Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica, Economia e Finanza

RELAZIONE FINALE

UN APPROCCIO PER LA COMBINAZIONE DI BIOMARCATORI

Relatore: Prof. Gianfranco Adimari
Dipartimento di Scienze Statistiche

Laureando: Pietro Belloni
Matricola N. 1073223

Anno Accademico 2015/2016

A Loredana e Giancarlo

Indice

Cap. 3: Un'applicazione su dati reali

Introduzione

Con il passare degli anni, la medicina ha progressivamente affinato le tecniche di diagnosi per minimizzare i casi in cui uno strumento diagnostico non produce una valutazione dello stato del paziente corrispondente alla realtà. Di recente, l'utilizzo di test diagnostici e di biomarcatori sempre più accurati ha ridotto sensibilmente il numero di casi classificati incorrettamente, anche se molti strumenti rimangono comunque soggetti a errore statistico. Nella prima parte di questa relazione, verrà spiegato nel dettaglio il funzionamento dei test diagnostici e dei parametri con cui essi vengono descritti (accuratezza, sensibilità e specificità). Successivamente si delineerà il principale strumento di valutazione di un test diagnostico, ovvero la curva *ROC* e l'area sottesa ad essa, esponendo anche i casi multidimensionali che ne derivano come la superficie *ROC* e il *VUS*. Nella seconda parte si discuterà di come combinare due o più biomarcatori al fine di aumentarne l'accuratezza diagnostica. Verranno esposti alcuni nuovi approcci per le combinazioni lineari ottimali di biomarcatori, approcci che hanno come obiettivo principale quello di diminuire la complessità computazionale dei metodi già esistenti. Saranno anche citati alcuni recenti studi effettuati su questi argomenti. Nella terza e ultima parte della relazione, gli approcci discussi in precedenza verranno applicati ad alcuni dati di natura medica provenienti dall'Università di Harvard. Lo scopo sarà quello di creare una combinazione ottimale di biomarcatori che abbia un'accuratezza diagnostica alta e che non richieda costi computazionali elevati in termini di tempo o memoria utilizzata. Per affrontare questo problema, verranno presentati alcuni algoritmi in linguaggio R.

CAPITOLO 1: Test diagnostici, accuratezza e biomarcatori

In questo capitolo verranno trattati i concetti chiave utili per definire in maniera accurata le ipotesi e gli approfondimenti trattati nelle successive parti della relazione.

1.1 I test diagnostici

In campo medico, un test diagnostico è un qualunque strumento di misurazione utile all'identificazione di uno stato di una determinata malattia: se l'esito di un test è positivo si sospetta la presenza della malattia, se invece l'esito è negativo si tende a escludere la presenza della malattia. Alcuni comuni test diagnostici sono: il Pap-test per diagnosticare il tumore del collo dell'utero, la mammografia per identificare il cancro alla mammella o Breath-test al sorbitolo, utile a riconoscere la celiachia. I test posso offrire diversi tipi di risultati:

- risultato qualitativo dicotomico (sano/malato, positivo/negativo, …);
- risultato qualitativo politomico a più classi (più stadi di avanzamento di una malattia);
- risultato quantitativo discreto o continuo.

Nell'ultimo caso si può individuare un valore soglia (*cut-off*) che discrimina i valori degli individui sani da quelli malati, o una serie di *cut-off* che discriminano vari stadi di avanzamento di una malattia. Molti test sono soggetti a errori sistematici: non

necessariamente classificano correttamente gli individui in sani e malati, ma possono classificare come sano un individuo malato o come malato un individuo sano. Ponendosi nella situazione di un test con risultato quantitativo dicotomizzato da un singolo valore di *cut-off*, si possono dunque distinguere quattro possibili esiti:

- vero positivo (*true positive*, *TP*): il valore del test è positivo e il paziente è malato, dunque la diagnosi coglie la reale situazione della patologia;
- vero negativo (*true negative*, *TN*): il valore del test è negativo e il paziente non è malato, dunque la diagnosi coglie la reale situazione della patologia;
- falso positivo (*false positive*, *FP*): il valore del test è positivo ma il paziente non è realmente malato, dunque la diagnosi non coglie la reale situazione della patologia;
- falso negativo (*false negative*, *FN*): il valore del test è negativo ma il paziente in realtà è malato, dunque la diagnosi non coglie la reale situazione della patologia.

Queste quattro situazioni possono essere riassunte in forma matriciale, in una tabella di errata classificazione o matrice di confusione:

		Vero stato del paziente		
Risultato del test		Malato	Sano	Totale
	Positivo	Veri positivi (*TP*)	Falsi positivi (*FP*)	*TP* + *FP*
	Negativo	Falsi negativi (*FN*)	Veri negativi (*TN*)	*FN* + *TN*
	Totale	*TP* + *FN*	*FP* + *TN*	

Tabella 1.1 – Matrice di confusione

Un'ulteriore rappresentazione degli esiti di un test con un risultato continuo si ha raffigurando le distribuzioni dei sani e dei malati, assieme al livello di *cut-off*:

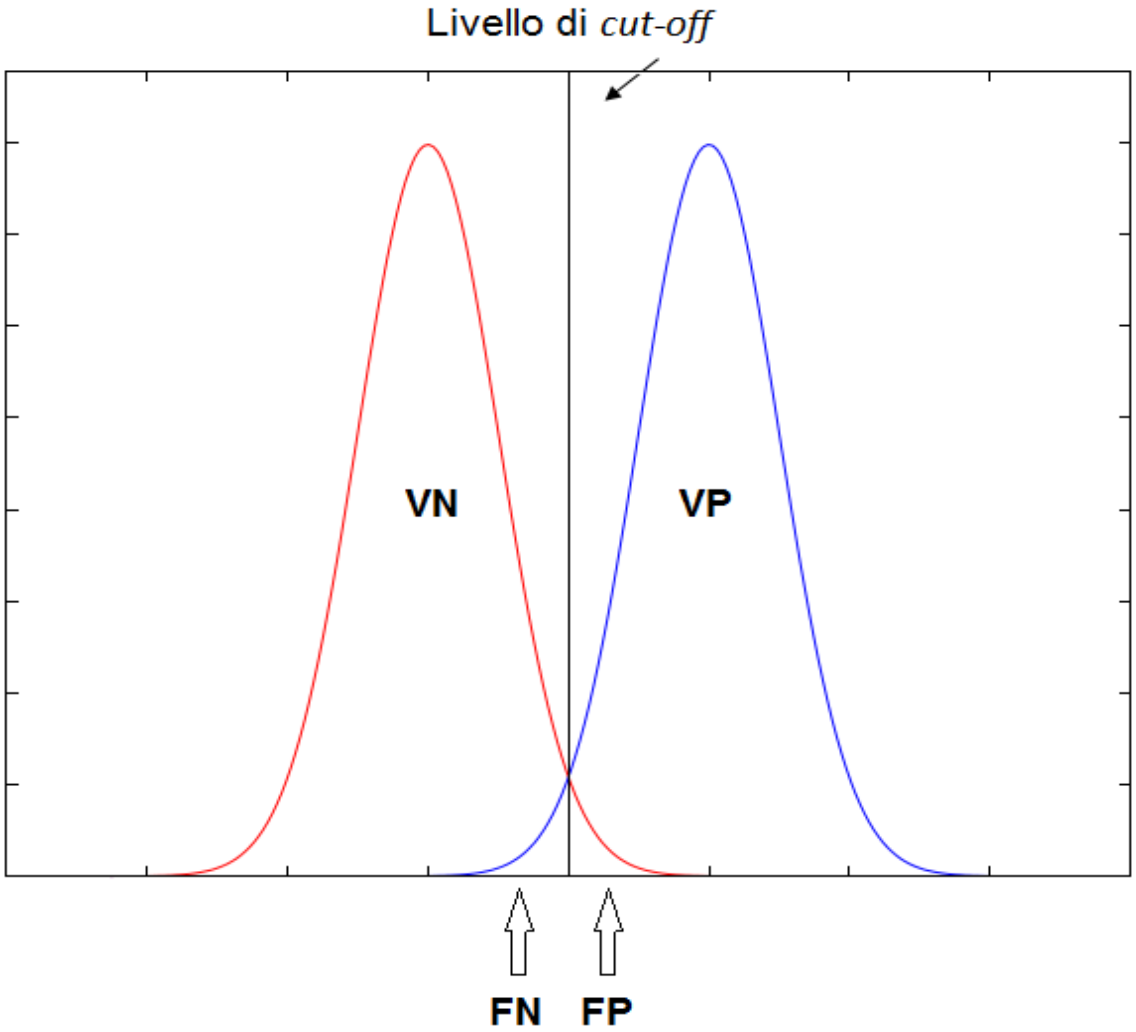

Figura 1.1 - Distribuzioni delle popolazioni di sani (in rosso) e malati (in blu)

Un test diagnostico che produce un numero di falsi positivi e falsi negativi uguale a zero è detto *gold standard*: questo test è privo di errore e discrimina correttamente i sani dai malati. Purtroppo, un test *gold standard* non sempre è applicabile (perché troppo invasivo o troppo costoso), dunque si ricorre all'utilizzo di altri test la cui affidabilità non è completa, ma sono meno costosi e sicuri per il paziente.

