

“

JNOTES

CREATIVE NOTES

>

DECISIONI BAYESIANE e MODEL EVALUGTION

LIMITI DELL'APPROCCIO MAXIMUM-A-POSTERIORI

FINORA abbiamo assunto che le stime tgg ai campioni basandosi sulle probabilità a posteriori massime. Questo approccio tuttavia non considera che diversi errori di classificazione possono avere un impatto differente.

Quando assumiamo una classe basandoci sulla probabilità più alta (ma) assumiamo implicitamente che tutti gli errori abbiano lo stesso costo. Nella realtà, le conseguenze di classificazioni errate, possono variare drasticamente.

NEL USO DI UN SISTEMA DIAGNOSTICO MEDICO:

O SE CLASSIFICHIAMO ERRORANTE UN PATIENTE MALATO (FALSO POSITIVO), QUESTO PUÒ ESSERE CATASTROFICO!

- ULTRASONI ESAMI NON NECESSARI
- STRESS PSICOLOGICO PER IL PATIENTE
- COSTI SANITARI AGLIUVANTI (MANENTI PERICOLOSI)

O SE CLASSIFICHIAMO ERRORANTE UN PATIENTE SANO (FALSO NEGATIVO) LE CONSEGUENZE PUÒ ESSERE:

- RITARDO NELLA DIAGNOSI E NEL TRATTAMENTO
- PROGRESSIONE DELLA MALATTIA SENZA INTERVENTI
- POSSIBILI DANNI PERMANENTI O ESITI FATALI

COME QUANTIFICARE GIÒ NELLA PIA DECISIONE FINALE?

SONO STATE DA NOI INTRODOTTE NEMIHE PER LE DECISIONI

$$\text{ACCURACY} = \frac{\# \text{ DI CAMPIONI CLASSIFICATI CORRETTAMENTE}}{\# \text{ TOTALE DI CAMPIONI}}$$

$$\text{ERROR RATE} = \frac{\# \text{ DI CAMPIONI CLASSIFICATI ERRORANTE}}{\# \text{ TOTALE DI CAMPIONI}} = 1 - \text{ACCURACY}$$

Perché l'accuracy non è un ottima metria?

- ① Non considera il costo dei diversi tipi di errore: se uno falso positivo, l'accuracy tutta tutti gli errori allo stesso modo ma alcuni errori potrebbero essere molto più costosi o pericolosi di altri

② **DIPENDENZA DISTRIBUZIONE CLASSE NEL DATASET:** L'ACCURATEZZA È INPIÙ
DIPENDENTE DALLA PROPORTIONE DELLE DIVERSE CLASSI NEL DATASET, CHE POSSANO ESSERE NON RIFLETTE LA
DISTRIBUZIONE DELL'APPPLICAZIONE REALE. (ALGORITMO: $P(C=c | X=x) = \frac{P(x|C=c) P(c)}{\sum_i P(x|C=i) P(i)}$, DIPENDE
DALLA PRIOR) SE NEL NOSTRO DATASET IL 95% DEI CAPIONI APPARTIENE ALLA CLASSE A,
UN CLASSIFICATORE CASUALE CHE ASSIGNA SEMPRE UNA CLASSE A OTTERREBBE UN ACCURATEZZA DEL 95%
APPARENDONE Evidentemente EFFICACE.

③ **Non è Normalizzata:** L'ACCURATEZZA DA SOLO NON PERMETTE DI GIUDicare SE UN
CLASSIFICATORE È EFFETTIVAMENTE "UTILE" (MIGLIORE OFL USO) o MIGLIORE DELLE DECISIONI
PRESE SENZA UNA CLASSIFICAZIONE.

Noi vorremo che il nostro sistema fornisse, prestazioni ottimali sui dati dell'applicazione. Tuttavia, non possano misurare a priori le prestazioni di un sistema sui dati dell'applicazione, poiché tipicamente non abbiamo né i dati né le etichette corrispondenti

Possano d'altra parte, utilizzare un set di valutazione per calcolare le prestazioni
di diversi modelli. Le metriche di valutazione hanno due scopi:

- ① MISURARE CORRISPONDENTEMENTE LE NOSTRE PRESTAZIONI SUL DATASET DATO
- ② OTTENERE UNA STIMA DI TUE NOSTRE PRESTAZIONI SU DATI DI APPLICAZIONE CHE SI
CORRISPONDANO AL NOSTRO SET DI VALUTAZIONE.

Per il secondo punto è importante che le nostre metriche non dipendano
dalla distribuzione empirica del set di valutazione, quando questa può differire dalla
distribuzione dell'applicazione reale.

Se stiamo addestrando un sistema per rilevare frodi bancarie usando un set di dati storici dove
le frodi rappresentano l'1% delle transazioni, ma nell'applicazione reale attuale rappresentano il
3% (a causa di nuove tecniche di frode), dobbiamo utilizzare metriche che non siano influenzate
da questa differenza di distribuzione.

MATRICE DI CONFUSIONE

La matrice di confusione, per ogni combinazione di etichette di classe C_i e etichetta predetta
 C_j : il numero di campioni C_i predetti come C_j :

- GLI ELEMENTI SULLA DIAGONALE rappresentano i campioni classificati correttamente
- GLI ELEMENTI FUORI DALLA DIAGONALE rappresentano i campioni classificati erroneamente.

	Class C_1	Class C_2	...	Class C_K
Prediction C_1	# samples of C_1 predicted as C_1	# samples of C_2 predicted as C_1		# samples of C_K predicted as C_1
Prediction C_2	# samples of C_1 predicted as C_2	# samples of C_2 predicted as C_2		# samples of C_K predicted as C_2
⋮	⋮	⋮	⋮	⋮
Prediction C_K	# samples of C_1 predicted as C_K	# samples of C_2 predicted as C_K		# samples of C_K predicted as C_K

Per capire nello quale l'accuracy fa male, consideriamo un problema di classificazione binario, in questo caso la matrice di confusione:

	Class H_F	Class H_T
Prediction H_F	True Negative	False Negative
Prediction H_T	False Positive	True Positive

Dove:

- True Negative: (n) Campioni della classe H_F correttamente classificati come H_F
- True Positive: (p) Campioni della classe H_T correttamente classificati come H_T
- False Negative: (pn) Campioni della classe H_T erroneamente classificati come H_F
- False Positive: (pp) Campioni della classe H_F erroneamente classificati come H_T

Accuracy e Error Rate:

$$\text{Accuracy} = (Tn + Tp) / (Tn + Fp + Tp + Fn)$$

$$\text{Error Rate} = (Fp + Fn) / (Tn + Fp + Tp + Fn) = 1 - \text{Accuracy}$$

Consideriamo l'esempio delle previsioni della pioggia in cui andò, su un anno, il modello effettua le seguenti previsioni:

Rain Clear

Prevision - Rainy 15 giorni 30 giorni

Prevision : Clear 20 giorni 300 giorni

L'accuracy è: $\frac{300+15}{365} \approx 86\%$

Anche se l'accuracy è 86%, il modello presenta problemi significativi:

- Alta proporzione di falsi positivi: 30 giorni su 35 previsioni di pioggia sono errate (66,7% di FP)
- Alta proporzione di falsi negativi: 20 giorni di pioggia su 35 reali non sono stati previsti (55,6% di FN)
- Scarsa capacità preveditriva per la classe minoritaria: in un cumulo di 35 giorni, prevedere correttamente la pioggia è più importante che prevedere giorni secchi, ma il modello è inefficace proprio in questo campo.

L'accuracy allora può diviso rispetto alle prestazioni su classi specifiche soprattutto quando queste sono solanzate

Let's now consider a diagnostic task, where we need to classify a medical exam as "positive" or "negative"

To assess the effectiveness of a classifier, we employ a dataset that contains both actual positive and negative samples (sample label)

To have a good coverage of both cases, we collect the results for few hundreds of patients of each kind. Let's assume we have 1000 positive and 1000 negative samples, and that two classifiers \mathcal{R}_1 and \mathcal{R}_2 have the following confusion matrices for the evaluation set:

	Pos.	Neg.
Pred. Pos.	940	20
Pred. Neg.	60	980

	Pos.	Neg.
Pred. Pos.	980	40
Pred. Neg.	20	960

The second recognizer has an overall larger accuracy, 97% vs. 96%

Tuttavia le false rate misurano il numero di errori per il dataset, ignorando il fatto che la nostra applicazione realizza, purtroppo avere una distribuzione a priori di pazienti positivi diversa dalla distribuzione fornita del set di valutazione (cioè la proporzione di pazienti positivi sul campione potrebbe differire dalla proporzione di pazienti positivi nel set di valutazione)

Io quando addeo il mio classificatore faccio comunque il passo davanti su un dataset con la sua priori, se mi più dati di una classe, colico pensare che il profilo sia più vicino a questa classe che a quelle che appartengono a quella classe. A sua volta quando mi calcolo la precisione, non faccio per l'100% del dataset di valutazione sto comunque valutando il tutto conoscendo la distribuzione del set di valutazione ipoteticamente simile a quella di addestramento. A livello applicativo la distribuzione può cambiare completamente però.

