



Università degli Studi dell'Insubria
Dipartimento di Scienze Teoriche e Applicate

Big Data
a.a. 2022/2023

Lezione05: Score e Ranking

Davide Tosi

Dipartimento di Scienze Teoriche e Applicate

davide.tosi@uninsubria.it



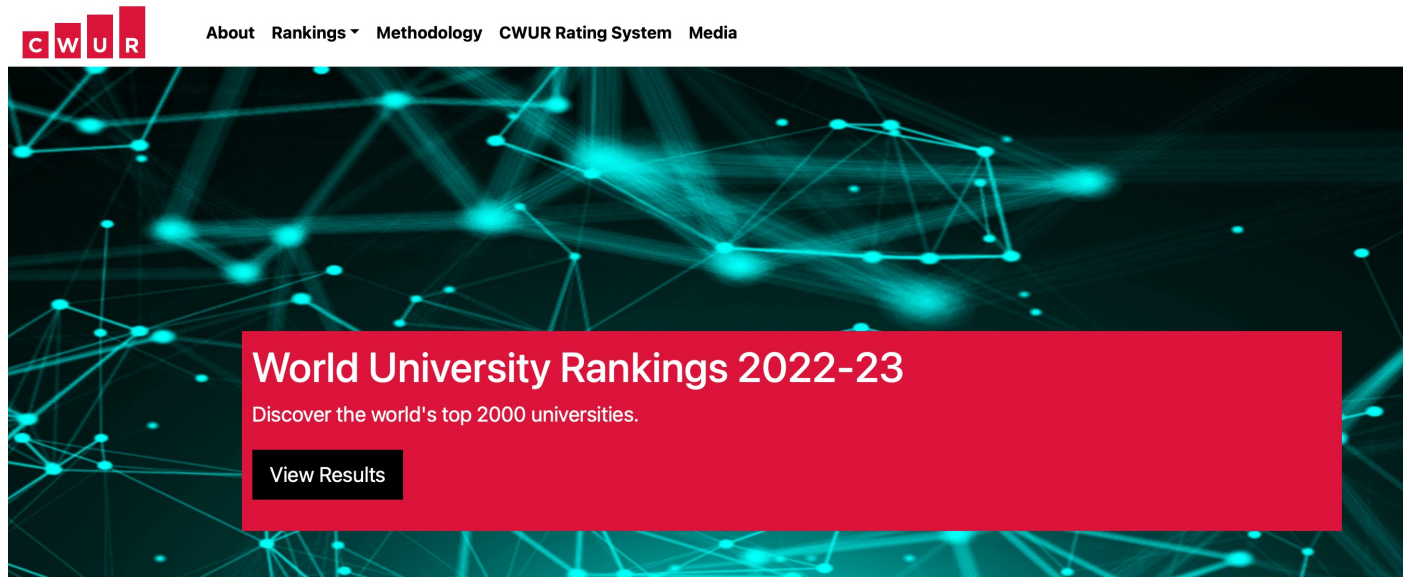
Funzioni di Scoring e Ranking

Funzioni di Scoring: misure che riducono record multi-dimensionali in un singolo valore per evidenziare alcune peculiarità importanti del dato

Funzioni di Ranking: ordinano elementi solitamente in base agli score

... **tecniche utili ed euristiche** per iniziare a stuzzicare la comprensione di grandi dataset (pur essendo tecniche difficilmente convalidabili scientificamente contrariamente ai metodi statistici)