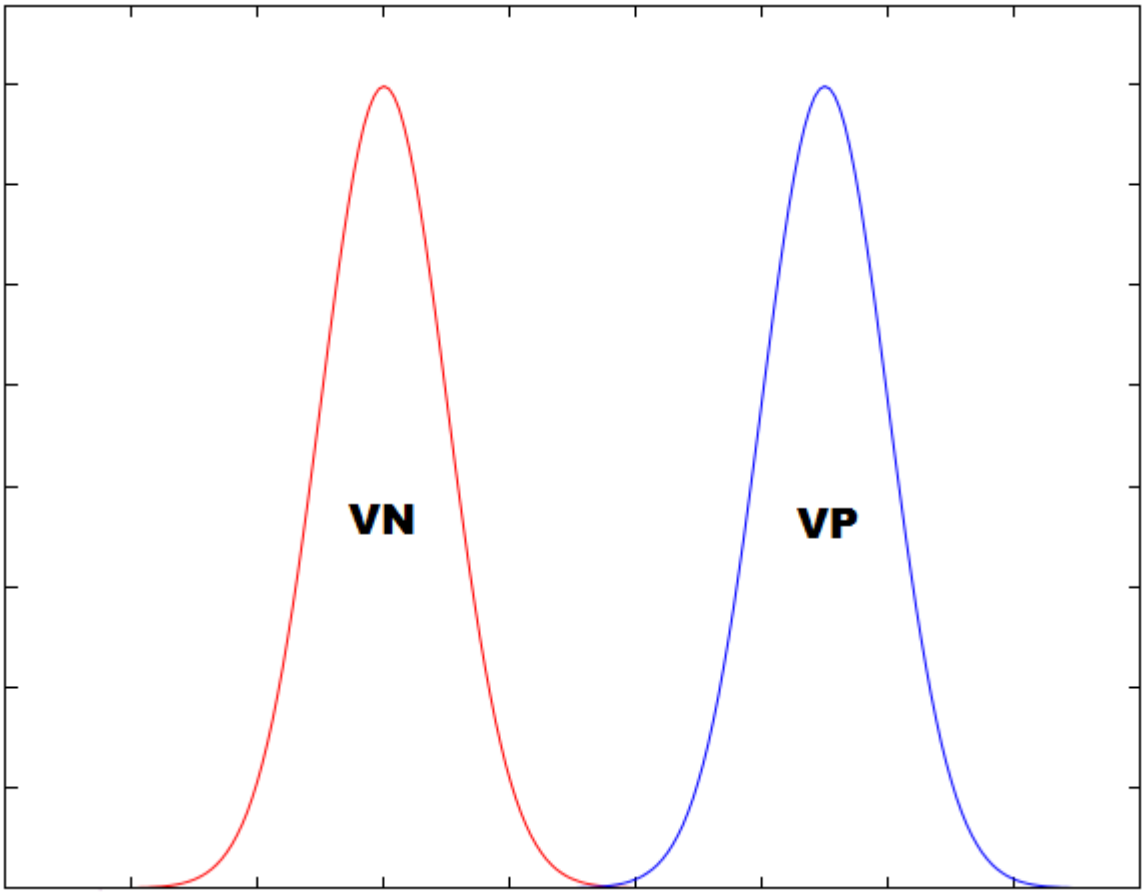

Figura 1.2 - Distribuzioni delle popolazioni di sani (in rosso) e malati (in blu) nel caso di un test gold standard

1.2 Accuratezza, sensibilità e specificità

L'affidabilità dei test viene misurata attraverso l'indice di accuratezza, ovvero la capacità di discriminare correttamente gli individui nei due gruppi (sani e malati):

$$Accuratezza = \frac{TP + TN}{TP + TN + FP + FN}.$$

L'accuratezza si identifica come la somma di veri negativi e veri positivi sul totale degli individui e assume valori nell'intervallo $[0, 1]$: un *gold standard* avrà l'accuratezza sempre pari a 1. Due ulteriori indici di bontà di un test diagnostico sono la sensibilità e la specificità, ovvero le proporzioni di veri positivi sul totale dei malati e falsi negativi sul totale dei sani:

$$Sensibilità = \frac{TP}{TP + FN}$$

$$Specificità = \frac{TN}{TN + FP}.$$

Un test è molto sensibile quando la sua proporzione di falsi negativi è bassa, di conseguenza saranno pochi i soggetti malati erroneamente classificati dal test come sani; al contrario, se il test ha un'ottima specificità, si ha un basso rischio di falsi positivi. Entrambe le misure dipendono dal livello di *cut-off*, il cui valore può essere "spostato" per favorire, all'interno dello stesso test, un maggior livello di sensibilità o specificità. Un'alta sensibilità è da preferirsi nel caso in cui si vogliano individuare malattie molto gravi o altamente infettive; quando invece si ha a che fare con malattie poco gravi o con cure estremamente invasive potrebbe risultare più conveniente aumentare la specificità. Chiamato *k* il livello di *cut-off*, si può pensare alla sensibilità e alla specificità come a due funzioni di *k*: dette F_1 e F_2 le funzioni di distribuzione cumulate riguardo ai valori assunti dal test rispettivamente per gli individui sani e per quelli malati, si ha che:

$$Sensibilità = 1 - F_2(k) = \Pr(T > k \mid paziente\ malato)$$

$$Specificità = F_1(k) = \Pr(T < k \mid paziente\ sano)$$

con *T* risultato del test.

1.3 La curva *ROC*

La curva *ROC* (*Receiver Operating Characteristic*) è uno strumento statistico utilizzato in molti campi della scienza e dell'ingegneria. È stata sviluppata per la prima volta durante la Seconda Guerra Mondiale per lo studio del rapporto fra segnale e disturbo

nell'analisi delle immagini radar, ma viene presto impiegata in altri campi di ricerca, tra cui quello medico (a partire dagli anni '70). Tramite la curva *ROC*, infatti, è possibile valutare la performance di un test diagnostico al variare del livello di *cut-off*. Matematicamente, la curva *ROC* è un luogo geometrico così descritto:

$$\{1 - F_1(k), 1 - F_2(k)\} \quad \forall\, k \in \mathbb{R}$$

con *k* livello di *cut-off* e F_1 , F_2 funzioni precedentemente descritte. In altri termini, la curva *ROC* presenta, al variare del valore di *cut-off*, in ascissa il complemento a 1 della specificità e in ordinata la sensibilità di un certo test diagnostico.

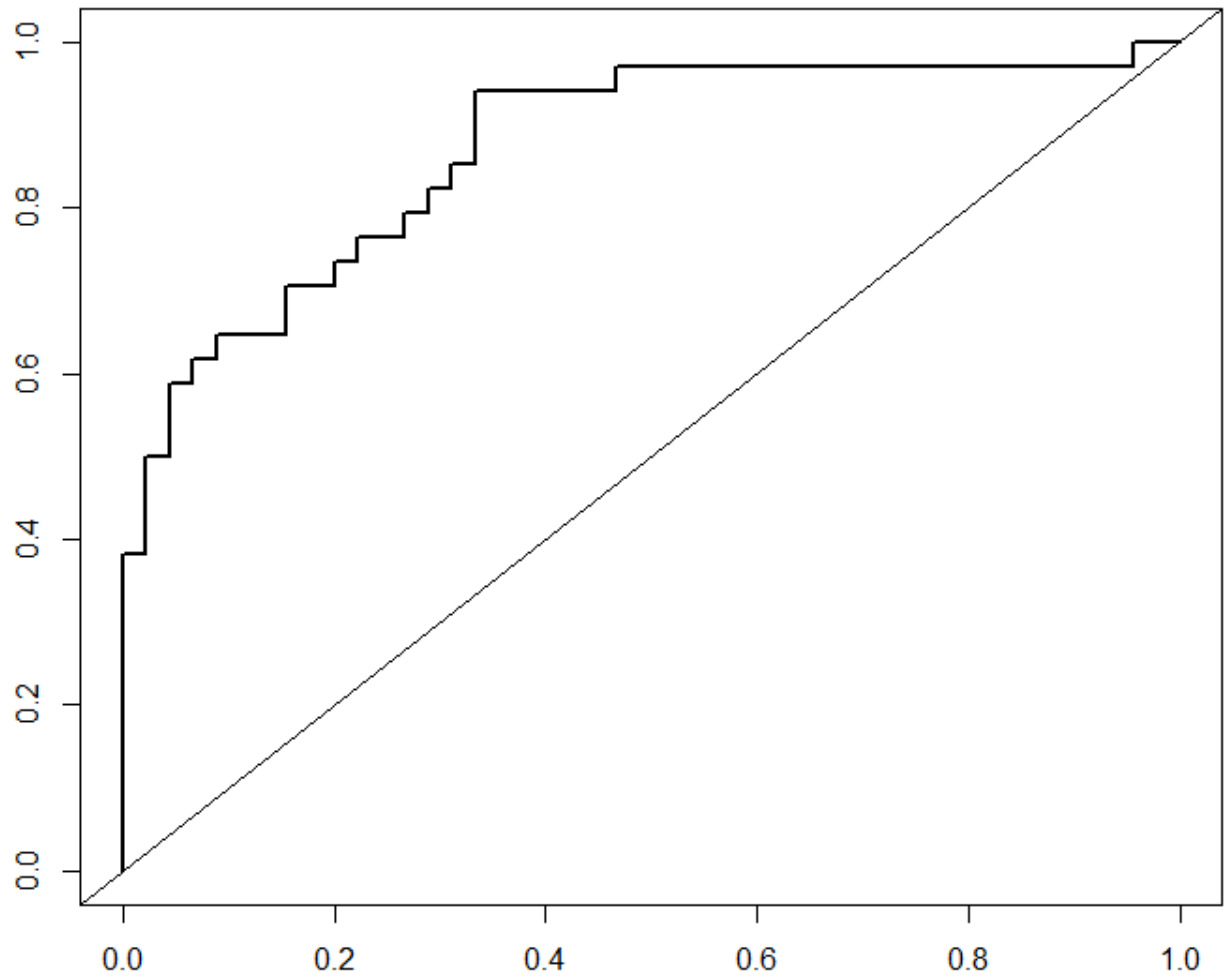

Figura 1.3 - Un esempio di curva ROC

Dato che la sensibilità e il complemento a 1 della specificità variano entrambi nell'intervallo [0, 1], la curva *ROC* si trova sempre all'interno del quadrato di vertici (0, 0), (0, 1), (1, 0), (1, 1). La bisettrice rappresenta il particolare caso in cui il test non è informativo poiché classifica casualmente i sani e i malati: più la curva *ROC* si trova al di sopra della bisettrice, migliore sarà il test diagnostico analizzato. La curva *ROC* di un test *gold standard* sarà composta dall'unione del segmento di estremi (0, 0) e (0, 1) e del segmento di estremi (0, 1) e (1, 1).