Metri di errore per classe

① False negative rate (FNR) : (Tasso di falso negativo / Tasso di mancata elevazione)

$$P_{FN} = \frac{FN}{FN+TP}, \text{ errore FNR in classe positiva.}$$

② False Positive rate (FPR) (tasso di falso accettazione)

$$P_{FP} = \frac{FP}{FP+TN}, \text{ errore FPR in classe negativa}$$

③ True Positive Rate (TPR) (richiamo, sensibilità)

$$TPR = \frac{TP}{TP+FN} = 1 - FNR, \text{ rappresenta la proporzione di campioni positivi correttamente identificati.}$$

④ True Negative Rate (TNR) (specificità)

$$TNR = \frac{TN}{TP+TN} = 1 - FPR$$

NAPOLI PER LA PROPORTIONE DI CAMPIONI NERATIVI CORRISPONDENTI AI TESTIFICATI

QUESTE METRICHE DIPENDONO DAI CAMPIONI DI UNA SINGOLA CLASSE, NON SONO INFLUENZATE DALLA PROPORTIONE DI CAMPIONI DI OLTRE CLASSE.

Empirical Prior

L'EMPIRICAL PRIOR DEL EVALUATION SET È DISTRIBUZIONE EMPIRICA. SI RIFERISCE ALLA PROPORTIONE EFFETTIVA DELLE DIVERSE CLASSI, PRESENTI ALL'INTERNO DEL DATASET CHE STIAMO UTILIZZANDO. (SIMILE ALA PRIOR PROBABILITA' MA NON È UNA CONOSCENZA NERA, PENSARE È SICURAMENTE).

NEL CASO DI UN PROBLEMA DI CLASSIFICAZIONE DINAMICA LA DISTRIBUZIONE EMPIRICA DEL SET DI VALUTAZIONE ($T_{t^{emp}}$) MAPPARESINA LA PROPORTIONE DI CAMPIONI APPARTENENTI ALLA CLASSE POSITIVA, RISPETTO AL TOTALE DEI CAMPIONI.

$$T_{t^{emp}} = \frac{TP+FN}{TN+FP+TP+FN}$$

QUESTA MAPPARESINA TUTTI I PROBLEMI DELL'ALGORITMO/ERRORE.

$$\text{err-rate} = \text{err} = \frac{\# \text{ di errori}}{\# \text{ di campioni}} = \frac{FP + FN}{TN + FP + TP + FN} = \frac{PP}{TN + PP + TP + FN} + \frac{FN}{TN + PP + TP + FN} \geq \\ = \frac{FP}{TN + PP} \cdot \frac{TN + PP}{TN + PP + TP + FN} + \frac{FN}{TP + FN} \cdot \frac{TP + FN}{TN + FP + TP + FN} = P_{FP} (1 - T_{t^{emp}}) + P_{FN} T_{t^{emp}}$$

IL TASSO DI ERRORE È UNA MEDIA PUNZONATA DI TASSI DI ERRORE SPECIFICI PER CLASSE, DOVE I PESI SONO DETERMINATI DALL'EMPIRICAL PRIOR (CHE CIARNO DAL' EVALUATION SET E CABA PRESENTELLATE NEL CONTESTO APPLICATIVO).

Distribuzione Empirica vs Distribuzione Appuntiva

Consideriamo uno scenario più realistico per il nostro classificatore di ninostilo (di cui acciamo anche un minimo). Immaginiamo che venga indetto in un contesto dove la probabilità A priori di un risultato positivo è molto bassa, ad esempio l'1%.

QUESTO SCENARIO È REALISTICO, PRATICAMENTE LE MALATTIE HANNO UNA PREVALENZA RELATIVAMENTE BASSA NELL'POPOLAZIONE GENERALE. SE VOLIAMO CHE IL NOSTRO SET DI VALUTAZIONE RISPECCHIASSE ESATTEMENTE QUESTA DISTRIBUZIONE APPUNTIVA ($T_{t^{app}}=0,01$) DOVETRÀ RACCOLGERE 99 CAMPIONI NEGATIVI E 1 POSITIVI.

SE VOLIAMO ALMENO 1000 CAMPIONI PER LA CLASSE POSITIVA (PER MANUTENERE UNA STIMA AFFIDABILE PER LE PRESTAZIONI SU OGNI CLASSE) AVREMOSO DISOLDO DI RACCOLGERE BEN 95000 CAMPIONI PER LA CLASSE NEGATIVA.

FORTUNATAMENTE, SE ASSUMIAMO CHE I TASSI DI ERRORE PER LA CLASSE AVANZOSSINO APPROSSIMATIVAMENTE GLI STESSI DI QUELLI MISURATI NEL NOSTRO SET DI VALUTAZIONE DINAMICO ($T_{t^{app}}=0,5$), POSSIAMO SINTETIZZARE IL TASSO DI ERRORE CHE AVREMO CON UNA DISTRIBUZIONE APPUNTIVA ($T_{app}=0,01$) SENZA DOVER EFFETTUARE NIENTE RACCOLGIRE L'ERRORE QUANTITATIVO DI CAMPIONI NELL'AVV.

La forma generale di una STEM (riflette quella del set di valutazione)

$\text{err}_{app} = P_{R_1}(l-\bar{l}) + P_{R_2}(\bar{l})$, $\bar{l}=90$ in questo caso, probabilità a priori, o una classe positiva.

$$P_{R_1}(R_1) = \frac{70}{20+90} = 0.07$$

Consideriamo i 2 classificatori R_1 e R_2 con le stesse probabilità a priori

$$P_{R_1}(R_1) = \frac{70}{20+90} = 0.07, \quad P_{R_1}(R_2) = \frac{60}{60+90} = 0.06$$

$$P_{R_2}(R_1) = \frac{60}{60+90} = 0.06, \quad P_{R_2}(R_2) = \frac{20}{20+90} = 0.07$$

In distribuzione empirica nel set di valutazione ottimale FNG $P_{\hat{f}}^{emp} = \frac{1000}{2000} = 0.5$, da cui si rilevano i tassi di errore otimali $err(R_1) = 0.03$, $err(R_2) = 0.03$

Se cambiamo la priori empirica con $\bar{l} = 0.02$, con gli stessi errori per classe, avremo $err_{app}(R_1) = 0.0204$, $err_{app}(R_2) = 0.0333$

Quindi in un contesto applicativo con questo classificatore R_1 sarebbe meglio di R_2 . Questo dimostra chiaramente come la distribuzione delle classi influenza la valutazione delle prestazioni di classificazione.

COSTO

Infine, il tasso di errore non tiene conto del costo dei diversi tipi di errori. Per il compito diagnostico, abbiamo visto che, se la probabilità a priori di essere positivo è bassa, il primo classificatore (R1) comporta un numero totale di errori inferiore. Tuttavia, etichetta erroneamente come negativi più pazienti positivi rispetto al secondo sistema.

Possiamo immaginare che etichettare erroneamente come negativo un paziente positivo sia più pericoloso del contrario, cioè possa avere un costo maggiore.

Il concetto di costo non è limitato a un costo monetario, ma è una generalizzazione che dovrebbe quantificare gli effetti di una classificazione errata, e in particolare riflettere gli effetti relativi dei diversi tipi di errore. Tipicamente, i costi cambieranno da applicazione ad applicazione.

Per esempio, in un contesto medico:

Un falso negativo potrebbe significare che un paziente malato non riceve le cure necessarie, con potenziali conseguenze gravi per la salute.

Un falso positivo potrebbe comportare esami aggiuntivi o trattamenti non necessari, con costi economici e possibili effetti collaterali, ma generalmente con conseguenze meno gravi.

In altre parole, anche se R1 ha un tasso di errore complessivo inferiore con la distribuzione applicativa, R2 potrebbe comunque essere preferibile se il costo di un falso negativo è significativamente maggiore di quello di un falso positivo, poiché R2 ha un tasso di falsi negativi più basso (0.02 vs 0.06).

Questa considerazione evidenzia l'importanza di andare oltre le semplici metriche di accuratezza o tasso di errore quando si valutano i classificatori per applicazioni reali, tenendo conto sia della distribuzione delle classi sia del costo relativo dei diversi tipi di errori.