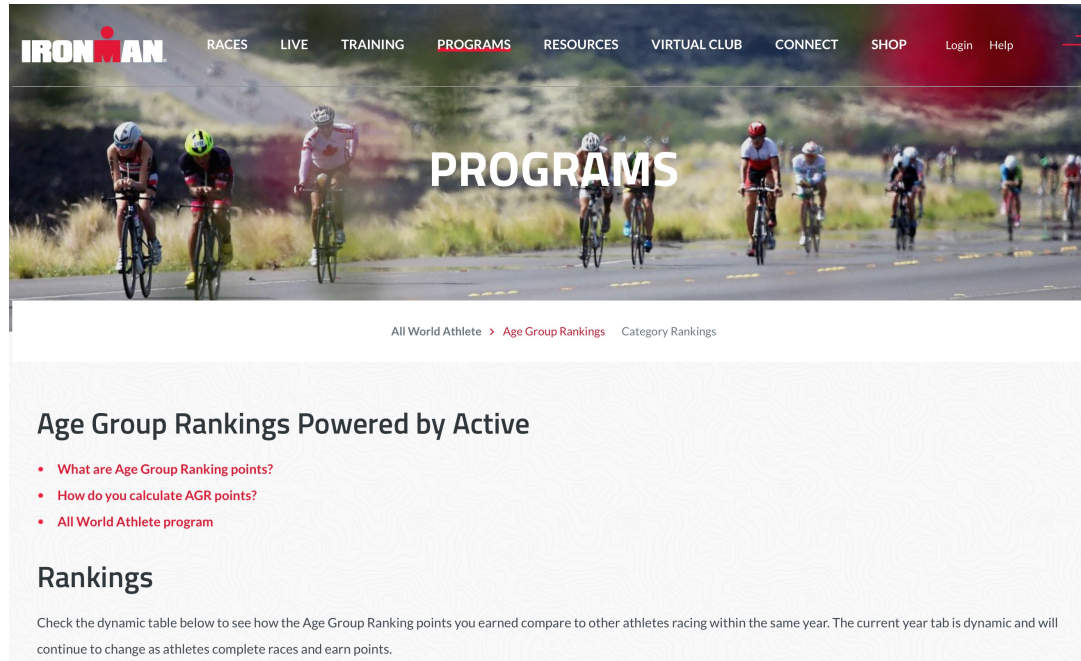
Esempio: Funzioni di Scoring e Ranking



Funzioni di Scoring: “CWUR uses seven objective and robust indicators grouped into four areas to rank the world’s universities (see <https://cwur.org> - Methodology)”

Funzioni di Ranking: 5 possibili ordinamenti per:
National, Education, Employability, Faculty, Research Rank

Esempio: Funzioni di Scoring e Ranking



Funzioni di Scoring: ogni gara, posizionamento, tempo di gara, ecc danno dei punti, che vengono sommati alla fine dell'anno

Funzioni di Ranking: diversi possibili ordinamenti per: Overall, Gender, Swim, Bike, Run, ... Ranks

Funzioni di Scoring e Ranking

Altro tipico esempio è l'assegnazione del voto finale ad uno studente del mio esame.

Il voto finale è una semplice funzione che mette insieme diversi aspetti, al fine di riflettere la reale performance dello studente:

$$\sum \text{punteggio task pratici} + \text{eventuale prova orale}$$

Da tenere bene a mente:

- **Grado di Arbitrarietà:** ogni docente usa differenti variabili di valutazione
- **Mancanza di Dati di Validazione:** non esiste il gold standard per valutarvi
- **Robustezza Generale:** diversi sistemi di valutazione producono risultati simili

Funzioni di Scoring e Ranking

Il problema critico nella progettazione di Funzioni di Scoring è che non esiste un gold standard/risposta giusta.

Le tecniche di machine learning come la regressione lineare possono apprendere una funzione di scoring dalle funzionalità se si dispone di dati di training, cosa che in genere non è possibile.

Gli algoritmi di ranking di Google, ad esempio, si addestrano sui dati dei clic.