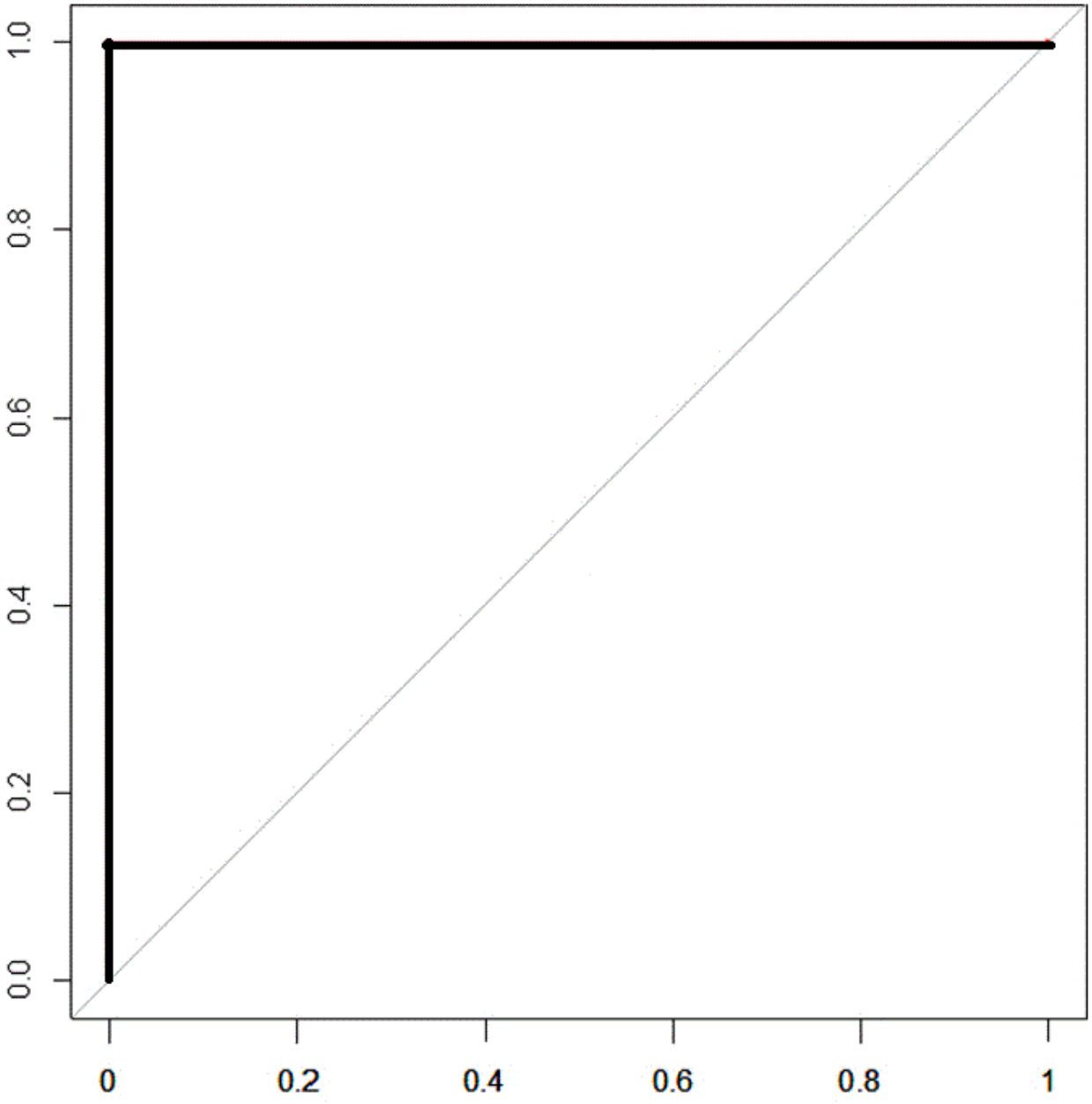

Figura 1.4 – La curva ROC di un test gold standard

1.4 L'*AUC*

La curva *ROC* può essere utilizzata per valutare la performance di un test diagnostico o per confrontare tra loro più test differenti nell'ambito della stessa malattia: la curva più vicina alla bisettrice corrisponde al test peggiore. Una misura sintetica e precisa per eseguire queste valutazioni è l'area sottesa alla curva *ROC* (detta *AUC*, *Area Under Curve*), che rappresenta l'efficacia del test in esame. Nel caso di un test non informativo l'*AUC* assume un valore pari a 0.5, mentre nel caso di un test privo di errore l'*AUC* sarà uguale a 1; di conseguenza, l'*AUC* assume sempre valori nell'intervallo [0.5, 1]. Per l'interpretazione dei valori assunti dall'*AUC*, si usa la seguente classificazione (Swets, 1998):

- *AUC* = 0.5: test non informativo;
- 0.5 < *AUC* ≤ 0.7: test poco accurato;
- 0.7 < *AUC* ≤ 0.9: test moderatamente accurato;
- 0.9 < *AUC* < 1.0: test altamente accurato;
- *AUC* = 1.0: test perfetto.

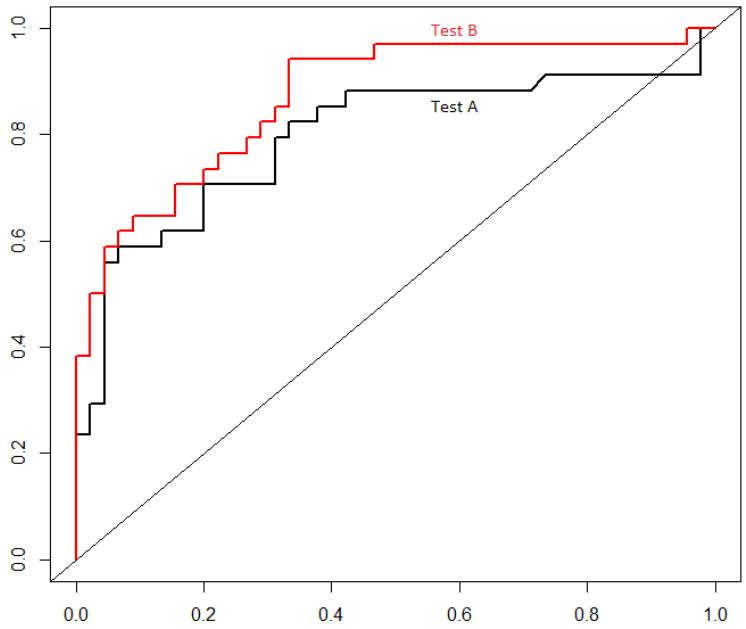

Figura 1.5 – Confronto fra test diagnostici mediante curva ROC

Nella figura 1.5 si nota facilmente come due test diagnostici possano essere confrontati tramite la curva *ROC* e l'*AUC*: il test A ha una performance peggiore del test B (dato che la sua curva *ROC* si avvicina maggiormente alla bisettrice), infatti si ha che $AUC_{Test\ A} = 0.7951$ mentre $AUC_{Test\ B} = 0.8706$.

Data S_1 la variabile che rappresenta i risultati di un test diagnostico nella popolazione dei pazienti sani e data S_2 la corrispondente variabile per il gruppo dei pazienti malati, si calcola il corrispondente *AUC* attraverso la formula (Bamber, 1975):

$$AUC = \Pr(S_1 < S_2),$$

o calcolando l'integrale:

$$AUC = \int_0^1 \left[1 - F_2\left(F_1^{-1}(1-t)\right)\right] \mathrm{d}t$$

con F_1 e F_2 funzioni prese in esame nel paragrafo 1.2. Queste espressioni (matematicamente equivalenti) evidenziano che l'*AUC* non è altro che la probabilità che un soggetto estratto a caso dalla popolazione dei malati abbia il risultato del test diagnostico maggiore di un soggetto estratto a caso dalla popolazione dei sani. In molte situazioni è corretto assumere la normalità nella distribuzione dei risultati dei test diagnostici: si ha quindi che $S_i \sim N\left(\mu_i, \sigma_i^2\right)$ con $i = 1, 2$. Sotto questa ipotesi si può calcolare l'*AUC* attraverso la formula:

$$AUC = \Phi\left(\frac{\mu_2 - \mu_1}{\sqrt{\sigma_2^2 + \sigma_1^2}}\right).$$

Nel caso in cui i test diagnostici non rispettino l'ipotesi di normalità, si può stimare l'*AUC* contando quante volte il valore del test nella popolazione dei sani precede quello della popolazione dei malati e dividendo il tutto per il prodotto della numerosità dei due gruppi:

$$\widehat{AUC} = \frac{1}{n_1 \cdot n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left(S_{1,i} < S_{2,j}\right)$$

con $I(\cdot)$ funzione indicatrice. Si nota facilmente lo stretto legame presente tra la stima non parametrica dell'*AUC* e la statistica di Mann-Whitney:

$$T_{MW} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left(S_{1,i} < S_{2,j}\right) \qquad \text{dunque} \qquad \widehat{AUC} = \frac{T_{MW}}{n_1 n_2}.$$

Il test di Mann-Whitney è il corrispondente non parametrico del t-test a due campioni. Il vantaggio rispetto al t-test è che l'unica assunzione riguardo la distribuzione di provenienza dei dati è che essi possano essere ordinati completamente. Il test si basa sull'assunzione che, se si estraggono due campioni dalle due popolazioni S_1 e S_2, e non è vera l'ipotesi sulla loro uguaglianza, il campione complessivo non produrrà una sequenza di ranghi equamente distribuita tra gli elementi della prima e della seconda popolazione ma, seguendo l'ipotesi alternativa, presenterà un addensamento di ranghi corrispondenti a S_1 (S_2) tra quelli più bassi (più alti). Per scrivere un sistema d'ipotesi per il test di Mann-Whitney, si può utilizzare la mediana delle distribuzioni:

$$\begin{cases} H_0 : Me_{S_1} = Me_{S_2} \\ H_1 : Me_{S_1} \neq Me_{S_2} \end{cases}$$

o, alternativamente, si possono usare le funzioni cumulate:

$$\begin{cases} H_0 : F_1 = F_2 \\ H_1 : F_1 \neq F_2 \end{cases}.$$

1.5 I biomarcatori

Per definizione, un marcatore biologico (o biomarcatore) è un "indicatore di un processo fisiologico, patologico o di risposta biologica a un intervento terapeutico"

(Agenzia Italiana del Farmaco, 2014). Di fatto, i biomarcatori sono delle sostanze o delle attività adoperate per misurare uno stato biologico, possono essere utilizzati da soli o in combinazione tra loro e, perché un biomarcatore sia valido, i risultati prodotti devono essere accurati e riproducibili. Nella medicina moderna è utilizzato un largo numero di biomarcatori: per ogni sistema biologico interno all'organismo umano esistono uno o più biomarcatori atti a misurarne le caratteristiche e lo stato di salute. Molti biomarcatori sono semplici misure utilizzate giornalmente negli esami medici di routine: pressione sanguigna, temperatura corporea, livello di colesterolo nel sangue…

Il risultato di un biomarcatore perde valore quando dipende da un elevato numero di variabili che possono alterarne l'efficienza. Al contrario, un buon biomarcatore possiede le seguenti caratteristiche:

- una specifica correlazione con la malattia che si vuole individuare;
- un'adeguata riproduttività sul tipo di trattamento e sulla risposta;
- la possibilità di effettuare la determinazione con precisione e in tempi brevi;
- essere il più possibile insensibile a errori di campionamento.