Menù Coperni

Abbiamo bisogno di una menù che superi tutti i limiti discorsi in precedenza e che ci permetta di riassumere le prestazioni di un classificatore con un solo numero che:

- ① DIPENDA DALLA DISTRIBUZIONE APPPLICATIVA A PALE, PIUTTOSTO CHE DALLA DISTRIBUZIONE EMPIRICA DEL SET DI VALUTAZIONE.
- ② TENGÀ CONTO DEI DIVERSI COSTI DEI ERROI DI CLASSIFICAZIONE.
- ③ CONSPERA DI CONFRONTARE CLASSIFICATORI DIVERSI.
- ④ SIA NORMALIZZATA, PERMETTENDO DI VALUTARE SE UN CLASSIFICATORE È EFFETTIVAMENTE INFORMATIVI UTILI DAGLI OBIETTIVI.

QUESTA RETEVA È IL RISCHIO BAYESIANO.

Bayes Risk

L'OGOLOTTIVO PONERMENTALE DI UN PROBLEMA DI CLASSIFICAZIONE (DI UN CLASSIFICATORE) È CONSENTIRE DI SCEGLIERE UN'AZIONE (= DECISIONE) ADEGUATA TRA UN INSERTE DI POSSIBILI AZIONI:
Per esempio, nostro caso)

- ASSIGNARE UN'ETICHETTA SPECIFICA AL CAMPIONE.

A CLASSELLA AZIONE POSSIAMO ASSOCIARE UN COSTO $C(x_i, h)$ CHE DOBBIANO "PALARE" QUANDO SCEGLIEMO L'AZIONE h ED IL CAMPIONE APPARTENEVA ALLA CLASSE x_i . [COSTO DI DELLA CLASSE (CONFERMO CHE SCELTO È NO - GENERE)]

SEBBENÈ NON ABBIANO OFFERTO CON LE OTTENERE DECISIONI OTTIMALI, ASSURDO CHE UN CLASSIFICATORE SIATE IN GRADO DI PRODURRE UNA DECISIONE (ETICHETTATURA) PER OGNI CAMPIONE SECONDO UNA CERTA REGOLA.

Denotiamo con $\hat{a}(x_i, h)$ la decisione presa dal classificatore R per il campione x_i , la cui etichetta di classe corrente è c , il costo di tale decisione è
$$(\hat{a}(x_i, h)/c).$$

AZIONE DI VALUTAZIONE
DI UNO ERRORE

Corrisponde al costo di errore l'etichetta $\hat{a}(x_i, h)$ quando la classe corrente è c .

Possiamo ESAMINARE IL COSTO ATTESO (Bayes Risk) DELLE DECISIONI PRESE DAL NOSTRO CLASSIFICATORE SULLA POPOLAZIONE MATERIALE (APPPLICAZIONE),cioè il costo che ci ASPETTIANO DI PALARE UTILIZZANDO IL NOSTRO SISTEMA SULLA POPOLAZIONE DELL'APPPLICAZIONE.

$$B = E \left[\sum_{x_i \in E} \sum_{h \in H} (\hat{a}(x_i, h)/c) \right]$$

Dove:

- E è l'evaluation o environment. rappresenta il contesto dell'applicazione materiale in cui il classificatore viene utilizzato.

Oltre ad B dico $\frac{1}{|E|}$ significa considerare l'insieme completo di tutti i possibili campioni che

IL NOSTRO SISTEMA CLASSIFICA I NEI RISULTATI, NON SOLO QUelli PRESENTI NEL SET DI VALUTAZIONE.

- $X, C/\varepsilon$ È UNA DISTRIBUZIONE CHE INDICA COME SONO DISTRIBUITI I COSTI NELLA POPOLAZIONE DI APPUNTAMENTO.

$X, C/\varepsilon$ DISTRIBUE CON LE CARATTERISTICHE: I LF (C/I) SONO DISTRIBUITI INSIEME NEL MEDIO RISULTATO DELL'APPUNTAMENTO.

$$B = E_{X, C/\varepsilon} [C(a(x, \eta))|c] = \int_c \int_{x, \eta} C(a(x, \eta))|c f_{X, C/\varepsilon}(x, c) dx dc \stackrel{\text{C è costante}}{=} \\ = \sum_{c=1}^n \int_{x, \eta} C(a(x, \eta))|c f_{X, C/\varepsilon}(x, c) dx$$

PURTROPPO NON POSSIAMO CONOSCERE DELL'A DISTRIBUTIONE CONCERNA $X, C/\varepsilon$. Tuttavia se assumiamo di conoscere le probabilità a milioni dell'appuntamento $\pi_c = P(C=c)$, possiamo ESPRIMERE IL RISCHIO BATTENDO PER L'APPUNTAMENTO TABELLE COME

$$B = \sum_{c=1}^n \pi_c \int f_{X|C=c}(x|c) C(a(x, \eta))|c dx = \sum_{c=1}^n \pi_c E_{X|C=c} [C(a(x, \eta))|c]$$

STIMO SONO NOI I COSTI PER OGNI POSSIBILE COMBINAZIONE DI CARATTERISTICHE C/I, PRATICI SECONDO LA LORO PROBABILITÀ DI OCCORRENZA NELL'AMBIENTE DI APPUNTAMENTO.

DATO CHE PENSAMO NON POSSIAMO ACCEDERE ALLA DISTRIBUZIONE $X|C, \varepsilon$ E ANCHE A $f_{X|C, \varepsilon}(x|c)$ NON POSSIAMO CALCOLARE OMELTAMENTE IL RISCHIO BATTENDO (CONTENUTO APPUNTAMENTO). TUTTAVIA SE DISPOSIAMO DI UN INSERTE DI CAMPIONI DI VALUTAZIONE ESTREMAMENTE $(x_1, c_1), \dots, (x_n, c_n)$ POSSIAMO APPROSSIMARE E (L'INSERTEVITA) CALCOLANDO LA MEDIA DEI COSTI SUI CAMPIONI.

SE OLTRE I CAMPIONI x_i SONO GENERATI DA $X|C, \varepsilon$ DANNO VANO CHE IL NUMERO DI CAMPIONI ALCUNSE DIVENTA GRANDE ACCORDO CHE

$$E_{X|C, \varepsilon} [C(a(x, \eta))|c] = \int C(a(x, \eta))|c f_{X|C, \varepsilon}(x|c) dx = \\ = \frac{1}{N_c} \sum_{i=1}^{N_c} C(a(x_i, \eta))|c$$

APPROSSIMANDO ALLA FINE DEI COSTI CALCOLATI SUI CAMPIONI DI CAMPIONI CHIUSI DEL SET DI VALUTAZIONE.

Gramm Crises Risk

$$\text{GCR} = \sum_{c=1}^n \frac{1}{N_c} \sum_{i=1}^{N_c} C(a(x_i, \eta))|c$$

IL RISCHIO GRISSA I COSTI DELLE NOSTRE DECISIONI PER UN APPUNTAMENTO TABLATI SUL CAMPIONE DI VALUTAZIONE. POSSIAMO USARE GCR PER CONFRONTO I CLASSIFICATORI: COSTI INFERIORI COMPARANDO A PRESTAZIONI MILLARI.

SE IMPONIAMO $\pi_c = \pi_c^*$, IL GRAMM RISK ENTRICO CORISPONDE AI COSTI TOTALI DI MISCLASSIFICATION PER I

CAMPIONI DI VALUTAZIONE (SIMILI AI TASSI DI ERRORE IN IMPRESA COMO DIT COSTI).

Il rischio bayesiano empirico può essere calcolato direttamente utilizzando la matrice di confusione e la matrice dei costi. Questo è particolarmente utile nei problemi con più di due classi, dove la notazione diventa più complessa ma il principio rimane lo stesso.

Esempio di un Problema a 3 Classi:

Consideriamo un problema con 3 classi, dove abbiamo:

Matrice dei costi C:	$\begin{matrix} 1 & 203 & 119 & 56 \\ 2 & 865 & 93 & 92 \\ 3 & 50 & 92 & 225 \\ \hline L & 2 & 3 & \end{matrix}$
$\Pi = \text{Prob}(C FO)$	$\begin{matrix} 1 & 0.3 & 0.4 & 0.3 \\ 2 & 0.4 & 0.3 & 0.3 \\ 3 & 0.3 & 0.2 & 0.5 \end{matrix}$
Matrice di confusione M:	$\begin{matrix} 1 & 145 & 199 & 121 \\ 2 & 50 & 92 & 225 \\ 3 & 11 & 2 & 3 \end{matrix}$

Questa matrice indica quanto "costa" ogni tipo di classificazione. Ad esempio: $C(1,3) = 2$ significa che classificare un campione di classe 1 come classe 3 ha un costo di 2. Notiamo che la diagonale è zero, il che significa che le classificazioni corrette non hanno costo (cosa che possiamo assumere senza perdita di generalità).

Probabilità a priori: Π_i

$$\Pi = [0.3 \ 0.4 \ 0.3]$$

Questo vettore rappresenta le probabilità a priori delle tre classi nell'ambiente dell'applicazione reale: 30% per la classe 1, 40% per la classe 2 e 30% per la classe 3.

