Normalizzazione

TRASFORM. TRA 0 e 1

$(0,1)$ → c d

$$x' = \frac{x-a}{b-a}$$

TRASF. TRA -1 e 1

$(-1,1)$ → c d

$$x' = -1 + \frac{x-a}{b-a}\,2 = 2\,\frac{x-a}{b-a} - 1$$

scaletura + traslazione del caso $(0,1)$

STANDARDIZZAZIONE

$$x' = \frac{x-\bar{x}}{s_x} \qquad \bar{x}' = \frac{\bar{x}-\bar{x}}{s_x} = 0 \qquad s_{x'}^2 = \frac{1}{s_x^2}\,s_x^2 = 1 \qquad s_{x'} = 1$$

La traslazione di $-\bar{x}$ non ha effetto

Non è tanto diversa dalla normalizzazione

$x' = \log x$ Si passa da dei dati che aumentano di tipo esponenziale a un qualcosa di lineare

(Scala logaritmica)

LATEX TBD end

LEZIONE 08 - 25/03/2025

ANOVA (ANALISI DELLA VARIANZA)

Tecnica che si utilizza su popolazione ~~statistica~~ suddivisa in sottopopolazioni.

Si abbiano n individui suddivisi in G gruppi. Vengano queste osservazioni ordinate in base ai gruppi: (Si lavora in G gruppi indipendentemente dalla loro composizione)

x_j^i → i gruppo, j → osserv. all'interno del gruppo

n_i ampiezza del gruppo i

$$x_1^1, x_2^1, \dots, x_{n_1}^1, x_1^2, \dots, x_{n_2}^2, \dots, x_1^G, \dots, x_{n_G}^G$$ → G ultimo gruppo

Media campionaria

$$\frac{1}{n_i}\sum_{j=1}^{n_i} x_j^i = \bar{x}^i$$ Media camp. gruppo i

$$n_i \bar{x}^i = \sum_{j=1}^{n_i} x_j^i$$

$$\rightarrow \sum_{i=1}^{G} n_i \bar{x}^i = \sum_{i=1}^{G}\sum_{j=1}^{n_i} x_j^i$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{G} n_i \bar{x}^i$$ Media totale

$$SS_T = \sum_{i=1}^{G}\sum_{j=1}^{n_i} (x_j^i - \bar{x})^2$$ Varianza rispetto alla media totale

Somma dei quadrati rispetto alla totale di osservazioni

SUM OF SQUARE TOTAL

$$\frac{1}{n-1}\,SS_T$$ VAR. TOTALE

$$SS_W = \sum_{i=1}^{G} \sum_{j=1}^{n} (x_j^i - \bar{x}^i)^2 \longrightarrow \frac{1}{n-G} SS_W$$ VARIANZA ENTRO I GRUPPI

SUM OF SQUARE WITHIN

$$SS_B = \sum_{i=1}^{G} n_i (\bar{x}^i - \bar{x})^2$$

SS B S. OF S. BETWEEN

Si calcola la varianza tra le medie campionarie dei gruppi. Gruppi con dentro + individui hanno un poco maggiore (x questo si moltiplica per n_i lo scarto)

$$\frac{1}{G-1} SS_B$$ VARIANZA TRA I GRUPPI

$SS_T = SS_W + SS_B$ è dimostrabile, non lo faremo. (c'è nelle note)

Vale sempre

$$\frac{SS_T}{n-1} = \frac{SS_W}{n-1} + \frac{SS_B}{n-1} = \frac{n-G}{n-G} \frac{SS_W}{n-1} + \frac{G-1}{G-1} \frac{SS_B}{n-1} =$$

$$\underbrace{\frac{1}{n-1} \cdot SS_T}_{\text{VAR. TOTALE}} = \underbrace{\frac{n-G}{n-1} \cdot \frac{1}{n-G} SS_W}_{\text{VAR. ENTRO I GRUPPI}} + \underbrace{\frac{G-1}{n-1} \cdot \frac{1}{G-1} SS_B}_{\text{VAR. TRA I GRUPPI}}$$

LA VAR. TRA I GRUPPI, se è trascurabile, significa che VAR. TOT. e VAR. ENTRO I GRUPPI sono approssimativamente simili

Se non è trascurabili, in uno dei gruppi, considerando quelle risorse, c'è una differenza significativa...

La divisione in gruppi è fornita a priori (non bisogna trovare una combinazione che rispetti le nostre aspettative, abbiamo i dati divisi in gruppi e li analizziamo)

(Test statistici)

CLASSIFICATORI

Automatismo che ci permette di classificare degli oggetti (es. ALBERI DI DECISIONE)

Bisogna valutare se questi classificatori siano buoni o meno. Sono dei predittori.

Si parlerà di classificatori binari x comodità.

Classe positiva / classe negativa.