Il BMI (Body-Mass Index)

Un esempio convincente di quanto detto è quello del BMI
(indice definito per verificare se il ns peso è sotto controllo)

$$BMI = peso/altezza^2$$

Peso è in Kg e altezza in m

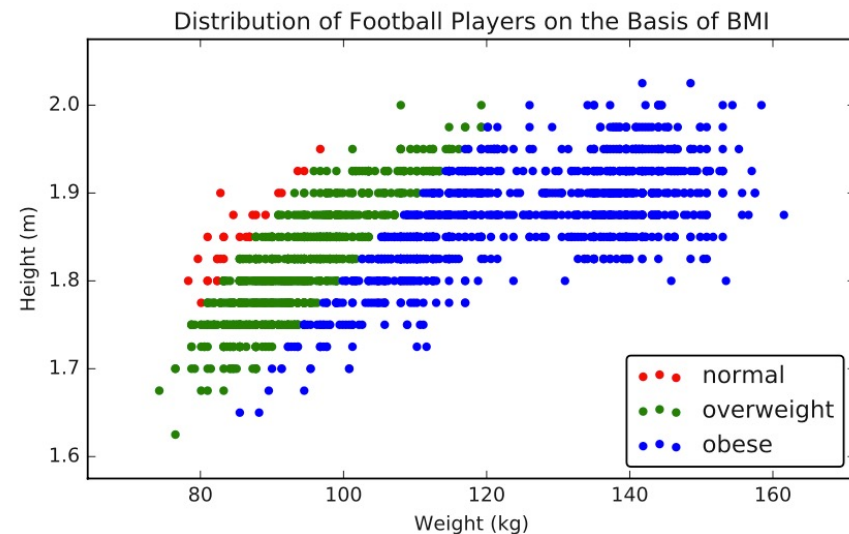
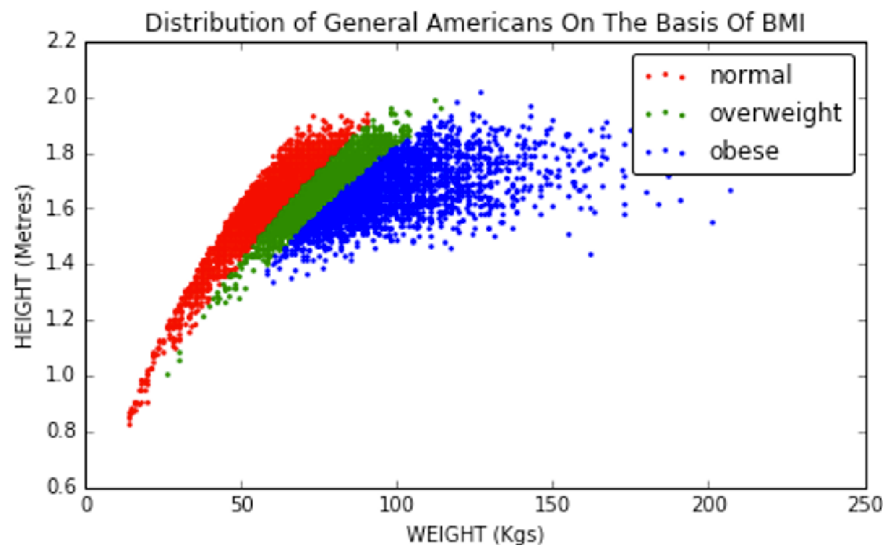
- Sottopeso: < 18.5
- Normopeso: 18.5 a 25.0
- Sovrappeso: 25.0 a 30.0
- Obeso: > 30.0

$$Tosi = \frac{67.0}{1.75^2} = 21.8$$

II BMI (Body-Mass Index)

BMI è facile da interpretare, ampiamente usato, e generalmente correla con il grasso corporeo (la massa dovrebbe scalare con il quadrato dell'altezza).

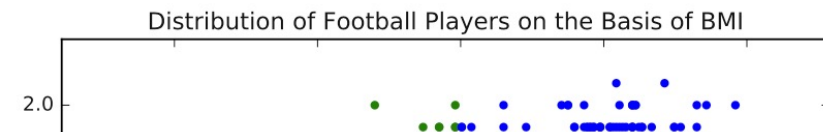
Il problema sta in quel “generalmente”. La massa infatti dipende dalla massa grassa (grasso corporeo) ma anche dalla massa magra (muscoli).