In genere, un biomarcatore può essere classificato all'interno di una delle tre categorie qui riportate:

1) biomarcatori che forniscono informazioni su una malattia o sul rischio di riscontrare una malattia;
2) biomarcatori che misurano gli effetti di un farmaco (o un'altra sostanza assunta);
3) biomarcatori che misurano l'interazione tra un farmaco e la sua molecola obiettivo.

In questa tesi, ci si soffermerà solamente sui biomarcatori della prima di queste tre classi. Infatti, un test diagnostico con risultato discreto o continuo si basa su una

misura effettuata da un biomarcatore del primo tipo. Più avanti, si parlerà quindi di accuratezza, sensibilità e specificità di un biomarcatore (o di una combinazione di biomarcatori) immaginando che risultati da esso prodotti possano essere utilizzati per costruire un test diagnostico.

1.6 Generalizzazione dei concetti in più dimensioni: (iper)superficie *ROC*

Spesso accade di dover utilizzare un biomarcatore per analizzare una malattia con diversi stadi di avanzamento, oppure di dover distinguere più di due situazioni molto differenti tra loro all'interno della stessa malattia. In questi casi si generalizzano i concetti matematici visti in precedenza aumentandone le dimensioni: la curva *ROC* diventa una superficie *ROC* (tre classi diagnostiche) o una ipersuperficie *ROC* (più di tre classi diagnostiche) e, rispettivamente, si parla di volume (*VUS*, *Volume Under Surface*) o di ipervolume (*HUM*, *Hypervolume Under Manifold*) sotteso a essa. In presenza di n classi diagnostiche ordinate il *cut-off* non è più un singolo valore al di sopra del quale i soggetti verranno classificati come malati, ma si avranno $k_1, k_2, \dots, k_{n-1}$ valori di *cut-off* tali che $k_1 \leq k_2 \leq \dots \leq k_{n-1}$. Di conseguenza, i soggetti saranno classificati in n classi di corretta classificazione (*TCF*, *True Class Fractions*) e in $n^2 - n$ classi di errata classificazione (*FCR*, *False Classification Rates*). Per fare un esempio, si può citare il caso molto comune della presenza di tre classi diagnostiche ordinate: i valori di *cut-off* saranno due e i soggetti verranno classificati nel seguente modo:

- TCF_1: soggetti della classe 1 classificati correttamente;
- TCF_2: soggetti della classe 2 classificati correttamente;

- TCF_3: soggetti della classe 3 classificati correttamente;
- FCR_1: soggetti della classe 2 erroneamente classificati come appartenenti alla classe 1;
- FCR_2: soggetti della classe 1 erroneamente classificati come appartenenti alla classe 2;
- FCR_3: soggetti della classe 3 erroneamente classificati come appartenenti alla classe 2;
- FCR_4: soggetti della classe 2 erroneamente classificati come appartenenti alla classe 3;
- FCR_5: soggetti della classe 1 erroneamente classificati come appartenenti alla classe 3;
- FCR_6: soggetti della classe 3 erroneamente classificati come appartenenti alla classe 1.

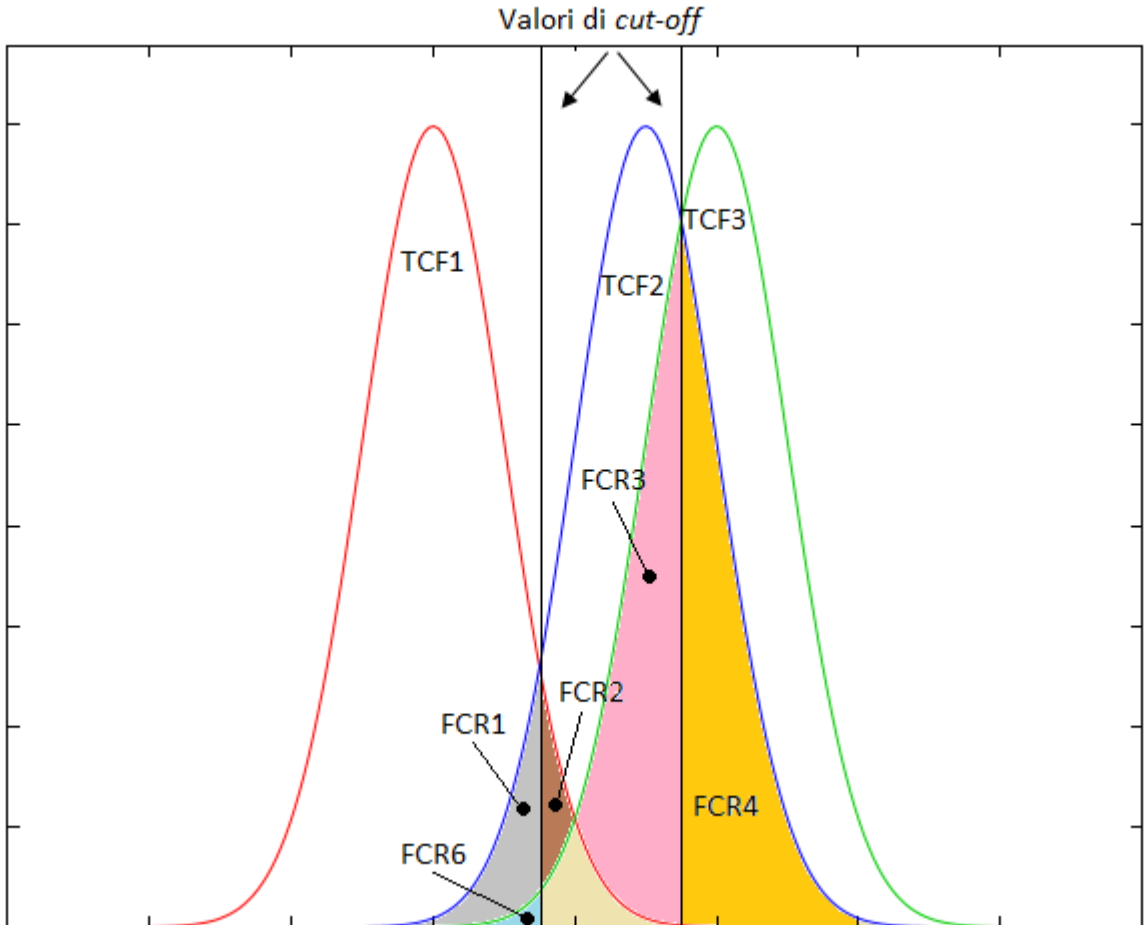

Figura 1.6 – Rappresentazione grafica delle possibili classificazioni in presenza di tre classi diagnostiche

Si vede facilmente che, all'aumentare delle classi diagnostiche, i problemi di classificazione crescono notevolmente. Dalla figura 1.6 si nota che non sempre si osservano tutte le classi diagnostiche: nel caso raffigurato è assente la classe FCR_5 perché la popolazione della classe 1 non supera i valori del secondo *cut-off*.

In analogia con il caso di diagnosi binaria, possiamo identificare con $X = (X_1, \dots, X_{n_1})^T$, $Y = (Y_1, \dots, Y_{n_2})^T$ e $Z = (Z_1, \dots, Z_{n_3})^T$ i risultati del biomarcatore per gli individui delle classi 1, 2 e 3. Ipotizzando che questi tre gruppi si distribuiscano in funzione dei due *cut-off* k_1 e k_2 seguendo le distribuzioni F_1, F_2 e F_3, possiamo definire la superficie *ROC* come un luogo geometrico così descritto:

$$\{ F_1(k_1), \qquad F_2(k_2) - F_2(k_1), \qquad 1 - F_3(k_2) \} \quad \text{con} \quad k_1 \leq k_2.$$

Sia per il caso in esame con tre classi diagnostiche, che per altri casi con più classi, il significato degli indici di accuratezza, sensibilità e specificità dei biomarcatori restano invariati.