Matrice di confusione M:

$$M = [205 \ 111 \ 56]$$

$$\begin{bmatrix} 145 & 199 & 121 \\ 50 & 92 & 225 \\ 11 & 2 & 3 \end{bmatrix}$$

Questa matrice mostra quanti campioni di ciascuna classe vera (colonne) sono stati classificati in ciascuna classe predetta (righe). Ad esempio: 145 campioni di classe 1 sono stati classificati come classe 2.

Calcolo del Rischio Bayesiano Empirico:

Ricordiamo la formula del rischio bayesiano empirico:

$$\text{Bemp} = \sum_{i=1}^3 \sum_{j=1}^3 \Pi_i \cdot \frac{\text{C}(i,j)}{N_c} \cdot \sum_{k=1}^3 \text{C}(i,k) \cdot \Pi_k$$

Per calcolare questo valore utilizzando la matrice di confusione, dobbiamo procedere classe per classe.

Calcolo per la Classe 1:

Probabilità a priori: $\Pi_1 = 0.3$

Numero totale di campioni della classe 1: $N_1 = 205 + 145 + 50 = 400$

Distribuzione delle predizioni per campioni della classe 1:

205 campioni classificati correttamente (costo 0)

145 campioni classificati come classe 2 (costo 1)

50 campioni classificati come classe 3 (costo 2)

Il contributo della classe 1 al rischio è:

$$\Pi_1 \cdot (1/N_c) \cdot \sum_{i=1}^3 \sum_{j=1}^3 \text{C}(i,j) \cdot \Pi_j = 0.3 \cdot (1/400) \cdot (0 \times 205 + 1 \times 145 + 2 \times 50) = 0.3 \cdot (1/400) \cdot 245 = 0.18375$$

Calcolo per la Classe 2:

Probabilità a priori: $\Pi_2 = 0.4$

Numero totale di campioni della classe 2: $N_2 = 111 + 199 + 92 = 402$

Distribuzione delle predizioni per campioni della classe 2:

111 campioni classificati come classe 1 (costo 1)

199 campioni classificati correttamente (costo 0)

92 campioni classificati come classe 3 (costo 2)

Il contributo della classe 2 al rischio è:

$$\Pi_2 \cdot (1/N_c) \cdot \sum_{i=1}^3 \sum_{j=1}^3 \text{C}(i,j) \cdot \Pi_j = 0.4 \cdot (1/402) \cdot (0 \times 111 + 0 \times 199 + 1 \times 92) = 0.4 \cdot (1/402) \cdot 92 = 0.20199$$

Calcolo per la Classe 3:

Probabilità a priori: $\Pi_3 = 0.3$

Numero totale di campioni della classe 3: $N_3 = 56 + 121 + 225 = 402$

Distribuzione delle predizioni per campioni della classe 3:

56 campioni classificati come classe 1 (costo 2)

121 campioni classificati come classe 2 (costo 1)

225 campioni classificati correttamente (costo 0)

Il contributo della classe 3 al rischio è:

$$\Pi_3 \cdot (1/N_c) \cdot \sum_{i=1}^3 \sum_{j=1}^3 \text{C}(i,j) \cdot \Pi_j = 0.3 \cdot (1/402) \cdot (2 \times 56 + 1 \times 121 + 0 \times 225) = 0.3 \cdot (1/402) \cdot 233 \approx 0.17388$$

Rischio Bayesiano Empirico Totale:

Sommando i contributi di tutte le classi:

$$\text{Bemp} = 0.18375 + 0.20199 + 0.17388 = 0.55962$$

Interpretazione del Risultato:

Il valore 0.55962 rappresenta il costo medio atteso per campione quando utilizziamo questo classificatore nell'ambiente dell'applicazione reale. È un singolo numero che riassume le prestazioni del classificatore, tenendo conto sia dei diversi tipi di errori (attraverso la matrice di costi) sia della distribuzione reale delle classi (attraverso le probabilità a priori).

Questo approccio ha diversi vantaggi:

Incorpora le probabilità a priori dell'applicazione reale.

Considera i diversi costi associati ai diversi tipi di errori.

Fornisce una metrica interpretabile e confrontabile.

Si può calcolare facilmente dalle matrici di confusione e dei costi.

Bates Risk Binary Problem

CONFORMA UN PROBLEMA DI CLASSIFICAZIONE BINARIA

IN QUESTO HANNO 3 COSTI PER OGNI TIPO DI CLASSIFICAZIONE CHE POSSONO AVERE:
 TP, TN, FP e FN

	Class \mathcal{H}_F	Class \mathcal{H}_T
Prediction \mathcal{H}_F	$C(\mathcal{H}_F \mathcal{H}_F)$	$C(\mathcal{H}_F \mathcal{H}_T)$
Prediction \mathcal{H}_T	$C(\mathcal{H}_T \mathcal{H}_F)$	$C(\mathcal{H}_T \mathcal{H}_T)$

H_F/H_T IL NUMERO È UNO ENTRAMBI APPARTENGONO ALLA CLASSE
 FALSA ANCHE È IN REALTÀ E ALLA CLASSE VERA.

PER SEMPLIFICAZIONE ASSUMONO CHE I COSTI PER LE CLASSIFICAZIONI CORrette SONO 0
 $(\mathcal{H}_F|\mathcal{H}_F)=0, C(\mathcal{H}_T|\mathcal{H}_T)=0$. INTUITIVAMENTE NON PARLANO QUANDO LA PREVISIONE È GIUSTA.

INOLTRE X SEMPLICITÀ ASSUMONO $C(\mathcal{H}_F|\mathcal{H}_T) \geq 0$ E $C(\mathcal{H}_T|\mathcal{H}_F) \geq 0$ [I PER UN PUNTO NELLA FORMULA POSSONO ESSERE
 GENISMO NEGATIVI]



	Class \mathcal{H}_F	Class \mathcal{H}_T
Prediction \mathcal{H}_F	0	$C(\mathcal{H}_F \mathcal{H}_T) = C_{fp}$
Prediction \mathcal{H}_T	$C(\mathcal{H}_T \mathcal{H}_F) = C_{fn}$	0

DOVE: C_{fp} È IL COSTO DI UN FALSO POSITIVO, RENDE C_{fn} È IL COSTO DI UN FALSO NEGATIVO.

CALCOLANO ORA IL RISCHIO DELL'ADDESTRAMENTO, PER FARLO DOVEMMO CONFORMARE IL COSTO ATTESO DELLE
 NOSTRE DECISIONI NELL'AMBIENTE DELL'APPLICAZIONE REALE:

$$B_{emp} = \frac{\pi_1}{N_T} \sum_{i|G_i \geq H_T} C(c_i^*|H_T) + \frac{1-\pi_1}{N_F} \sum_{i|G_i > H_F} C(c_i^*|H_F)$$

• π_1 È UNA PROBABILITÀ A MIGLIARE DELLA CLASSE H_T NELL'APPRAZIAMENTO ALTA

• $\frac{1}{N_T} \sum_{i|G_i > H_T} C(c_i^*|H_T)$ È IL COSTO AVERIO PER I CAMPIONI OTTIENUTI DA CLASSE H_T NELL'AMB. SET DI VALUTAZIONE

• N_T È IL NUMERO TOTALE DI CAMPIONI H_T NELL'SET DI VALUTAZIONE

• LA SOMMA CONSIDERA TUTTI I CAMPIONI IN CLASSE VERA $c_i^* = H_T$

• PER OGNI CAMPIONE, CALCOLANO IL COSTO $C(c_i^*|H_T)$ DELL'ESTRAZIONE c_i^*

• c_i^* È LA CLASSE PREFERITA PER IL CAMPIONE x_i

• STESSA LOGICA PER IL COSTO DI FN

Ora consideriamo che per i campioni H_T , il costo è C_{fp} se la classificazione è errata con le
 classi H_F , altrimenti 0.

Per i campioni H_F , il costo è C_{fn} se la classificazione è errata con le
 classi H_T , altrimenti 0.

Possiamo risolvere la formula così

$$B_{emp} = \pi_T \sum_{i:i \in H_T} \frac{C_{f_m} I[c_i = H_T]}{N_T} + (1 - \pi_T) \sum_{i:i \in H_F} \frac{C_{f_F} I[c_i = H_T]}{N_F}$$

Dove I [condizione] è la funzione indicatrice che vale 1 se la condizione è vera o altrimenti.

Ora osserviamo che: $\sum_{i:i \in H_T} \frac{C_{f_m} I[c_i = H_T]}{N_T}$ è la porzione di campioni H_T classificati correttamente, cioè il tasso di FN P_{f_m} .

e $\sum_{i:i \in H_F} \frac{C_{f_F} I[c_i = H_T]}{N_F}$ è la porzione di campioni H_F classificati correttamente, cioè il tasso di FN P_{f_F} .