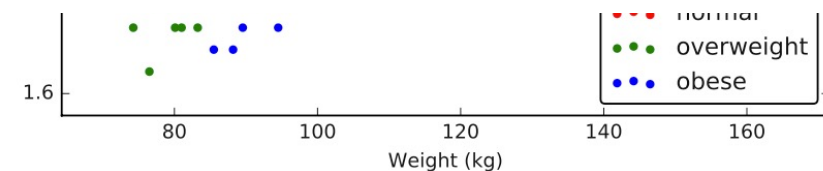
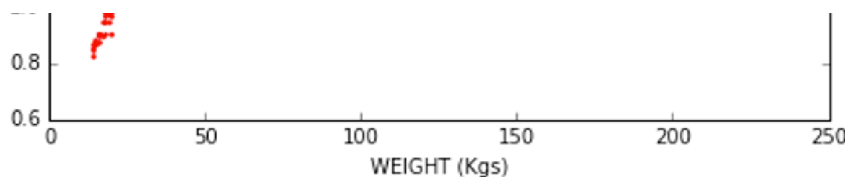
Il BMI (Body-Mass Index)

BMI è facile da interpretare, ampiamente usato, e generalmente correla con il grasso corporeo (la massa dovrebbe scalare con il quadrato dell'altezza).

Il problema sta in quel “generalmente”. La massa infatti dipende dalla massa grassa (grasso corporeo) ma anche dalla massa magra (muscoli).



Quindi veramente i Giocatori professionisti di Football Americano sono quasi tutti obesi?



Sviluppare Sistemi di Scoring e Ranking

Il BMI è un **Proxy** (i.e., una approssimazione del mondo reale) che descrive facilmente e correla abbastanza bene con la variabile grasso corporeo della popolazione media.

I **ranking** sono permutazioni che ordinano n entità per merito, generalmente costruite ordinando l'output di un sistema di punteggio.

In generale, l'ordinamento dei risultati di un sistema di punteggio produce una classifica numerica.

Dualmente, la posizione in classifica di ciascun elemento ha associato anche un punteggio numerico.

Chi quindi fornisce una miglior rappresentazione del dato?

Chi quindi fornisce una miglior rappresentazione del dato?

E' più interessante sapere che l'Università Insubria si colloca in 882esima posizione nel CWUR World University Ranking o che ha un score CWUR di 71.0?

1. Qual è la distribuzione sottostante dei punteggi?

(i.e., quanto cambia lo score dalla posizione #881 a #882?)

2. Le differenze nelle classifiche sono lineari?

(i.e., la differenza tra #1 e #2 è uguale alla differenza tra #881 e #882?)

3. Hai considerato gli estremi e i valori di mezzo?

(i.e., #882 su quanti? Che differenza c'è tra score #882 e #1?, ecc.)

Sviluppare Sistemi di Scoring e Ranking

I sistemi di punteggio ben progettati hanno spesso una distribuzione a campana. Con i punteggi concentrati attorno alla media, piccole differenze di punteggio possono significare grandi differenze di rango.

In una distribuzione normale, aumentando il punteggio dalla media di una deviazione standard (1σ) si passa dal 50° percentile all'84° percentile. Ma lo stesso cambiamento di dimensioni da 1σ a 2σ ti porta solo dall'84° al 92,5° percentile.

L'essere scivolati dalla posizione #877 (21/22) a #882 (22/23) per la ns. Università non è dipeso da noi (il ns. score è rimasto 71.0)

Buone Funzioni di Scoring sono:

- Facilmente calcolabili
(es. BMI: 2 sole var legate da una funzione matematica semplice)
- Facilmente comprensibili
(alla popolazione affinché siano poi ampiamente riconosciute e usate)
- Danno interpretazione monotona delle variabili
(univoca indicazione di come le variabili correlano con l'output)
- Producono risultati soddisfacenti sui valori anomali
(che se si conosce il dominio dovrebbero essere facilmente identificabili)
- Usano variabili sistematicamente normalizzate
(vedi slide a seguire)
- Con pochi "pareggi" a livello di indici
(classificare la *praticità delle persone* in base al numero di dita non è granché rivelatore)

Normalizzazione e Z-score:

Le variabili tratte da distribuzioni a campana possono impattare sensibilmente nelle funzioni di punteggio: ci saranno valori anomali alle code di entrambe le estremità che corrispondono agli elementi migliori/peggiori, oltre a un picco nel mezzo di elementi i cui punteggi dovrebbero essere tutti relativamente simili.