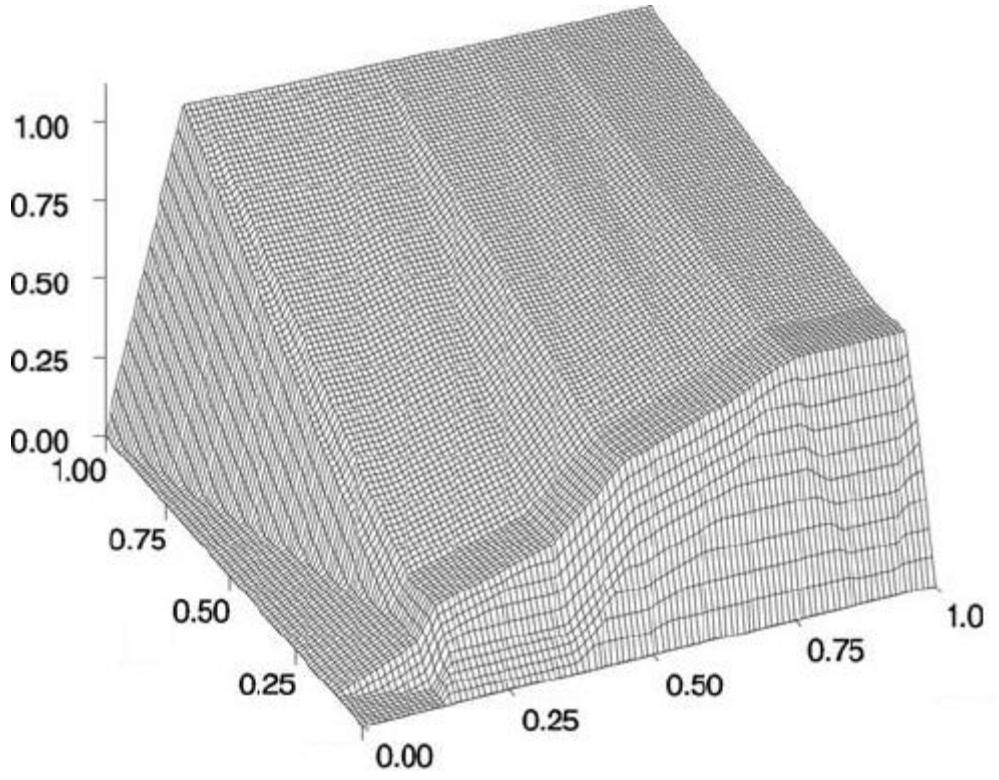

Figura 1.7 - Un esempio di superficie ROC

CAPITOLO 2: Combinazioni lineari ottimali di biomarcatori per diagnosi multi-categoriali

In questo capitolo, si discute di alcuni metodi per combinare in maniera ottimale due o più biomarcatori al fine di incrementarne il potere diagnostico. In particolare si approfondiscono due indici, proposti da M. J. Hsu e Y. H. Chen nel 2014, pensati per ottenere una combinazione lineare ottima di biomarcatori. Questi indici sono collegati all'ipervolume sotteso all'ipersuperficie *ROC* (*HUM*, *Hypervolume Under Manifold*), ma non ne prevedono il calcolo. Dopo aver definito queste due misure, si procederà presentando dei metodi parametrici e non parametrici per ottenere, per mezzo di esse, una combinazione lineare ottima di biomarcatori.

2.1 Le combinazioni ottime di biomarcatori

Nel campo medico, in presenza di un set di biomarcatori atti a diagnosticare una malattia, è comune aumentare l'accuratezza diagnostica ricorrendo ad una combinazione lineare dei biomarcatori a disposizione. La combinazione ottima dei biomarcatori è quella corrispondente al maggior *HUM* sotteso all'ipersuperficie *ROC* costruita utilizzando la combinazione dei biomarcatori come fosse un biomarcatore unico. Infatti, l'*HUM* (così come i suoi corrispondenti *AUC* e *VUS*, rispettivamente nel

caso di due o tre classi diagnostiche) è un indice dell'efficacia di un biomarcatore. È noto che il calcolo dell'*HUM* in presenza di più di tre dimensioni o di un cospicuo numero di biomarcatori da combinare diventa molto lungo e complesso. Questa difficoltà computazionale ha fatto sì che molti ricercatori proponessero misure alternative alla stima dell'*HUM* per la valutazione dell'accuratezza diagnostica. In questo capitolo verranno analizzate due di queste misure che prevedono solo il calcolo dell'area sottesa alla curva *ROC* per il confronto di due categorie diagnostiche adiacenti; la complessità computazionale è quindi ridotta notevolmente sia in termini di tempo che in termini di memoria richiesta. Queste misure sono collegate al limite superiore e al limite inferiore dell'*HUM* e riflettono, rispettivamente, il peggior caso e il caso medio dell'accuratezza diagnostica nel confronto a coppie tra biomarcatori.

2.2 Le disuguaglianze di Fréchet

In questo paragrafo si presenta il concetto necessario per definire le suddette misure di accuratezza diagnostica alterative all'*HUM*: le disuguaglianze di Fréchet. Senza assumere indipendenza o un qualunque tipo di dipendenza, esse delimitano un intervallo nel quale rientra la probabilità di due o più eventi legati tra loro da un operatore logico di intersezione o di unione. Nel caso più semplice, ovvero quello con solo due eventi A e B, le disuguaglianze di Fréchet sono definite come segue:

$$max(0, P(A) + P(B) - 1) \leq P(A \cap B) \leq min(P(A), P(B))$$

$$max(P(A), P(B)) \leq P(A \cup B) \leq min(1, P(A) + P(B)).$$

Generalizzando: siano $A_1, A_2, \dots, A_n$ degli eventi, le disuguaglianze di Fréchet affermano che la probabilità della loro intersezione è compresa nell'intervallo:

$$max(0, P(A_1) + \cdots + P(A_n) - (n-1)) \leq P(A_1 \cap \dots \cap A_n) \leq min(P(A_1), \dots, P(A_n)),$$

e che la probabilità della loro unione è compresa nell'intervallo:

$$max(P(A_1), \dots, P(A_n)) \leq P(A_1 \cup \dots \cup A_n) \leq min(1, P(A_1) + \dots + P(A_n)).$$

Il primo di questi intervalli può essere applicato alla formula per il calcolo dell'*HUM*, dato che essa si basa su una serie di intersezioni di probabilità.

2.3 Due nuove misure per l'accuratezza diagnostica

Supponiamo che esista una malattia con *M* categorie diagnostiche (per esempio, un tumore con *M* stadi di avanzamento) e chiamiamo $X_1, \dots, X_M$ i vettori *p*-dimensionali contenenti i valori delle diagnosi eseguite da *p* biomarcatori diagnostici nelle *M* classi. Chiamiamo β un vettore, sempre *p*-dimensionale, contenente i "pesi" della combinazione lineare ($\beta^T X_1 + \dots + \beta^T X_M$) dei biomarcatori sopracitati. L'*HUM* riflette la potenza diagnostica della combinazione dei biomarcatori, ed è calcolabile attraverso la formula:

$$HUM = \Pr(\beta^T X_M > \beta^T X_{M-1} > \dots > \beta^T X_1).$$

Le misure proposte da M. J. Hsu e Y. H. Chen alternativamente al calcolo dell'HUM sono le seguenti:

$$P_A = \sum_{i=1}^{M-1} \frac{\Pr(\beta^T X_{i+1} > \beta^T X_i)}{M-1}$$

$$P_M = min_{1 \leq i \leq M-1}(\Pr(\beta^T X_{i+1} > \beta^T X_i))$$

ne segue, per le disuguaglianze di Fréchet, che

$$max\,(0,\ (M-1)P_A - (M-2)\,) \leq HUM \leq P_M.$$

È quindi chiaro come P_A e P_M siano relazionate, rispettivamente, al valore minimo e al valore massimo dell'*HUM*. Inoltre, si nota immediatamente come le difficoltà computazionali per queste due quantità siano decisamente inferiori rispetto al calcolo dell'*HUM*, dato che prevedono solo confronti di coppie di probabilità (calcolabili con un integrale) e non un confronto congiunto di M probabilità (calcolabile con un integrale M-dimensionale). In altre parole, si può dire che P_A rappresenta il caso medio e P_M il peggior caso di accuratezza diagnostica misurata sulle coppie adiacenti di biomarcatori: è questa stessa idea a suggerire che le due quantità servono da indice per la valutazione dell'accuratezza nelle diagnosi multi-categoriali. Di seguito verranno presentati tre metodi per ottenere la combinazione lineare ottima di biomarcatori usando gli indici P_A e P_M appena visti.

2.4 Metodi per ottenere la combinazione ottima di biomarcatori

2.4.1 Il metodo parametrico

Il primo metodo è basato sull'assunzione di normalità dei biomarcatori diagnostici. Supponiamo che ogni variabile X_i si distribuisca come una normale p-variata di media μ_i e matrice di varianze-covarianze Σ_i:

$$X_i \sim N_p(\mu_i, \Sigma_i) \qquad \text{con} \qquad i = 1, \dots, M.$$

Sempre chiamando β il vettore dei pesi per la combinazione lineare, si otterrà:

$$\beta^T X_i \sim N_p(\beta^T \mu_i, \beta^T \Sigma_i \beta) \qquad \text{con} \qquad i = 1, \dots, M.$$

Con queste premesse, siano X_h e X_g i vettori dei biomarcatori relativi a due categorie diagnostiche adiacenti. Si potrà calcolare l'$AUC_{h,g}$ in funzione di β sfruttando l'ipotesi di normalità attraverso la formula:

$$AUC_{h,g}(\beta) = \Pr\left(\beta^T X_h > \beta^T X_g\right) = \int_0^1 \Phi\left(\frac{\beta^T(\mu_h - \mu_g) - c(u)\sqrt{\beta^T \Sigma_g \beta}}{\sqrt{\beta^T \Sigma_h \beta}}\right) du$$

con $c(u) = \Phi^{-1}(1-u)$ e μ_i , Σ_i medie e matrici varianze-covarianze stimate dai risultati dei rispettivi biomarcatori.

È possibile inserire i valori di *AUC* così calcolati nelle espressioni di P_A e P_M ottenendo una funzione (funzione obiettivo) da massimizzare relativamente a β: il β che massimizza la funzione sarà quello corrispondente alla combinazione ottimale dei biomarcatori. Generalmente, il vettore che massimizza P_A è diverso da quello che massimizza P_M , salvo il caso specifico in cui:

$$\mu_2 - \mu_1 = \mu_3 - \mu_2 = \cdots = \mu_M - \mu_{M-1} = \delta \qquad \text{e} \qquad \Sigma_i = \Sigma \qquad \forall i = 1, \dots, M.$$

In questo caso il β che massimizza P_A è lo stesso che massimizza P_M e sono entrambi proporzionali alla quantità $\Sigma^{-1}\delta$.

