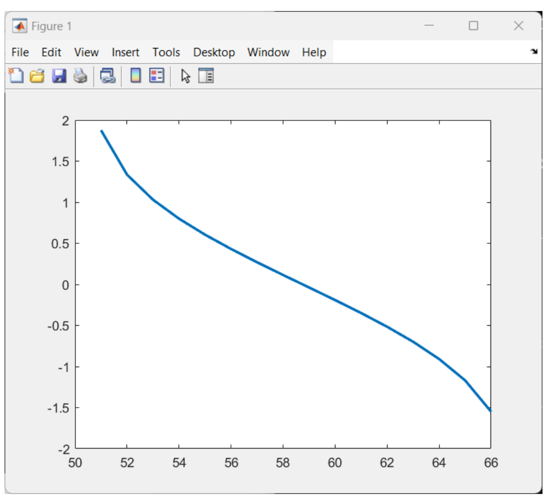
Proviamo a disegnare la ROC. Se α cresce la soglia si abbassa e la proabilità di superarla si alza.

Se cresce μ_1 la differenza tra μ_0 e μ_1 cresce (nel nostro schema $\mu_1 > \mu_0$), decresce la soglia e diventa più facile superarla, ha senso perché se sono più vicine è più facile distinguerle e la ROC deve alzarsi.

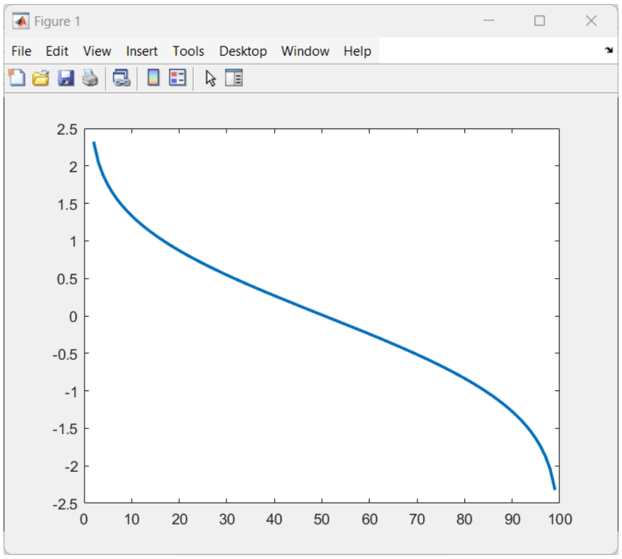
```
x = linspace(-3,3)
plot(qfunc(x), 'LineWidth', 2)
```



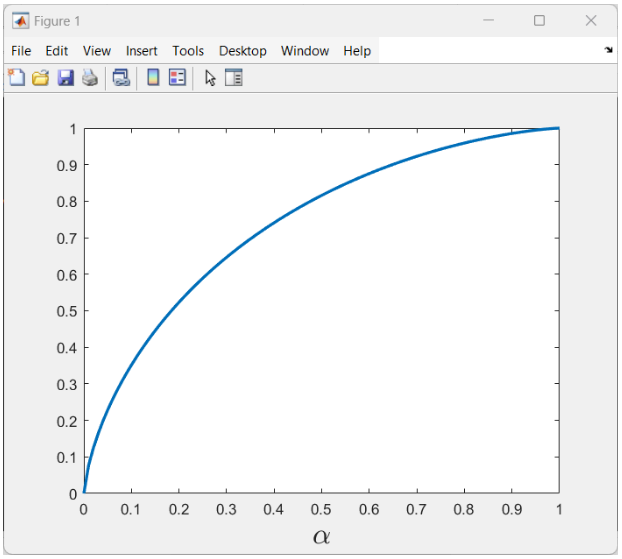

```
x = linspace(-3,3)
plot(qfuncinv(x), 'LineWidth', 2)
```



```
a = linspace(0,1)
plot(qfuncinv(a), 'LineWidth', 2)
```




```
m1 = 1.1
m0 = 0.2
sig = 1
Pcorr = qfunc( -(m1-m0)/sig + qfuncinv(a) )
plot(a, Pcorr, 'LineWidth', 2)
xlabel('$\alpha$', 'Interpreter','latex', 'FontSize', 20)
```



DISCLAIMER

Questi appunti sono stati realizzati a scopo puramente educativo e di condivisione della conoscenza. Non hanno alcun fine commerciale e non intendono violare alcun diritto d'autore o di proprietà intellettuale.

I contenuti di questo documento sono una rielaborazione personale di lezioni universitarie, materiali di studio e concetti appresi, espressi in modo originale ove possibile. Tuttavia, potrebbero includere riferimenti a fonti esterne, concetti accademici o traduzioni di materiale didattico fornito dai docenti o presente in libri di testo.

Se ritieni che questo documento contenga materiale di tua proprietà intellettuale e desideri richiederne la modifica o la rimozione, ti invito a contattarmi. Sarò disponibile a risolvere la questione nel minor tempo possibile.

In quanto autore di questi appunti non posso garantire l'accuratezza, la completezza o l'aggiornamento dei contenuti e non mi assumo alcuna responsabilità per eventuali errori, omissioni o danni derivanti dall’uso di queste informazioni. L’uso di questo materiale è a totale discrezione e responsabilità dell’utente.