Queste variabili normalmente distribuite dovrebbero essere normalizzate prima di sommarle, in modo che tutte le caratteristiche abbiano medie e varianza comparabili → nessuna singola caratteristica ha un impatto troppo dominante sui risultati.

In generale, si calcola lo **Z-score** e si sommano i contributi (+ con variabili che correlano positivamente, - con quelle che correlano negativamente) con un peso uniforme.

Normalizzazione e Z-score

Z-score trasforma un arbitrario set di valori in un range uniforme:

$$Z_i = (X_i - \bar{X}) / \sigma$$

Z-score ha *media* = 0 e *dev.std* = 1

Il segno identifica se il valore è sopra o sotto media.

Esempio: voti presi nell'ultimo mio appello d'esame

Score | 19, 22, 24, 22, 30, 22, 26, 24, 22

$\mu(\text{Score}) = 23.4$

$\sigma(\text{Score}) = 2.95$

Z | -1.51, -0.49, 0.19, -0.49, 2.22, -0.49, 0.87, 0.19, -0.49

$\mu(Z) = 0.0$

$\sigma(Z) = 1.00$



Sviluppare Sistemi di Scoring e Ranking

Esercizio1 di Normalizzazione

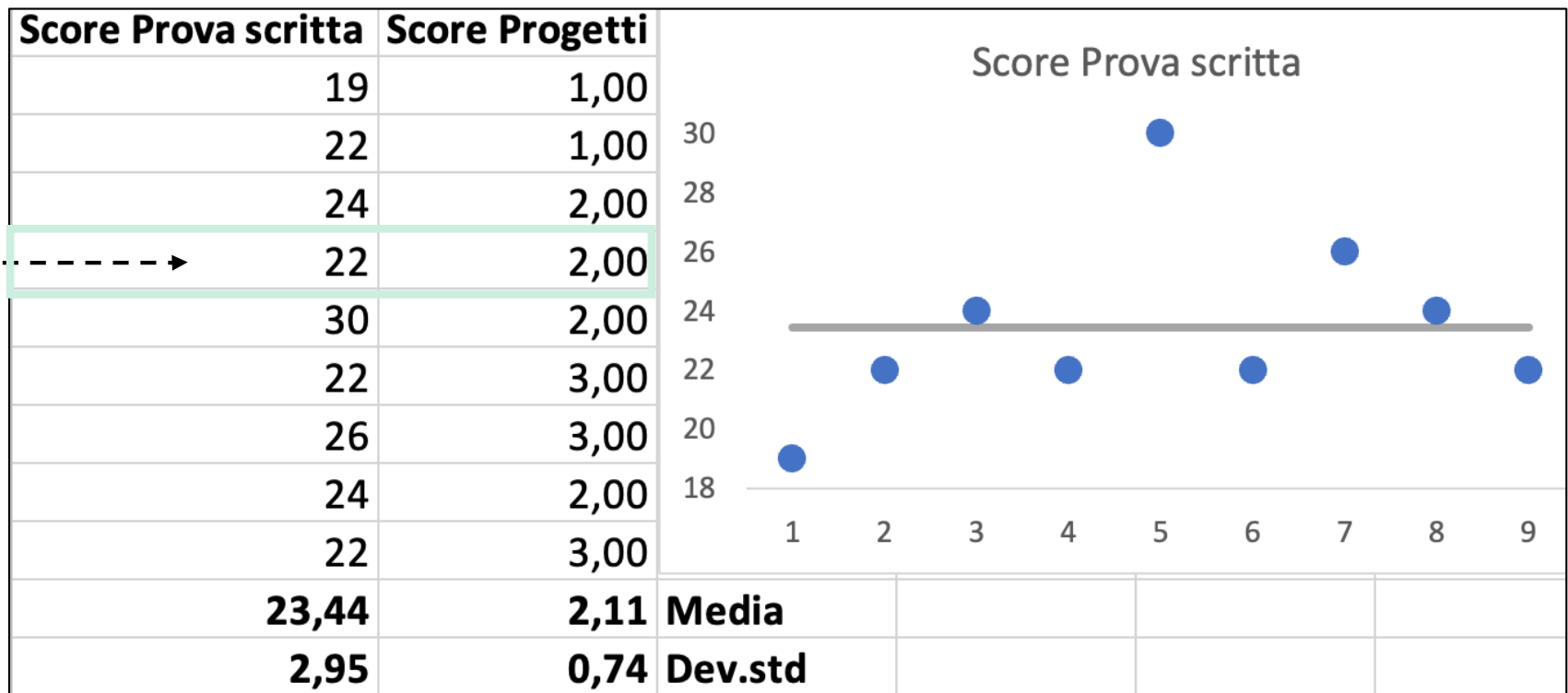
Definire un indice in base ai decessi per covid19 in:

Bulgaria, Croazia, San Marino, USA, Italia, UK, Spania, Tunisia, Francia

Ordinare poi chi, secondo voi, ha fatto “meglio” nella gestione dei decessi covid19 in questa pandemia

Sviluppare Sistemi di Scoring e Ranking

Esercizio2 di Normalizzazione con Z-score



In quale delle due prove Paolo ha performato meglio rispetto all'andamento della classe?



Tecniche Avanzate di Scoring e Ranking

Combinazioni lineari di valori normalizzati generalmente ritornano score ragionevoli.

Altre tecniche più avanzate esistono:

- Elo ranking
- Unione di ordinamenti di ranking
- Ordinamento di grafi orientati

Tecniche Avanzate di Scoring e Ranking

Como interpretare le raccolte di vittorie, se ci sono molti candidati e non tutte le coppie di giocatori si affrontano?

Non è corretto dire che quello con il maggior numero di vittorie vince, perché (a) potrebbero aver gareggiato in più confronti rispetto ad altri giocatori, e (b) potrebbero aver evitato avversari forti e battuto solo la concorrenza inferiore.

Elo-ranking inizia valutando tutti i giocatori allo stesso modo, e quindi regola in modo incrementale il punteggio di ciascun giocatore in risposta al risultato di ogni partita.



Prof. Arpad Elo

L'Elo-ranking di un giocatore è:

$$r'(A) = r(A) + k(S_A - E_A)$$

dove $r'(A)$ è l'aggiornamento del suo punteggio, $r(A)$ il suo attuale punteggio, E_A il punteggio atteso, S_A il punteggio acquisito col nuovo match e k un coeff. correttivo di variazione Elo.

Come determinare E_A ?