2.4.2 Il metodo non parametrico

Il secondo metodo non assume la normalità dei biomarcatori. Per ogni coppia di vettori di biomarcatori X_h e X_g si ricorre alla stima dell'*AUC* legata alla statistica di Mann–Whitney tramite la relazione:

$$\widehat{AUC}_{h,g} = \frac{T_{MW}}{n_g n_h} \qquad \text{con} \qquad T_{MW} = \sum_{s=1}^{n_g} \sum_{r=1}^{n_h} d_{sr}$$

dove

$$d_{sr} = \begin{cases} 1 \; se \; \beta^T X_{h,r} > \; \beta^T X_{g,s} \\ 0 \; se \; \beta^T X_{h,r} < \; \beta^T X_{g,s} \end{cases}.$$

n_i denota la numerosità del campione tratto dalla *i*-esima categoria diagnostica e $X_{i,t}$ denota il vettore dei risultati dei biomarcatori per il *t*-esimo soggetto dell'*i*-esima categoria diagnostica. Si potrebbe, come nel caso parametrico, inserire il valore $\widehat{AUC}$ così calcolato all'interno delle espressioni di P_A e P_M, ottenendo una funzione obiettivo, e trovare il vettore β che la massimizza. Tuttavia, in questo caso si presenterebbero delle funzioni non lisce (la presenza di una funzione indicatrice crea dei veri e propri "gradini" nella funzione obiettivo) che risultano molto complesse da massimizzare poiché potrebbero non essere differenziabili in tutti i punti. Per ovviare a questo problema, si propongono due procedure iterative: la procedura *step-down* e la procedura *step-up*.

Procedura *step-down*:

1. Per ognuno dei *p* biomarcatori, si trova l'accuratezza diagnostica attraverso l'indice P_A (o P_M) usando la stima non parametrica dell'*AUC* tra coppie di classi diagnostiche adiacenti.
2. Osservando l'indice P_A (o P_M) si ordinano i *p* biomarcatori in ordine decrescente: per primo si metterà il biomarcatore con il maggior P_A (o P_M), per secondo quello con il secondo valore di P_A (o P_M) …
3. Si crea una combinazione lineare con i primi due biomarcatori: $V = X_{(1)} + \alpha X_{(2)}$. Si utilizza la stima non parametrica dell'*AUC* in P_A (o P_M) per trovare la combinazione ottima dei biomarcatori $X_{(1)}$ e $X_{(2)}$. Essa corrisponde a trovare il termine $\alpha \in [-1, 1]$ che massimizza una funzione obiettivo non liscia ma, siccome l'elemento da massimizzare è uno scalare e non un vettore, la complessità computazionale diminuisce sensibilmente.

4. Una volta trovato il termine α del punto precedente, si ritorna al punto 3. rimpiazzando $X_{(1)}$ con il nuovo biomarcatore *V* ottenuto dalla precedente combinazione lineare e $X_{(2)}$ con $X_{(3)}$.
5. Si ripete il punto 4. finché tutti i *p* biomarcatori sono inclusi nella combinazione lineare *V*.

La procedura *step-up* si realizza nello stesso modo della procedura *step-down*, fatta eccezione per il punto 2.: i biomarcatori vengono ordinati rispetto al loro indice P_A (o P_M) in maniera crescente.

2.4.3 Procedura min-max

La procedura min-max, esposta da C. Liu et altr. in un articolo del 2011, viene qui impiegata per trovare la combinazione ottima di biomarcatori attraverso i nuovi indici P_A e P_M . Anziché considerare l'intera combinazione lineare dei biomarcatori ordinati attraverso P_A (o P_M), si crea una combinazione lineare che comprende solo il primo e l'ultimo dei biomarcatori nell'ordinamento di cui al punto 2 del paragrafo 2.4.2:

$$V = X_{(max)} + \alpha X_{(min)}$$

Dunque, la stima non parametrica dell'*AUC* tra le classi diagnostiche *h* e *g* diventa:

$$\widehat{AUC}_{h,g} = \frac{\sum_{s=1}^{n_g}\sum_{r=1}^{n_h} I\left(X_{h,r,(max)} + \alpha X_{h,r,(min)} > X_{g,s,(max)} + \alpha X_{g,s,(min)}\right)}{n_g n_h}$$

con *I(x)* funzione indicatrice. Il valore di α che massimizza la funzione obiettivo in termini di P_A (o P_M) corrisponde alla combinazione ottima dei biomarcatori $X_{(max)}$ e $X_{(min)}$. Questa procedura è molto approssimativa rispetto ai metodi precedenti, ma è decisamente più semplice dal punto di vista computazionale.

I metodi precedentemente esposti sono stati sperimentati da M. J. Hsu e Y. H. Chen nel 2014 su due dataset, il primo con tre categorie diagnostiche ordinate e il secondo con cinque categorie diagnostiche non ordinate. Quando le categorie diagnostiche non sono ordinate, il loro ordinamento è quello che corrisponde al massimo indice P_A (o P_M).

2.4.4 Dataset "AL"

Nel primo dataset, contenuto all'interno del pacchetto R "DiagTest3Grp", vengono utilizzati 14 biomarcatori neuropsicologici su 118 pazienti soggetti ad Alzheimer. Il dataset è composto da 118 osservazioni e 15 variabili: la prima variabile indica lo stato della malattia (assenza, fase intermedia, malattia avanzata) e le rimanenti sono vettori contenenti i risultati dei 14 biomarcatori. Su questi dati, viene calcolata la combinazione ottimale dei 14 biomarcatori attraverso i seguenti criteri:

- criterio *naïve*: a tutti i biomarcatori viene dato lo stesso peso. Se uno dei criteri dovesse generare una combinazione di biomarcatori con un'accuratezza più bassa di questo criterio, sarebbe sicuramente da scartare;
- criterio P_A parametrico: viene utilizzato l'indice P_A come funzione obiettivo e viene assunta la normalità dei biomarcatori;
- criterio P_M parametrico: viene utilizzato l'indice P_M come funzione obiettivo e viene assunta la normalità dei biomarcatori;
- criterio P_A non parametrico: viene utilizzato l'indice P_A come funzione obiettivo e non viene assunta la normalità dei biomarcatori (procedura non parametrica *step-down*);
- criterio P_M non parametrico: viene utilizzato l'indice P_M come funzione obiettivo e non viene assunta la normalità dei biomarcatori (procedura non parametrica *step-down*);

- criterio del massimo *HUM*: è sicuramente il criterio più preciso, ma comporta difficoltà computazionali elevate (come evidenziato nei capitoli precedenti).

Per ognuna di queste combinazioni, viene calcolata l'accuratezza diagnostica tramite la stima degli indici P_A e P_M e tramite la stima dell'*HUM*. Dai risultati riportati da Hsu e Chen, si nota che le combinazioni lineare ottenute tramite i criteri P_A (parametrico e non parametrico) presentano i maggiori indici di accuratezza rispetto agli altri criteri.

2.4.5 Dataset "processed.cleveland"

Nel secondo dataset, reperibile all'indirizzo *archive.ics.uci.edu/ml/machine-learning-databases/heart-disease*, vengono utilizzati 4 biomarcatori su 303 soggetti malati di cuore. Il dataset è composto da 303 osservazioni e 14 variabili. Tra queste sono d'interesse le variabili `trestbps`, `chol`, `thalach` e `oldpeak`, che corrispondono ai vettori con i risultati dei 4 biomarcatori, e la variabile `num`, che divide i soggetti in 5 categorie (ognuna corrispondente a una diversa modalità della malattia). A differenza del caso precedente, la distribuzione dei biomarcatori non segue un ordine specifico in relazione ai 5 stadi della malattia. Anche in questo caso, viene calcolata la combinazione ottima dei biomarcatori attraverso i metodi *naïve*, P_A parametrico, P_M parametrico, P_A non parametrico (procedura *step-down*), P_M non parametrico (procedura *step-down*) e del massimo *HUM*. Come nel caso precedente, viene calcolata l'accuratezza diagnostica stimando P_A , P_M e *HUM* delle varie combinazioni di biomarcatori. I risultati ricalcano i precedenti: le combinazioni ottenute tramite i metodi P_A parametrico e P_A non parametrico hanno un'accuratezza diagnostica migliore delle combinazioni ottenute attraverso gli atri metodi. In questo caso, l'utilizzo del criterio del massimo *HUM* per trovare la combinazione ottimale di biomarcatori sarebbe impossibile da applicare senza calcolatori molto potenti: è richiesto, infatti, lo svolgimento di una sommatoria cinque-dimensionale. Gli stessi

autori non calcolano la combinazione con il criterio del massimo *HUM*, ma fanno ricorso a un risultato ottenuto da J. Li, Y. Chow, W. K. Wong e T. Y. Wong nel 2014.

2.4.6 Conclusioni

Visti i risultati delle applicazioni su dataset, tra le due misure proposte gli autori notano che P_A opera in maniera più efficiente, creando combinazioni con maggiore accuratezza diagnostica. Inoltre, da studi di simulazione non riportati in questa relazione, si nota che l'approccio parametrico funziona molto bene quando i biomarcatori seguono una distribuzione normale, mentre l'approccio non parametrico produce ottimi risultati sia quando i biomarcatori seguono una distribuzione normale sia quando ciò non accade.

CAPITOLO 3: Un'applicazione su dati reali

In questo capitolo si applicheranno i metodi analizzati precedentemente per combinare dei biomarcatori in maniera ottimale. Tutte le funzioni e le analisi sono state eseguite in linguaggio R nella versione del software 3.3.0; il livello di significatività fissato è pari a $\alpha = 0.05$.

3.1 I dati: descrizione e analisi esplorative

3.1.1 Descrizione dei dati

I dati utilizzati compongono un dataset che raccoglie le misure di 12 biomarcatori effettuate all'Università di Harvard su 279 pazienti affette da tumore alle ovaie. L'obiettivo è cercare di ottenere una combinazione ottimale dei 12 biomarcatori, ovvero una combinazione che abbia un'accuratezza diagnostica maggiore di ogni biomarcatore (questa combinazione sarà di tipo lineare e si otterrà con uno dei metodi presentati in precedenza). Viste le conclusioni tratte nel paragrafo 2.5.3, si adopererà P_A come funzione obiettivo, dato che si è dimostrata più efficiente rispetto a P_M. Le pazienti sono state divise in gruppi, come segue:

- gruppo 1: pazienti sani (gruppo di controllo);
- gruppo 2: pazienti con tumore di tipo benigno;

- gruppo 3: pazienti con tumore di tipo maligno.