Quindi $B_{emp} = \pi_T C_{f_m} P_{f_m} + (1 - \pi_T) C_{f_F} P_{f_F}$ questa è anche chiamata (Un-)monotone Detection Cost Function (UDCF).

Detection Cost Function

La UDCF è quindi essenzialmente un altro modo di chiamare l'external data risk, questa ci permette di calcolare sistematicamente le prestazioni di un classificatore (chimicoxa) tenendo conto sia del costo dei vari tipi di errore sia della distinzione tra le classi nell'applicazione.

In formula si scrive:

$$DCF_U(C_{f_m}, C_{f_F}, \pi_T) = \pi_T (P_{f_m} C_{f_m} + (1 - \pi_T) C_{f_F} P_{f_F})$$

Dove:

$C_{f_m} e C_{f_F}$ è il costo rispettivamente di un falso negativo e falso positivo

π_T è invece la probabilità a priori della classe positiva

$P_{f_m} e P_{f_F}$ è il tasso di FN e FP del classificatore che dipende dalla soglia di decisione e che via il classificatore

Per comprendere perché se un classificatore è effettivamente utile, possono confrontarlo con dei sistemi di riferimento semplici chiamati dummy system.

① Sistema che accetta sempre (classifica sempre $c_i = H_T$)

$$P_{f_F} = 0 \quad (\text{tutti i campioni negativi sono classificati erroneamente})$$

$$P_{f_m} = 1 \quad (\text{nessun campione positivo è classificato erroneamente})$$

$$DCF_U = (1 - \pi_T) C_{f_F}$$

② Sistema che rifiuta sempre (classifica sempre $c_i = H_F$)

$$P_{f_F} = 1$$

$$P_{f_m} = 0$$

$$DCF_U = \pi_T C_{f_m}$$

Il motivo di questi sistemi finti è rappresenta ciò che potremo ottenere senza nemmeno guardare i dati, basandoci solo sulla probabilità a priori. Il nostro classificatore dovrebbe allora fare meglio

DI QUESTI. MA COME QUANTIFICARLO?

DCF Normalizzata

PER VALUTARE QUANTO MEGLIO IL NOSTRO CLASSIFICATORE PERFORMA RISPETTO AL MIGLIOR SISTEMA FINITIMO, INTRODUCIAMO LA DCF NORMALIZZATA;

$$DCF(\pi_T, C_F, C_P) = \frac{DCF_U(\pi_T, C_F, C_P)}{\min(\pi_T C_F, (1-\pi_T)C_P)} \text{ IL MILIONE DI GEL STIMATO}$$

SE LA DCF < 1 INDICA CHE IL CLASSIFICATORE È MEGLIO DEL SISTEMA FINITO.

DCF > 1 INDICA CHE IL CLASSIFICATORE NON E' PIÙ INFORMATIVO. VULNERABILE.

DCF > 2 INDICA CHE IL CLASSIFICATORE FA LAVORO.

Nota: IL MILIONE SISTEMA CORRISPONE ALLA DECISIONE CAUTELARE OTTIMALE, BASATO SOLO SULLE INFORMAZIONI A PRIORI, SENZA CONSIDERARE I DATI SPECIFICI DEI SAMPLI (SERIE R, LIKELIHOOD...) [LUNGO LA PONTEVRA]

UNA PROPRIETÀ INNOVANTE DELLA DCF NORMALIZZATA È CHE È INVARIANTE RISPETTO ALLA SCALA. QUESTO SIGNIFICA CHE POSSIAMO AGGIUNGERE I COSTI SENZA CAMBIARE IL VALORE DELLA DCF.

PAGO INVALORIATI TUTTI I COSTI X UNA COSTANTE λ , E LA DCF RESTA EGUALE LA STESSA.



ADESSO VEDO CHE ADDOSSO TUTTI VALORI SEPARATI (C_F, C_P, π_T) POSSANO ALCUNA IL TUTTO CONSIDERANDO UN VALORE CHE CONDENA IN UN UNICO NUMERO L'EFFETTO DELLA PREDICTION ATTRAVERSO LA LASCIA POSITIVA (π_T) E L'INFORMAZIONE RELATIVA DEGLI ERROI (C_F, C_P). → EFFECTIVE PRIOR $\tilde{\pi}$

$$\tilde{\pi} = \frac{\pi_T C_F}{\pi_T C_F + (1-\pi_T) C_P}, \quad 1-\tilde{\pi} = \frac{(1-\pi_T) C_P}{\pi_T C_F + (1-\pi_T) C_P}$$

L'INNOMINARIA DCF DIVENTA:

$$DCF_U(\tilde{\pi}) = \tilde{\pi} P_F + (1-\tilde{\pi}) P_F$$

E LA CORRISPONDENTE DCF NORMALIZZATA MANTIENE LO STESSO VALORE (GRAZIE ALLA PROPIETÀ), QUINDI LA $DCF(\tilde{\pi}, C_F, C_P) = \text{CONTRARIO}(\tilde{\pi}, 1, 1)$.

POTREMO quindi INTERPRETARE $\tilde{\pi}$ L'EFFECTIVE PRIOR CON UN PUNTO EFFETTIVO: SE LA PROBABILITÀ A PIENO DI H_T FOGLI $\tilde{\pi}$ E ASSUMESSO COSTI UNIFORMI ($C_F=C_P$) OTTEREMMO LUSSI STESSI COSTI NORMALIZZATI O UNA NUOVA APPLICAZIONE UNIVALENTI (SO UGUALI DAL PUNTO DI VISTA DELLA DCF).

Relazione con il Tasso di Gnoce

$$err = \frac{\# \text{ DI ERROI (CLASSIFICATORI GRADUATI)}}{\# \text{ TOTALI OBIETTIVI}} = 1 - \text{ACCURACY} =$$

$$= \frac{N_T P_{Tn} + N_F P_{Fn}}{N} = \frac{N_T}{N} P_{Tn} + \frac{N_F}{N} P_{Fn}$$

Dove $\frac{N_T}{N}$ è la portione di campioni positivo nel set di valutazione ovvero il prior empirico.

Quindi (dalle equazioni DCF) possiamo vedere che l'EMPIRICAL RISK COMBINATO A UNO DI UN FATTORE DI SUIA, ALIA DCF, DI UN'applicazione con prior $(\frac{N_T}{N}, 1, 1)$ dove questo prior potrebbe essere diverso dal prior empirico dell'applicazione.

Possiamo considerare il Weighted error rate: $e = \frac{1}{2}(P_{Fn} + P_{Fa})$, ovvero l'errore media dei tassi di errore (considerando una $\pi_T = 0.5$ (normalmente da 0.5)) in questo caso $DCF_u > 0 (\frac{1}{2}, 1, 1)$

Immaginiamo un sistema di autenticazione biometrica con le seguenti caratteristiche:

La probabilità che un utente sia un impostore è $N_F = 0.999$ (quindi $\pi_T = 0.001$)

Il costo di accettare un impostore è $C_{fp} = 10$

Il costo di rifiutare un utente legittimo è $C_{fn} = 1$

Supponiamo che il nostro sistema abbia i seguenti tassi di errore:

$P_{Fn} = 0.02$ (2% di utenti legittimi vengono rifiutati)

$P_{fp} = 0.001$ (0.1% di impostori vengono accettati)

Calcoliamo la DCF non normalizzata:

$$DCF_u = 0.001 \times 1 \times 0.02 + 0.999 \times 10 \times 0.001 = 0.00002 + 0.00999 = 0.01001$$

Ora calcoliamo il costo del sistema fittizio:

Sistema che accetta sempre: $DCF_u = 0.999 \times 10 = 9.99$

Sistema che rifiuta sempre: $DCF_u = 0.001 \times 1 = 0.001$

Il miglior sistema fittizio è quello che rifiuta sempre, con un costo di 0.001.

La DCF normalizzata è quindi:

$$DCF = 0.01001 / 0.001 = 10.01$$

Questo valore ci suggerisce che il nostro sistema performi peggio del sistema fittizio! Questo può sembrare contraintuitivo, ma è dovuto alla combinazione di un prior molto sbilanciato (solo 0.1% di utenti legittimi) e costi asimmetrici.

Il prior effettivo in questo caso è:

$$\pi = (0.001 \times 1) / (0.001 \times 1 + 0.999 \times 10) = 0.001 / 10 = 0.0001$$

Questo valore estremamente basso suggerisce che, dati i costi e i prior specificati, la strategia ottimale sarebbe quasi sempre rifiutare tutti gli utenti.