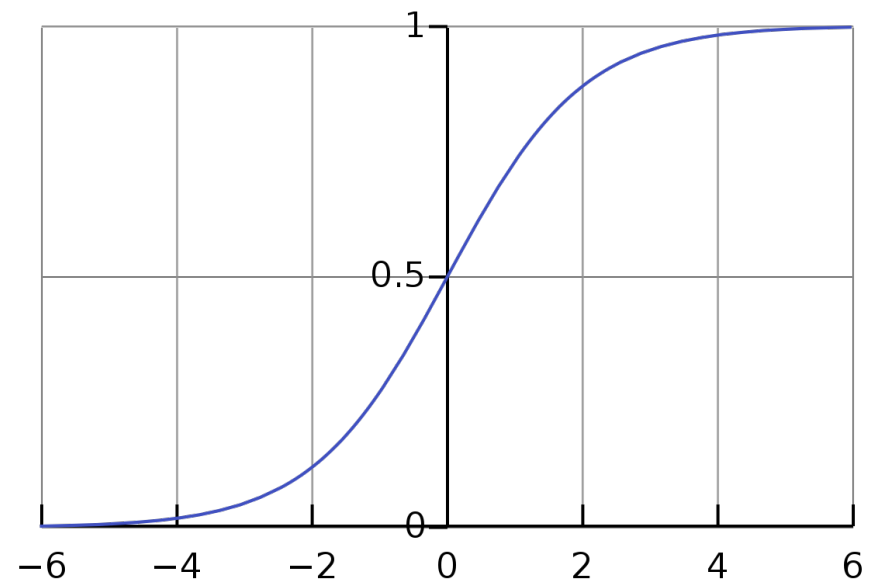
$$E_A = 1 * P_{A>B} + (-1)(1 - P_{A>B})$$

dove: $P_{A>B}$ è la probabilità che giocatore A vinca contro giocatore B. Questa probabilità ovviamente dipende dalle differenti skill tra A e B:

$$\text{cioè: } x = r(A) - r(B)$$

Ci resta ora solo da capire come determinare x e convertirla in una probabilità. Per fare questo dobbiamo introdurre la **funzione logit** che converte valori $-\infty < x < \infty$ in probabilità $0 \leq p \leq 1$

dove: $p = f(x) = \frac{1}{1+e^{-cx}}$



$f(0) = 0.5$ (i.e., quando A è forte come B, allora la prob di vincere è 0.5)
 $f(\text{inf}) = 1$ (i.e., quando A è infinitamente più forte di B, allora la sua prob di vincere è 1)
 $f(-\text{inf}) = 0$ (i.e., quando A è infinitamente più debole di B, allora la sua prob di vincere è 0)

Negli scacchi, $f(x) = E_A = \frac{1}{1+10^{(R_B-R_A)/400}}$ e $k = 32$.

Esempio di calcolo dell'Elo-ranking:

Il sottoscritto ha un ELO di 1580, e gioca 5 partite:


- perde contro un giocatore quotato 1700,
- vince con un giocatore che ha 1577 punti,
- patta con un giocatore con 1388 punti,
- vince con un giocatore con 1620 e infine
- perde con un giocatore che ha 1710 punti.


Il suo punteggio S_A è $(0 + 1 + 0.5 + 1 + 0) = 2.5$



Quanto sarà il nuovo ELO?

Tecniche Avanzate di Scoring e Ranking

L'Elo-ranking DEL giocatore di scacchi è:

Carlsen, Magnus  Age 32



Title	Grandmaster	Born	30 Nov 1990
Live Rating	2859.0		Tønsberg, Vestfold, Norway
FIDE Rating	2859		FIDE
World	#1 (active)		Wikipedia
Norway	#1 (active)		Facebook
FIDE Peak	2882 - 01 May 2014		Twitter
Rapid	2834 World #1 (active)		Instagram
Blitz	2830 World #4 (active)		Google

Source:
<http://www.tatasteelchess.com/>

Il **Metodo Borda** è utile per combinare insieme, in un unico score, diversi ranking (che possono essere calcolati con diverse funzioni e su diverse caratteristiche).

Borda assegna un costo o un peso a ciascuna delle n posizioni della permutazione. Quindi, per ciascuno degli n elementi, riassume i pesi delle sue posizioni su tutte le k classifiche di input. L'ordinamento di questi n punteggi determina la classifica finale del consenso.