Per facilitare le operazioni successive, l'intero dataset viene ordinato in base al gruppo di appartenenza dei soggetti.

3.1.2 Analisi esplorative

Le pazienti non sono distribuite in modo bilanciato all'interno dei tre gruppi: il gruppo di controllo conta 134 pazienti, il gruppo di affette da tumore benigno conta 65 pazienti e il gruppo di affette da tumore maligno ne conta 80. Le distribuzioni percentuali sono visibili nella figura 3.1.

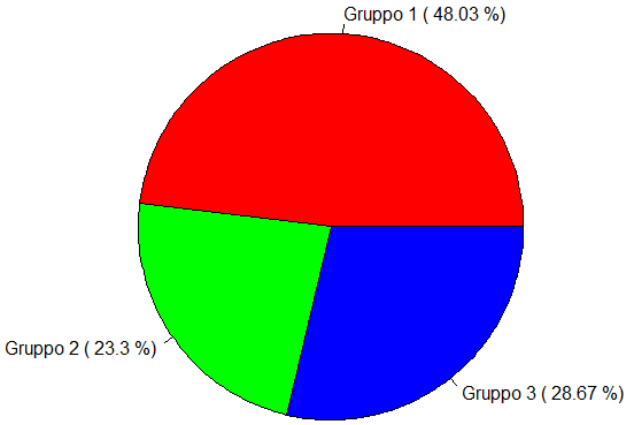

Figura 3.1 – Distribuzione percentuale dei tre gruppi di pazienti

I 12 biomarcatori, identificati dalle sigle CA153, CA125, KLK6, CA724, HE4, DD.X065, YKL40, DD.0110, DD.C248, DD.P108, IGF2 e SMRP, restituiscono risultati di tipo continuo. Le principali statistiche descrittive relative alle loro distribuzioni sono:

CA153	CA125	KLK6	CA724	HE4	DD.X065
Min. :-1.8379	Min. :-1.19152	Min. :-3.6119	Min. :-0.4105	Min. :-1.5332	Min. :-2.0049
1st Qu.:-0.2027	1st Qu.:-0.08759	1st Qu.:-0.3876	1st Qu.:-0.4105	1st Qu.:-0.1368	1st Qu.:-0.5405
Median : 0.3029	Median : 0.87755	Median : 0.3182	Median :-0.4105	Median : 0.5628	Median : 0.3358
Mean : 0.6225	Mean : 1.43042	Mean : 0.7125	Mean : 1.2683	Mean : 1.5478	Mean : 0.2501
3rd Qu.: 1.0214	3rd Qu.: 2.90919	3rd Qu.: 1.3228	3rd Qu.: 1.5817	3rd Qu.: 3.1501	3rd Qu.: 0.9280
Max. : 6.5056	Max. : 7.40016	Max. : 9.0863	Max. :12.6502	Max. : 6.2839	Max. : 2.8001

YKL40	DD.0110	DD.C248	DD.P108	IGF2	SMRP
Min. :-2.0684	Min. :-0.9450	Min. :-3.8192	Min. :-3.7357	Min. :-8.6428	Min. :-2.82041
1st Qu.:-0.1606	1st Qu.:-0.9450	1st Qu.:-0.5117	1st Qu.:-0.3501	1st Qu.:-0.7353	1st Qu.:-0.54091
Median : 0.5056	Median :-0.0414	Median : 0.1334	Median : 0.2214	Median :-0.2061	Median : 0.03111
Mean : 0.5801	Mean : 0.6111	Mean : 0.1350	Mean : 0.3350	Mean :-0.4181	Mean : 0.51467
3rd Qu.: 1.3380	3rd Qu.: 1.2678	3rd Qu.: 0.5092	3rd Qu.: 0.9452	3rd Qu.: 0.1908	3rd Qu.: 1.35834
Max. : 3.0025	Max. : 5.1979	Max. : 4.3965	Max. : 3.7422	Max. : 1.6940	Max. : 7.65117

Tabella 3.1 – Principali statistiche descrittive relative ai biomarcatori

Si deve scegliere se utilizzare il metodo parametrico o il metodo non parametrico per la combinazione dei biomarcatori. A tal scopo si esegue su ogni biomarcatore il test di normalità di Shapiro-Wilk, i cui risultati sono riportati nella tabella 3.2. Si potrebbero eseguire molte altre analisi esplorative sui risultati dei 12 biomarcatori, ma in questa specifica situazione siamo interessati esclusivamente alla loro normalità.

Biomarcatore	**Test di Shapiro-Wilk**	**Biomarcatore**	**Test di Shapiro-Wilk**
CA153	$W = 0.893$ *p-value* < 0.001	YKL40	$W = 0.990$ *p-value* = 0.061
CA125	$W = 0.922$ *p-value* < 0.001	DD.0110	$W = 0.807$ *p-value* < 0.001
KLK6	$W = 0.868$ *p-value* < 0.001	DD.C248	$W = 0.940$ *p-value* < 0.001
CA724	$W = 0.657$ *p-value* < 0.001	DD.P108	$W = 0.975$ *p-value* < 0.001
HE4	$W = 0.842$ *p-value* < 0.001	IGF2	$W = 0.824$ *p-value* < 0.001
DD.X065	$W = 0.977$ *p-value* < 0.001	SMRP	$W = 0.918$ *p-value* < 0.001

Tabella 3.2 – Test di Shapiro-Wilk sui risultai dei 12 biomarcatori

Si nota immediatamente che tutti i biomarcatori si allontanano notevolmente dall'ipotesi di normalità, fatta eccezione per YKL40, per il quale non si rifiuta l'ipotesi nulla se si considera il livello di significatività fissato. Considerando la preponderante non normalità nelle distribuzioni dei biomarcatori, si opta per il metodo non parametrico con procedura *step-down* per creare la combinazione lineare.

3.2 Funzione per il calcolo dell'indice P_A

In questo paragrafo si riporta la funzione R in grado di calcolare l'indice P_A di un biomarcatore in presenza di tre classi diagnostiche. Prima di procedere il biomarcatore IGF2 viene cambiato di segno (verrà indicato con la sigla - IGF2) perché, al contrario di tutti gli altri biomarcatori, i suoi risultati più alti corrispondono ai pazienti con tumore allo stadio benigno mentre quelli più bassi corrispondono allo stadio maligno avanzato. La prima parte della funzione è un ciclo che ha lo scopo di individuare il numero di individui che compongono i tre gruppi:

```
P_a <- function(x) {
  i = n_1 = n_2 = n_3 = 0
  for(i in 1:length(Gruppo)) {
    if(Gruppo[i]==1) {
      n_1 <- n_1 + 1
    } else {
      if(Gruppo[i]==2) {
        n_2 <- n_2 + 1
      } else {
        n_3 <- n_3 + 1
      }
    }
  }
```

Nelle variabili n_1, n_2 e n_3 è dunque presente il numero di soggetti appartenenti alle tre classi diagnostiche. Nella seconda parte della funzione, si calcola la statistica di Mann–Whitney relativamente ai gruppi 2 e 1 e ai gruppi 3 e 2, ovvero si contano le volte in cui un elemento del gruppo 2 supera un elemento del gruppo 1 e un elemento del gruppo 3 ne supera uno del gruppo 2. Le statistiche di Mann-Whitney serviranno a trovare le stime non parametriche dell'*AUC* fra le due coppie di gruppi:

```
j = k = l = z = t = y = 0
for(j in 1:n_1) {
  for(k in (n_1+1):(n_1+n_2)) {
    if(x[j]<x[k]) {
      t <- t + 1
    }
  }
}
for(l in (n_1+1):(n_1+n_2)) {
  for(z in (n_1+n_2+1):length(Gruppo)) {
    if(x[l]<x[z]) {
      y <- y + 1
    }
  }
}
```

Infine, vengono calcolate le stime non parametriche delle *AUC* dividendo le statistiche di Mann-Whitney per il prodotto delle numerosità dei gruppi confrontati. Come

ultimo passo, viene restituito l'indice P_A, ovvero la media tra le stime non parametriche delle *AUC*:

```
AUC_12 <- t/(n_1*n_2)
  AUC_23 <- y/(n_2*n_3)
  return((AUC_12+AUC_23)/2)
}
```

Questa funzione viene applicata a tutti e 12 i biomarcatori: i risultati si trovano nella tabella 3.3. È possibile confrontare gli indici P_A con i rispettivi *VUS*, calcolati in precedenza; inoltre, a partire da P_A si trova facilmente il limite inferiore del *VUS* dato che:

$$max(\,0, (M-1)P_A - (M-2)\,) \leq VUS \leq P_M.$$

Biomarcatore	**Indice P_A**	**Limite inferiore del *VUS***	***VUS***
CA153	0.654	0.308	0.355
CA125	0.771	0.543	0.566
KLK6	0.679	0.359	0.406
CA724	0.527	0.055	0.397
HE4	0 .752	0.504	0.517
DD.X065	0.584	0.168	0.271
YKL40	0.632	0.264	0.313
DD.0110	0.600	0.200	0.343
DD.C248	0.566	0.132	0.227
DD.P108	0.650	0.300	0.354
- IGF2	0.714	0.428	0.470
SMRP	0.646	0.292	0.352

Tabella 3.3 – Indici P_A e VUS dei biomarcatori

Si nota che il *VUS* rispetta il suo limite inferiore per ogni biomarcatore. Gli indici P_A, invece, sono sempre superiori al *VUS*, dato che si riscontra anche nelle analisi condotte in altri lavori.