Final remarks

Evaluation of multiclass tasks is more complex, and we cannot represent any application with a single parameter (effective prior)

However, we can compute the empirical Bayes risk

Also in the multiclass case we can then compute a normalized detection cost, obtained by scaling the empirical Bayes risk by the cost of the best dummy system — in this case, we have K dummy systems, each of them predicting a single class k regardless of the sample

Nella maggior parte dei casi, i classificatori binari non provvedono direttamente una decisione ma leverano un punteggio che risulta quanto forte è il modello "credere" che un campione appartenga alla classe positiva (HT). A seconda del tipo di classificatore abbiamo diversi modi per calcolare lo score.

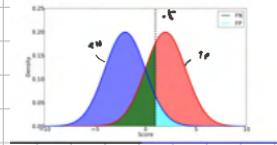
Modelli generativi: log-likelihood ratio $s = \log \frac{P(x|H_T)}{P(x|H_F)}$

Modelli discriminativi: posterior log probability $\approx s = \log \frac{P(H_T|x)}{P(H_F|x)}$

Non probabilistici: (SVM) $s = w^T x$

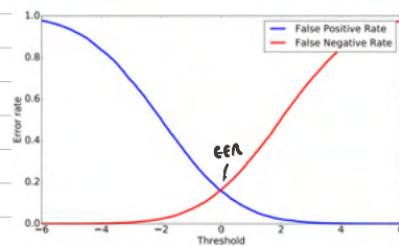
In tutti questi casi il punteggio più alto indica una maggiore propensione verso la classe HT. In decisione finale viene confrontato questo punteggio con una soglia (t) threshold, se il punteggio supera la soglia classifichiamo il campione come HT, se no come HF.

La scelta della soglia ha un impatto significativo sui tassi di errore, valori diversi di t corrispondono a error rate diversi per le 2 classi



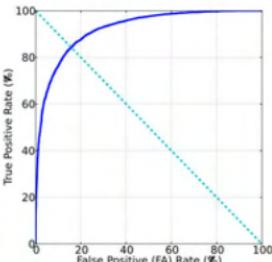
Possiamo anche misurare come il tasso di falsi positivi (FPR) e falsi negativi (FNR) variano al variare della soglia:

- La linea blu rappresenta il tasso di falsi positivi (FPR): quando la soglia è molto bassa (a sinistra del grafico), quasi tutti i campioni vengono classificati come HT, quindi il FPR è vicino a 1. Man mano che la soglia aumenta, il FPR diminuisce.
- La linea rossa rappresenta il tasso di falsi negativi (FNR): quando la soglia è molto bassa, pochi campioni HT vengono classificati erroneamente come HF, quindi il FNR è vicino a 0. Man mano che la soglia aumenta, il FNR cresce.

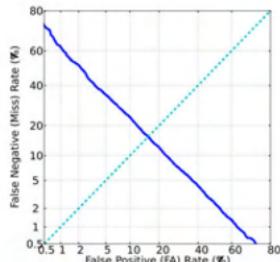


L'equal error rate (EER) corrisponde alla soglia per cui $FPR = FNR$.

- Receiver Operating Characteristic (ROC) curve



- Detection Error Trade-off (DET) curve



Curva ROC (Receiver Operating Characteristic): Questa curva traccia il tasso di veri positivi (TPR = 1 - FNR) rispetto al tasso di falsi positivi (FPR) al variare della soglia. Una curva ROC che si avvicina all'angolo superiore sinistro (alta TPR basso FPR) indica un classificatore con buone prestazioni. La diagonale rappresenta un classificatore casuale. L'area sotto la curva ROC (AUC-ROC) è una metrica popolare: un valore di 1 indica un classificatore perfetto, mentre 0.5 indica un classificatore casuale.

Curva DET (Detection Error Tradeoff): Questa curva traccia direttamente il tasso di falsi negativi (FNR) rispetto al tasso di falsi positivi (FPR). A differenza della curva ROC, entrambi gli assi rappresentano errori, quindi una curva DET che si avvicina all'origine indica un classificatore migliore. La curva DET viene spesso tracciata con scale normali/inverse (probit) per evidenziare meglio le differenze tra classificatori con buone prestazioni.

Bayes Decisions

Le prestazioni di un classificatore dipendono dalla solta selezionata. L'obiettivo è ora trovare la solta ottimale per una data applicazione o più in generale trovare un modo efficace per trasformare i puntelli del classificatore (sia binario che multiclasse) in un'decisione efficaci.

Per fare ciò partiamo dall'assunzione che il nostro classificatore sia probabilistico, cioè in grado di fornire la probabilità a posteriori $P(C=n|x, R)$ per un dato campione x . In realtà il classificatore, sotto l'ipotesi di normalità della sua distribuzione: classificazioni diversi potrebbero avere probabilità diverse per stesse classi, basandosi sulle loro "convintioni" interne.

Il costo atteso di un'azione a secondo le probabilità a posteriori fornite dal classificatore

$$C_{n,R}(a) = E_{C|x,R} [C(a|n)|x, R] = \sum_{n=1}^N C(a|n) P(C=n|x, R)$$

(costo fare l'azione quando la vera classe è n ...)

È una sorta ponderata dei costi di ogni possibile errore dove i pesi sono la probabilità che il campione appartenga a ciascuna classe.

La decisione bayesiana ottimale consiste nello scegliere l'azione $a^*(x, R)$ cui è minima il costo atteso

$$a^*(x, R) = \arg \min_a C_{n,R}(a)$$

AZIOVE (EILLUMINATURA) CHE CONDIZIONA IL MINOR COSTO ATTESO, SECONDO LE CONVENZIONI DEL CLASSIFICATORE.

For example, let's consider we have a 3-class problem, with cost matrix and priors given by

$$C = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}, \quad \pi = \begin{bmatrix} 0.3 \\ 0.4 \\ 0.3 \end{bmatrix}$$

For a test sample x_t , we have computed the posterior class probabilities (using the prior π)

$$q_t = \begin{bmatrix} P(C=1|x_t, \mathcal{R}) \\ P(C=2|x_t, \mathcal{R}) \\ P(C=3|x_t, \mathcal{R}) \end{bmatrix} = \begin{bmatrix} 0.40 \\ 0.25 \\ 0.35 \end{bmatrix}$$

The expected cost of actions "Predict a " are

$$C_{x_t, \mathcal{R}}(1) = 0 \times 0.40 + 1 \times 0.25 + 2 \times 0.35 = 0.95$$

$$C_{x_t, \mathcal{R}}(2) = 1 \times 0.40 + 0 \times 0.25 + 1 \times 0.35 = 0.75$$

$$C_{x_t, \mathcal{R}}(3) = 2 \times 0.40 + 1 \times 0.25 + 0 \times 0.35 = 1.05$$

or, in matrix form:

$$\begin{bmatrix} C_{x_t, \mathcal{R}}(1) \\ C_{x_t, \mathcal{R}}(2) \\ C_{x_t, \mathcal{R}}(3) \end{bmatrix} = C q_t$$

The optimal decision would therefore to assign label 2, even though it has the lowest posterior probability, since the expected cost due to mis-classifications would be lower.

LE DECISIONI BAYESIANE SONO OTTIME DAL PUNTO DI VISTA DEL CLASSIFICATORE. SE IL VALORE DI ϵ È IL CLASSIFICATORE R CONVOLUTO (CIOÈ, APROPRANO LE STESE PROBABILITÀ A POSTERIORI DELLE CLASSI) PER QUALSIASI APPROXIMAZIONE, LE DECISIONI BAYESIANE MINIMIZZANO IL RISCHIO BAYESIANO.

QUANDO DICIAMO CHE "LE DECISIONI BAYESIANE SONO OTTIME DAL PUNTO DI VISTA DEL CLASSIFICATORE" INTENDIAMO CHE QUESTE DECISIONI MINIMIZZANO IL COSTO ATTESO ALLORNO ALLE PROBABILITÀ A POSTERIORI STIMATE DAL CLASSIFICATORE STESSO.

IL CLASSIFICATORE R COSTRUISE UN PROFILLO DEL MONDO BASEATO SUI DATI DI ADDESTRAMENTO E FORNIRÀ DELLE CREDENZE SU COME SONO DISTRIBUITE LE CLASSI INSPERATE ALLE QUANTITÀ X . QUESTE CREDENZE SONO MANIFESTATE DALLE PROBABILITÀ A POSTERIORI $P(C=i|x, R)$. LA DECISIONE BAYESIANA $\delta^*(x, R)$ È QUELLA CHE MINIMIZZA IL COSTO ATTESO SECONDO QUESTE CREDENZE.

MA C'È UN PROBLEMA POTENZIALE: SE L'ADDESTRAMENTO DEL CLASSIFICATORE POTESSE NON CORRISPONDERE ALLA REALTÀ, IL CLASSIFICATORE POTREbbe SOVRASTIMARE o SOTTOSTIMARE CERTE PROBABILITÀ Ecco PERCHE' DISTINGUERANO TRA

* IL VALORE DI ϵ : rappresenta il "mondo reale" o la "verità" sull'applicazione.