La mappatura tra posizioni e costi può avvenire tramite:

1. Pesi “lineari” (i.e., assegna i punti per apparire nell' i – *esima* posizione in ogni permutazione)
2. Pesi “non lineari” (es. scelti da una distribuzione normale)



Tecniche Avanzate di Scoring e Ranking

Esempio del Metodo Borda lineare:

StudentID	Rank Esame1	Rank Esame2	Rank Esame3	Rank Esame4	Chi è stato più bravo?
#0001	D	B	B	D	
#0002	C	D	D	B	
#0003	A	C	C	C	
#0004	B	A	A	A	

Esempio del Metodo Borda lineare:

StudentID	Rank Esame1	Rank Esame2	Rank Esame3	Rank Esame4	Chi è stato più bravo?
#0001	D	B	B	D	12
#0002	C	D	D	B	13
#0003	A	C	C	C	10
#0004	B	A	A	A	5

Applichiamo i punti per apparire nell' i – esima posizione:

- a A (prima posizione): 1pt.
- a B (seconda posizione): 2pt.
- a C (terza posizione): 3pt.
- ...

Altro Esempio del **Metodo Borda lineare (inversa)**:

Prendiamo in considerazione la possibilità di determinare il vincitore di un'elezione multipartitica in cui ogni votante classifica i candidati in ordine di preferenza. (Il primo prende 4pt, il secondo 3pt, il terzo 2pt e il quarto 1pt)

Abbiamo 4 votanti che si esprimono così:

Classifica	Votante1	Votante2	Votante3	Votante4
Primo	Meloni	Salvini	Conte	Letta
Secondo	Salvini	Meloni	Letta	Conte
Terzo	Letta	Letta	Meloni	Meloni
Quarto	Conte	Conte	Salvini	Salvini

Conte: $4+3+1+1=9$
 Letta: $4+3+2+2=11$
 Meloni: $4+3+2+2=11$
 Salvini: $4+3+1+1=9$

Metodo Borda:

Pesi “lineari” funzionano bene quando abbiamo ugual confidenza tra tutte le posizioni in classifica (i.e., dare una variazione di peso 1 al primo o secondo classificato è equivalente a dare 1 a chi si colloca a metà classifica)

Ci sono scenari o domini dove invece la parametrizzazione deve essere diversa. Es. premio i primi in classifica e penalizzo gli ultimi (distribuzione non simmetrica)

Nell'esempio dell'elezione multipartitica si potrebbe pensare ad assegnare i punteggi come segue:
il primo prende 6pt, il secondo 4pt, il terzo 2pt e il quarto 1pt)

Directed Graph Orderings:

Le reti forniscono un modo alternativo per pensare a una serie di voti della forma "A è davanti a B". Possiamo costruire un grafo/rete orientato in cui esiste un vertice corrispondente a ciascuna entità e un arco orientato (A,B) per ogni voto $A > B$.

Se i voti fossero totalmente coerenti, la permutazione ottimale dei vertici violerebbe esattamente zero archi (i.e., non ci sono cicli diretti nel grafo). Un ciclo diretto come (A, C), (C, E), (E, A) rappresenta una contraddizione intrinseca a qualsiasi ordine di classificazione.

Un grafo diretto senza cicli è chiamato **Grafo Aciclico Diretto** o DAG. Tendenzialmente però, i voti non sono mai perfettamente coerenti.

Directed Graph Orderings:

Arrow's Impossibility Theorem dimostra che non c'è sempre un sistema di ranking in grado di

- essere completo: dati A e B deve sceglierne uno o dare uguale preferenza
- essere transitivo, nel senso che se $A > B$ e $B > C$ allora $A > C$
- se ogni elettore preferisce A a B, allora A vince su B
- le preferenze non possono dipendere solo da un dittatore
- la preferenza di A su B deve essere indipendente dalle preferenze degli altri candidati.

Meloni batte Salvini,
Salvini batte Letta,
ma Letta batte Meloni



Classifica	Votante1	Votante2	Votante3
Primo	Meloni	Salvini	Letta
Secondo	Salvini	Letta	Meloni
Terzo	Letta	Meloni	Salvini

Ranking Example: Who's Bigger?

Analyzed Wikipedia to extract measures of historical significance: PageRank, length, hits...

- Mapped values to normal distributions
- Use linear combination (factor analysis)
- Corrected for time by decaying modern figures.
- Separate scores for celebrity/gravitas.