3.3 Algoritmo per la combinazione ottimale di biomarcatori

3.3.1 Prima parte: ordinamento dei biomarcatori

Una volta ottenuta una funzione in grado di calcolare l'indice P_A, il metodo non parametrico per la combinazione ottimale di biomarcatori prevede che essi vengano ordinati in ordine decrescente. Viene quindi creato un nuovo dataset (che verrà chiamato `dati`) contenente i risultati dei biomarcatori ordinati secondo l'indice P_A e, come ultima variabile, il gruppo diagnostico a cui appartiene il soggetto. Parte di questo dataset è visibile nella tabella 3.4:

Biomarcatori ordinati secondo l'indice P_A											
CA125	**HE4**	**- IGF2**	**KLK6**	**CA153**	**DD.P108**	**SMRP**	**YKL40**	**DD.0110**	**DD.X065**	**DD.C248**	**CA724**
0,112	-0,612	-0,391	-0,097	0,116	-0,161	-0,555	-0,801	-0,439	0,821	-0,027	1,184
-0,001	1,127	-0,075	-0,133	0,321	0,519	0,829	1,287	-0,407	1,377	-0,027	-0,410
0,278	0,191	0,167	0,587	-0,143	0,944	0,344	0,330	-0,183	0,939	-1,325	-0,410
…	…	…	…	…	…	…	…	…	…	…	…
2,375	0,196	-0,188	-0,226	-0,041	-0,766	-1,348	0,163	-0,368	0,520	-1,146	-0,410
0,168	1,487	0,417	0,028	0,815	1,201	-1,229	0,580	0,326	1,209	0,774	-0,410
2,207	3,567	-0,172	2,093	1,584	1,007	0,999	1,814	2,402	1,087	0,466	-0,410
…	…	…	…	…	…	…	…	…	…	…	…
1,960	3,955	0,278	1,568	-0,036	-0,005	-0,097	0,693	-0,408	0,448	-1,325	1,584
-0,026	1,457	-0,615	0,431	1,020	0,975	0,954	0,391	0,504	0,756	0,273	-0,410
4,399	6,284	0,719	3,553	4,640	3,271	5,821	2,193	3,002	0,697	2,705	5,123

Tabella 3.4 – Parte del dataset con i biomarcatori ordinati secondo l'indice P_A

Su questo dataset si applicherà l'algoritmo per la combinazione lineare ottimale dei biomarcatori.

3.3.2 Seconda parte: combinazioni lineari ricorsive fra biomarcatori

L'algoritmo prevede che vengano combinati in maniera lineare i primi due biomarcatori in modo tale che risulti:

$$V_1 = CA125 + \alpha_1 \cdot HE4 \,.$$

Giunti a questo punto si usa P_A come funzione obiettivo, ovvero si cerca un α_1^* che massimizzi l'indice P_A calcolato sulla combinazione V_1. La ricorsività dell'algoritmo consiste nel sostituire CA125 con $V_1(\alpha_1^*)$ e HE4 con il biomarcatore successivo, ovvero – IGF2, ottenendo la nuova combinazione lineare:

$$V_2 = V(\alpha_1^*) + \alpha_2 \cdot (-IGF2) \,.$$

Si prosegue di questo passo fino all'esaurimento dei biomarcatori (dunque i cicli dell'algoritmo saranno 11). La combinazione finale sarà quella ottimale e avrà un'accuratezza diagnostica migliore di ogni biomarcatore contenuto nel dataset.

L'implementazione di questo algoritmo si basa su tre elementi:

1. La funzione R "`optim ()`", che permette di trovare il minimo di una funzione obiettivo. In questo caso, la funzione obiettivo verrà cambiata di segno così da ottenere il massimo.
2. Un ciclo di tipo `for`, che permette all'algoritmo di ripetersi per 11 volte.
3. Una matrice, detta `matrice_combinazioni`, in cui viene inserito il primo elemento della combinazione lineare. La prima colonna sarà dunque uguale a quella del biomarcatore CA125 e le successive, inizialmente vuote, verranno riempite con la combinazione lineare ottimale che si ottiene combinando progressivamente i biomarcatori.

Alla fine dell'esecuzione dell'algoritmo, la combinazione ottimale di biomarcatori si troverà nell'ultima colonna della matrice, quindi su di essa verrà calcolato l'indice P_A per verificare che la sua accuratezza diagnostica sia più alta di ogni biomarcatore. Come già riportato in precedenza, si ha che $\alpha \in [-1, 1]$, quindi la ricerca del termine α_1^* viene ristretta attraverso i comandi aggiuntivi "lower" e "upper". Il codice R per implementare l'algoritmo è il seguente:

```
matrice_combinazioni <- matrix(nrow= 279, ncol=12)
matrice_combinazioni[,1] <- dati[,1]
combinazione_ottimale <- c(1:279)
q <- 0
system.time(
for(q in 1:11) {
  V <- function (alpha) {
    return(P_a(matrice_combinazioni[,q]+alpha*dati[,q+1]))
    }
  alpha_max <- optim(par = 0, fn = V, lower = -1, upper = 1,
                     method = "Brent",
                     control=list(fnscale=-1))$par
  matrice_combinazioni[,q+1] <- (matrice_combinazioni[,q]
                                 + alpha_max*dati[,q+1])
  print(alpha_max)
})
combinazione_ottimale <- matrice_combinazioni[,q+1]
```

Il metodo di ottimizzazione scelto è l'algoritmo di Brent, che si basa su una combinazione del metodo di bisezione e il metodo di interpolazione parabolica.

3.4 Risultati dell'algoritmo

La combinazione ottimale è contenuta nel vettore `combinazione_ottimale`. Ad ogni interazione, il ciclo è stato programmato per stampare a schermo il coefficiente ottimo α_i^* che massimizza la combinazione lineare V_i. Inoltre, è possibile calcolare gli indici P_A delle varie combinazioni lineari. Tutti i risultati ottenuti sono esposti nella tabella 3.5:

Passo	**Biomarcatori combinati**	**Coefficiente α_i^***	**Indice P_A**
1	$V_1 = CA125 + \alpha_1 \cdot HE4$	0.818	0.804
2	$V_2 = V_1 + \alpha_2 \cdot (-IGF2)$	0.927	0.830
3	$V_3 = V_2 + \alpha_3 \cdot KLK6$	0.611	0.836
4	$V_4 = V_3 + \alpha_4 \cdot CA153$	0.163	0.837
5	$V_5 = V_4 + \alpha_5 \cdot DD.P108$	- 0.153	0.837
6	$V_6 = V_5 + \alpha_6 \cdot SMRP$	0.003	0.837
7	$V_7 = V_6 + \alpha_7 \cdot YKL40$	- 0.292	0.838
8	$V_8 = V_7 + \alpha_8 \cdot DD.0110$	- 0.203	0.838
9	$V_9 = V_8 + \alpha_9 \cdot DD.X065$	- 0.167	0.839
10	$V_{10} = V_9 + \alpha_{10} \cdot DD.C248$	- 0.187	0.840
11	$V_{11} = V_{10} + \alpha_{11} \cdot CA724$	0.132	**0.840**

Tabella 3.5 – Risultati dell'algoritmo di combinazione di biomarcatori

Si nota immediatamente che l'accuratezza diagnostica, misurata attraverso l'indice P_A, cresce al crescere del numero di biomarcatori combinati fino alla combinazione finale (P_A = 0.840). Quest'ultimo indice è superiore a quelli calcolati sui biomarcatori non combinati riportati nella tabella 3.3. Analogamente, anche il *VUS* calcolato sulla combinazione ottima di biomarcatori risulta anch'esso essere maggiore dei *VUS* calcolati su gli altri biomarcatori (*VUS* = 0.689).

Ulteriori risultati sull'efficacia dell'algoritmo possono essere tratti osservando il suo tempo di esecuzione, calcolato attraverso la funzione "`system.time ()`". Il tempo impiegato è di circa 7 secondi, dunque il processo non è molto dispendioso dal punto di vista computazionale. Per citare altri esempi, in presenza di 5 classi diagnostiche e di una numerosità pari a 1000, questo metodo di combinazione di biomarcatori ha un tempo di esecuzione nell'ordine dei minuti, mentre sugli stessi dati il metodo "classico" di combinazione basato sul massimo *HUM* impiega circa tre settimane (M. J. Hsu e Y. H. Chen, 2014).

Bibliografia e sitografia

- Hsu M. J. e Chen Y. H. *Optimal linear combination of biomarkers for multi-category diagnosis*. "Statistics in Medicine" 35 (2016).
- Kang L., Aiyi L. e Tian L. *Linear combination methods to improve diagnostic/prognostic accuracy on future observations*. "Statistical Methods in Medical Research" (2014).
- Chunling L., Aiyi L. e Halabi S. *A min-max combination of biomarkers to improve diagnostic accuracy*. "Statistic in Medicine" 35 (2011).
- Su J. Q. e Liu J. S. *Linear combinations of multiple diagnostic markers*. "Journal of the American Statistical Association" 88 (1993).
- He X. e Frey E. C. *The meaning and use of the volume under a three-class ROC surface (VUS)*. "Transactions on Medical Imaging" 27 (2008).
- Hanley J. A. e McNeil B. J. *A method of comparing the areas under receiver operating characteristic curves derived from the same cases*. "Radiology" 148 (1983).
- AA. VV. *Use of biomarkers in predicting healt and mortality*. "Today's research on aging" 14 (2008).
- Strimbu K. e Tavel J. A. *What are biomarkers?*. "Current opinion in HIV and AIDS" 5 (2010).
- Piccolo D. (2004). *Statistica*. Il Mulino, Milano.

- Azzalini A. (2001). *Inferenza statistica*. Springer, Milano.
- Dispense didattiche di Statistica Medica, Università di Padova, A.A. 2015-2016. A cura della Prof.ssa Laura Ventura.
- Dispense didattiche di Statistica Computazionale, Università di Padova, A.A. 2015-2016. A cura del Prof. Matteo Grigoletto.
- <*www.news-medical.net/health/What-is-a-Biomarker.aspx*>. Consultato in Luglio 2016.
- <*marginalrevolution.com/marginalrevolution/2015/09/the-frechet-probability-bounds-super-wonky.html*>. Consultato in Luglio 2016.
- <*www.researchgate.net/publication/261178836_R_Algorithms_to_Calculate_VUS_and_Plot_ROC_Surface_An_Application_from_Three_Ordered_CategorieC_of_Eating_Disorders*>. Consultato in Agosto 2016.
- <*www.magesblog.com/2013/03/how-to-use-optim-in-r.html*>. Consultato in Agosto 2016.