* IL CLASSIFICATORE π : rappresenta il nostro modello di quel mondo reale

In FASE "SE IL VALUTAZIONE (ξ) E IL CLASSIFICATORE (π) CONCORDANO" S'INDICA CHE LE PROBABILITÀ A POSTERIORI STIMATE DAL CLASSIFICATORE $P(C=c|x, \pi)$ CORRISPONDONO ESATTAMENTE ALLE PROBABILITÀ A POSTERIORI "VERE" $P(C=c|x, \xi)$ NEL MONDO REALE:

In simboli avremo esempio che: $P(C=c|x, \pi) \sim P(C=c|x, \xi)$ DOVE " \sim " INDICA CHE LE DISTRIBUZIONI SONO EQUIVALENTE (CONSIDERAZIONE LOCALE)

IL RISCHIO ARTESIANO È IL COSTO ATTESO DELLE NOSTRE DECISIONI NEL MONDO REALE:

$$0 = \mathbb{E}_{x \sim f_{\pi}} [C(a(x, \pi)|c)] = \int f_{x|\pi}(x) \sum_{c \in C} C(a(x, \pi)|c) P(C=c|x, \pi) dx$$

$$= \int f_{x|\pi}(x) \mathbb{E}_{c|x, \pi} [C(a(x, \pi)|c)|x] dx$$

SE $C(x, \pi) = C(x, \xi)$ cioè il classificatore e il valutatore concordano sulle probabilità CONDITIONALI, POSSANO SOSTituIRE ξ con π e avremo (se stessa affermazione)

$$0 = \int f_{x|\pi}(x) \mathbb{E}_{c|x, \pi} [C(a(x, \pi)|c)|x] dx$$

ALESSO PER OVALSI UNA DECISIONE $a(x, \pi)$ SUPPORE CHE:

$$\mathbb{E}_{C|x, \pi} [C(a(x, \pi)|c)] = C_{x, \pi}(a(x, \pi)) \geq C_{x, \pi}(a^*(x, \pi)), \quad \forall x$$

PERMÈ $a^*(x, \pi)$ È LA DECISIONE CHE MINIMIZZA IL COSTO SECONDO π

QUINDI

$$0 = \int f_{x|\pi}(x) \mathbb{E}_{c|x, \pi} [C(a(x, \pi)|c)|x] dx \geq \int f_{x|\pi}(x) \mathbb{E}_{c|x, \pi} [C(a^*(x, \pi)|c)|x] dx$$

QUINDI LA DECISIONE ARTESIANA $a^*(x, \pi)$ MINIMIZZA IL RISCHIO ARTESIANO 0, A CONDIZIONE CHE IL CLASSIFICATORE E IL VALUTATORE CONCORDANO SULLE PROBABILITÀ A POSTERIORI.

Non:

Questo risultato NON DIPENDE DALLA DISTRIBUZIONE MARGINALE DELLE CAMMINISTI CHE $f(x)$. A CONDIZIONE CHE LE PROBABILITÀ A POSTERIORI SIANO CORrette, NON IMPORTA COME SONO DISTRIBUITE LE CAMMINISTI, (HE X NELLA POPOLAZIONE).

Per esempio, il nostro classificatore medico potrebbe essere applicato in popolazioni con diverse prevalenze della malattia o diverse distribuzioni dei sintomi, ma finché stima correttamente $P(\text{malattia} | \text{sintomi})$, le sue decisioni saranno comunque ottimali.

Possiamo intendersene quindi l'ottimismo in 2 ruoli distinti

- optimal Bayes decisions of a recognizer \mathcal{R} minimize the Bayes risk, as evaluated by the recognizer itself
- optimal Bayes decisions of a system that has complete data knowledge $\mathcal{R} = \mathcal{E}$ minimize the Bayes risk of the evaluator \mathcal{E}

In the latter case the Bayes risk represents the best possible cost we would pay for classifying test data (but, of course, since we don't have full data knowledge we cannot compute the risk in this case)

Nota: QUANDO PARLAVO DI "CONSEGUIMENTO COMPLETO DEL PROBLEMA" ($n=2$), PONENDO DENTRO CHE QUESTO INTRALUOGO LA CAPACITÀ DI CLASSIFICARE IN MANIERA CORRETTA OGNI UN'UNIONE, ASSUMEVAMO CHE QUESTO È PIÙ CUSSE CONFERITO. O ALLE ALTRE M NON È COSÌ. $0 < p(C=1|x, \{\}) < 1$ C'È SEMPRE UN'INCERTITUDINE SU PROBABILITÀ NON È MAI ASSOLUTA

Let's now consider again a binary problem with cost matrix

	Class \mathcal{H}_F	Class \mathcal{H}_T
Prediction \mathcal{H}_F	0	$C(\mathcal{H}_F \mathcal{H}_T) = C_{fn}$
Prediction \mathcal{H}_T	$C(\mathcal{H}_T \mathcal{H}_F) = C_{fp}$	0

where C_{fn} is the cost of false negative errors, C_{fp} is the cost of false positive errors

The expected Bayes cost for action \mathcal{H}_T (i.e. for predicting \mathcal{H}_T) is

$$C_{x,\mathcal{R}}(\mathcal{H}_T) = C_{fp}P(\mathcal{H}_F|x, \mathcal{R}) + 0 \cdot P(\mathcal{H}_T|x, \mathcal{R}) = C_{fp}P(\mathcal{H}_F|x, \mathcal{R})$$

whereas the cost for action \mathcal{H}_F (i.e. for predicting \mathcal{H}_F) is

$$C_{x,\mathcal{R}}(\mathcal{H}_F) = C_{fn}P(\mathcal{H}_T|x, \mathcal{R}) + 0 \cdot P(\mathcal{H}_F|x, \mathcal{R}) = C_{fn}P(\mathcal{H}_T|x, \mathcal{R})$$

The optimal decision is the labeling that has lowest cost

LA DECISIONE OTTIMALE È QUELLA CON IL MINOR COSTO ATTESO. QUINDI:

CLASSIFICATORE CON H_T SE: $C_{fp}P(\mathcal{H}_F|x, \mathcal{R}) < C_{fn}P(\mathcal{H}_T|x, \mathcal{R})$

CLASSIFICATORE CON H_F SE: $C_{fp}P(\mathcal{H}_F|x, \mathcal{R}) > C_{fn}P(\mathcal{H}_T|x, \mathcal{R})$

Se sono uguali entrambe le azioni vanno bene

Possiamo esprimere le decisioni ottimali anche in un altro modo, definendo una funzione discriminante

$$r(x) = \log \frac{C_{fn}P(\mathcal{H}_T|x, \mathcal{R})}{C_{fp}P(\mathcal{H}_F|x, \mathcal{R})}, \quad a^*(x, \mathcal{R}) = \begin{cases} H_T & r(x) > 0 \\ H_F & r(x) \leq 0 \end{cases}$$

Se H è un nuovo evento, possiamo esprimere $P(H/x)$ e $P(H^c/x)$ in termini di likelihood priori.

$$r(x) = \log \frac{\pi_T C_{fn}}{1 - \pi_T C_{fp}} \frac{f_{X|H,R}(x|H)}{f_{X|H,R}(x|H^c)}$$

Dove: $\pi_T = P(H=H_1)$ è la prior probability della classe H_1 .

Il primo termine $\frac{\pi_T C_{fn}}{1 - \pi_T C_{fp}}$ è costante (non dipende da x) e rappresenta l'effetto condizionato delle proporzioni a priori, ovvero costi. Il secondo termine è il $r(x)$ rapporto di verosimiglianza logaritmico, che misura quanto più probabile sia osservare le caratteristiche x in un'azione di classe H_1 rispetto a quella H^c .

y

La regola di decisione diventa $r(x) \leq 0 \iff \log \frac{f_{X|H,R}(x|H)}{f_{X|H,R}(x|H^c)} \leq -\log \frac{\pi_T C_{fn}}{(1 - \pi_T) C_{fp}}$

The triplet (π_T, C_{fn}, C_{fp}) represents the working point of an application for a binary classification task.

We can show that, as for the Bayes empirical risk, the triplet is actually redundant, in the sense that we can build equivalent applications $(\tilde{\pi}_T, C'_{fn}, C'_{fp})$ which have the same decision rule as the original application, but different costs and priors.