Si organizzano come delle matrici:

- RIGHE: VALORE EFFETTIVO Positivo Negativo
- COLONNE: PREDIZIONE Positivo Negativo

VALORE EFFETTIVO

PREDIZIONE	P	N
P	True Positive	False Positive
N	False Negative	True Negative

- Se si somma sulla prima colonna si ha la cardinalità degli elementi realmente positivi
- Idem x la seconda (negativi)
- se si somma tutto si ottiene la cardinalità del campione

ACCURATEZZA: $\frac{TP + TN}{TP + FP + FN + TN}$ $\frac{\#\text{PREDIZIONI CORRETTE}}{\#\text{CASI TOTALI}}$

→ 1 BUONO
→ 0 ~~PESSIMO~~ NON ACCURATO
TENDERA

Metrica naturale e facile da interpretare. In alcuni casi è però poco significativa. In una situazione con uno sbilanciamento forte nei valori effettivi dei casi, l'accuratezza perde di utilità (x esempio se si ha 1 solo positivo ogni 10k) (A)

Bisogna vedere quanto si è bravi a
- prevedere i positivi guardando solo i positivi
- " i negativi " " negativi

} vanno fatte entrambe

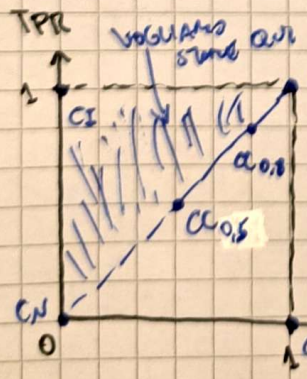
Si fanno 2 verifiche separate:

$TruePosiveRate = \frac{TP}{TOT\ P}$ DA 0 A 1 → BUONO A PREDIRE I POSITIVI

$TrueNegativeRate = \frac{TN}{TOT\ N}$

[TPR detto anche recall o sensitivity
TNR detto anche specificity]

Considerando entrambe TPR e TNR si possono fare delle considerazioni sul nostro classificatore.



(A) CP classificatore costante positivo

	P	N
P	TOT P	TOT N
N	0	0

1 - TNR = 1 - 0 = 1
TPR = 1

CN

0	0	
TOT P	TOT N	

1 - TNR = 1 - 1 = 0
TPR = 0

CE classif. errato

0	TOT N
TOT P	0

1 - TNR = 1 - 0 = 1
TPR = 0

CI Classificatore ideale

TOT P	0
0	TOT N

1 - TNR = 1 - 1 = 0
TPR = 1

Sulla diagonale

$CC_{0,5}$ classificatore casuale con 0,5 di probabilità

$\frac{1}{2}$ TOTP	$\frac{1}{2}$ TOTN
$\frac{1}{2}$ TOTP	$\frac{1}{2}$ TOTN

$1 - TNR = \frac{1}{2}$

$TPR = \frac{1}{2}$

Se si è sulla diagonale si ha un comportamento ~~equo~~ equiparabile a quello del lancio della moneta

⇒ poco significativo

$CC_{0,8}$

0.8 TOTP	0,8 TOTN
0.2 TOTN	0,2 TOTN

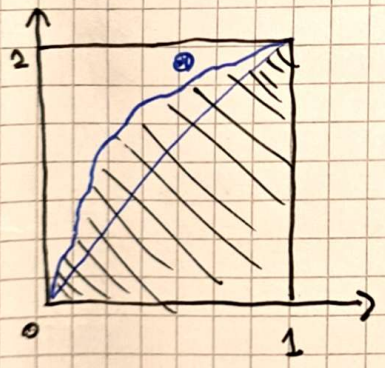
$1 - TNR = 0,8$

$TPR = 0,8$

- Se si ha un classificatore che sta nella parte bassa, si inverte la predizione.

CLASSIFICATORE A SOGLIA Se sto sotto la soglia negativo, sopra positivo (esempio)

Come si fissa la soglia?



$$c(x) = \begin{cases} P & \text{se } x > \tau \text{ (tau)} \\ N & \text{se } x \leq \tau \end{cases}$$

Se si prende come τ = valore minimo possibile nel campione si è nel caso CP

Se τ = valore max possibile nel campione si è nel caso CN

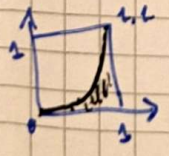
se τ cresce gradualmente si ottiene la curva ROC

Più l'area III è vicino a 1 (si tende a essere nel caso migliore), più si ottiene una curva detta ROC. L'AREA SOTTO LA CURVA ROC PRENDE IL NOME DI AUROC (AREA UNDER ROC)

↳ Serve a valutare la bontà di un classificatore indipendentemente dalla soglia.
> AREA = MIGLIOR CLASSIFICATORE INDIPENDENTE DALLA SOGLIA SCELTA

Si prendano due classificatori: Se la curva ROC di uno domina l'altra, allora quel classificatore ha risultati migliori

SI PUÒ FARE IL DISCORSO INVERSO & COMPLEMENTARE



Se lo si spacchia è buono

$B = A \cup (B \setminus A)$ sono disgiunti;

$P(B) = P(A) + P(B \setminus A) \geq 0$ ecc...

ESERCIZIO LEZIONE 13

$P(E|M) = 1 - P(\bar{E}|M)$ Dimostrazione

$P(A|M) = \frac{P(A \cap M)}{P(M)}$ $P(\bar{A}|M) = \frac{P(\bar{A} \cap M)}{P(M)}$

Dato che $(A \cap M) \cup (\bar{A} \cap M) = M$ e i due eventi sono disgiunti:

$P(M) = P((A \cap M) \cup (\bar{A} \cap M)) = P(A \cap M) + P(\bar{A} \cap M)$

Dividendo entrambi i membri per $P(M)$:

$$\frac{P(A \cap M)}{P(M)} + \frac{P(\bar{A} \cap M)}{P(M)} = 1 \Rightarrow P(A|M) = 1 - P(\bar{A}|M)$$