For example, we can represent a binary application in terms of the effective prior

$$\tilde{\pi} = \frac{\pi_T C_{fn}}{\pi_T C_{fn} + (1 - \pi_T) C_{fp}}$$

as application $(\tilde{\pi}, 1, 1)$. The application $(\tilde{\pi}, 1, 1)$ is indeed equivalent to the application (C_{fn}, C_{fp}, π_T) , since

$$\frac{\tilde{\pi}}{1 - \tilde{\pi}} = \frac{\frac{\pi_T C_{fn}}{\pi_T C_{fn} + (1 - \pi_T) C_{fp}}}{1 - \frac{\pi_T C_{fn}}{\pi_T C_{fn} + (1 - \pi_T) C_{fp}}} = \frac{\pi_T C_{fn}}{(1 - \pi_T) C_{fp}}$$

Immaginiamo un'applicazione con:

$$\pi_T = 0.01 \text{ (malattia rara)}$$

$$C_{fn} = 50 \text{ (costo alto per mancata diagnosi)}$$

$$C_{fp} = 1 \text{ (costo basso per falso allarme)}$$

Il prior effettivo sarebbe:

$$\tilde{\pi} = (0.01 \times 50) / (0.01 \times 50 + 0.99 \times 1) \approx 0.34$$

Quindi, l'applicazione $(0.01, 50, 1)$ è equivalente all'applicazione $(0.34, 1, 1)$. Sebbene la prima rappresenti una malattia rara, con costi fortemente asimmetrici, la seconda rappresenta una condizione più comune con costi uguali. Entrambe porterebbero alle stesse decisioni ottimali!

For systems producing well-calibrated log-likelihood ratios

$$s = \log \frac{f_{X|C}(x|\mathcal{H}_T)}{f_{X|C}(x|\mathcal{H}_F)}$$

the optimal threshold (optimal Bayes decision) becomes, in terms of effective prior:

$$t = -\log \frac{\tilde{\pi}}{1 - \tilde{\pi}}$$

QUESTA FORMULA SEPARA CHIARAMENTE IL RUOLO DEL CLASSIFICATORE (CHE CALCOLA LLR) DAL RUOLO DELL'APPLICAZIONE (CHE ORGANIZZA IL PRIOR EFFETTIVO E QUINDI LA SOGGLA)

- Se $\pi = 0.5$ (prior effettivo equilibrato), la soglia è $t = 0$. Ciò significa che classificheremo come positivo se il LLR è positivo, ovvero se le caratteristiche sono più probabili nella classe positiva.
- Se $\pi > 0.5$ (prior effettivo favorisce la classe positiva), la soglia è negativa. Ciò significa che classificheremo come positivo anche con evidenze leggermente a favore della classe negativa, perché il prior o i costi ci spingono verso la classe positiva.
- Se $\pi < 0.5$ (prior effettivo favorisce la classe negativa), la soglia è positiva. Ciò significa che richiederemo evidenze più forti per classificare come positivo.

GLI LLRs PERMETTONO DI SEPARARE NETTAMENTE IL CLASSIFICATORE DALL'APPLICAZIONE. IN TECNICA CUNN UNA PERFEZIONE CALCOLATA POSSERMO ADATTARE IL NOSTRO SISTEMA A OLTRENSI NUOVA APPLICAZIONE SENZA DOVERE REFERNIRSI A SOGGLA DI DECISIONE. PERO'

In general, systems often do not produce well-calibrated LLRs

- Non-probabilistic scores (e.g. SVM)
- Mis-match between train and test populations
- Non-accurate model assumptions

In these cases, we say that scores are **mis-calibrated**

The theoretical threshold $-\log \frac{\tilde{\pi}}{1 - \tilde{\pi}}$ is not optimal anymore

ANALISI: QUANDO
APPROSSIMIAMO A
UNA DISTRIBUZIONE
LAUSIANA DATI CHE
NON LO SO ANCONTRI
NIS-CALIBRAZIONE

For a given application, we can measure the additional cost due to the use of mis-calibrated scores

We can define the **minimum** cost DCF_{min} corresponding to the use of the optimal threshold for a given evaluation set

We consider varying the threshold t to obtain all possible combinations of P_{fn} and P_{fp} for the evaluation set

We select the threshold corresponding to the lowest DCF

The corresponding value DCF_{min} is the cost we would pay if we knew before-hand the optimal threshold for the evaluation set

We can think of this value as a measure of the quality of the classifier

We can also compute the **actual DCF** obtained using the threshold corresponding to the effective prior π

The difference between the actual and minimum DCF represents the loss due to score mis-calibration

Il DCF attuale e il DCF minimo usano la stessa distribuzione a priori (prior), ma differiscono nella soglia di decisione che viene applicata.

Facciamo chiarezza su questa distinzione fondamentale:

DCF attuale (DCFact): Questo valore si calcola utilizzando la soglia teorica derivata dal prior effettivo π . La formula per questa soglia è $-\log(\pi/(1-\pi))$. Il DCF attuale misura il costo che pagheremmo se applicassimo questa soglia teorica, assumendo che i punteggi del classificatore siano perfettamente calibrati.

DCF minimo (DCFmin): Questo valore si calcola trovando empiricamente la soglia ottimale che minimizza il DCF sul set di valutazione, a prescindere dalla formula teorica. Esploriamo tutte le possibili soglie e scegliamo quella che dà il DCF più basso.

In entrambi i casi usiamo la stessa formula per calcolare il DCF:

$$DCF = \pi T \times C_{fn} \times P_{fn} + (1 - \pi) \times C_{fp} \times P_{fp}$$

dove πT , C_{fn} e C_{fp} sono identici in entrambi i calcoli. La differenza sta nei valori di P_{fn} e P_{fp} , che dipendono dalla soglia di decisione scelta.

Per capire meglio, immaginiamo un esempio:

Abbiamo un classificatore che produce punteggi tra -10 e +10

Il prior effettivo è $\pi = 0.3$, quindi la soglia teorica è $-\log(0.3/0.7) \approx 0.85$

Utilizzando questa soglia teorica di 0.85, otteniamo $P_{fn} = 0.20$ e $P_{fp} = 0.15$.

Il DCF attuale sarebbe quindi: $0.3 \times 1 \times 0.20 + 0.7 \times 1 \times 0.15 = 0.165$

Ma se testiamo tutte le possibili soglie, potremmo scoprire che:

Con una soglia di 1.5, otteniamo $P_{fn} = 0.25$ e $P_{fp} = 0.10$

Il DCF sarebbe: $0.3 \times 1 \times 0.25 + 0.7 \times 1 \times 0.10 = 0.145$

In questo caso, il DCF_{min} sarebbe 0.145, ottenuto con una soglia diversa da quella teorica.

La differenza tra $DCFact$ e DCF_{min} ($0.165 - 0.145 = 0.02$ in questo esempio) rappresenta il costo aggiuntivo dovuto alla miscalibrazione dei punteggi. Se i punteggi fossero perfettamente calibrati, la soglia teorica sarebbe identica alla soglia ottimale, e $DCFact$ coinciderebbe con DCF_{min} .

We can also compare different systems over different applications through Bayes error plots

These plots can be used to report actual and / or minimum DCF for different applications

A binary application is parametrized by a single value π

We can thus plot the DCF as a function of prior log-odds $\log \frac{\pi}{1-\pi}$, i.e. the negative of the Bayes optimal threshold.

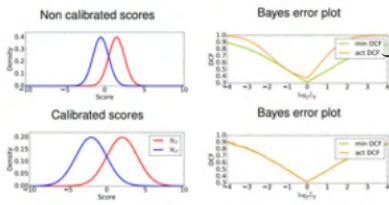
abbiamo alzato i valori di P_{fp} , P_{fn} cambiando il t
Fatto ciò, calcolando la DCFatt con $T=1$: addiamo
ottimale vicini migliori (cioè significa che il
classificatore continua
ancora al suo interno
informazione
prior)

AVVIENE AL SVO INTERNO
INFORMAZIONE
prior

DCF TUTTI SETT CONVVAE

CONSIDERANDO IL PROFILO
APPLICATIVO (CON TUTTI I

costi viene una soluz
e spesso è molto x
quel'applicazione)



Unmodelled
threshold learned

IL SISTEMA È PERN CALIBRATO
E LA SOLLA TEORIA FORMALE
PRESENTA QUASI OTIMALI

!

PER L'APPALUTAZIONE
SU CUI APPARISCONO IL
NODEO,

La calibrazione è cruciale in molti contesti applicativi:

Sistemi di decisione in ambito medico: Una buona calibrazione garantisce che le probabilità stimate di malattia riflettano accuratamente le probabilità reali, permettendo decisioni terapeutiche appropriate.

Sistemi di rilevamento delle frodi: La calibrazione permette di adattare il sistema a diversi contesti (es. diversi tipi di transazioni) senza dover riaddestrare il modello.

Sistemi biometrici: In contesti di sicurezza una buona calibrazione permette di stimare correttamente il rischio di falsi positivi e falsi negativi.

Combinazione di più classificatori: Quando combiniamo diversi classificatori, la calibrazione dei loro punteggi è essenziale per dare il peso appropriato a ciascun sistema.

In tutti questi casi un sistema ben calibrato non solo fornisce decisioni accurate, ma anche stime affidabili dell'incertezza associata a tali decisioni, permettendo un processo decisionale più informato e robusto.



