Math 187 F24

Worksheet

Inner Product Application: Ranking Baseball Player Seasons

We are to the home stretch of the baseball season! While there is still about a month left, many of the broad contours of the season are well-determined. Below is current data for 8 MLB players from www.baseball-reference.com (retrieved on 9/5/24), along with historical full-season data for three of the top 40 individual seasons ever.

Player	Age	Team	G	PA	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	WAR
Aaron Judge	32	NYY	138	618	107	159	32	1	51	124	7	0	113	150	0.323	0.455	0.702	9.5
Shohei Ohtani	29	LAD	137	630	111	158	30	7	44	99	46	4	72	142	0.290	0.375	0.613	7.0
Anthony Santander	29	BAL	135	580	82	126	23	2	39	91	2	0	48	112	0.242	0.312	0.519	3.0
Juan Soto	25	NYY	136	624	112	149	28	4	38	98	5	4	112	97	0.295	0.423	0.592	7.5
Brent Rooker	29	OAK	123	514	71	135	24	2	33	94	8	2	52	149	0.297	0.372	0.576	4.7
Pete Alonso	29	NYM	140	603	79	128	29	0	31	78	3	0	59	149	0.240	0.323	0.469	2.2
Vladimir Guerrero Jr.	25	TOR	138	603	85	175	40	1	28	92	2	2	61	84	0.328	0.400	0.564	5.6
Bryce Harper	31	PHI	123	539	73	132	34	0	26	76	5	4	66	115	0.282	0.372	0.521	4.0
1947 Jackie Robinson	28	BRO	151	701	125	175	31	5	12	48	29	11	74	36	0.297	0.383	0.427	4.1
1985 Rickey Henderson	26	NYY	143	654	146	172	28	5	24	72	80	10	99	65	0.314	0.419	0.516	9.9
2004 Barry Bonds	39	SFG	147	617	129	135	27	3	45	101	6	1	232	41	0.362	0.609	0.812	10.6

G	Games	HR	Home Runs	BA	Batting Average
PA	Plate Appearances	RBI	Runs Batted In	OBP	On-Base Percentage
R	Runs	SB	Stolen Bases	SLG	Slugging
H	Hits (singles + 2B + 3B + HR)	CS	Caught Stealing	WAR	Wins Above Replacement
2B	Doubles	BB	Base on Balls		
3B	Triples	SO	Strike Outs		

(a) With each player we can associate a "feature vector" v consisting of the numerical data we have for them. For what d are these vectors in $\mathbb{R}^d$? Should any of the available information be excluded from our feature vectors? Explain.

(b) Suppose you wanted to create a new column in the data measuring the total number of bases (say, "TNB") a player offensively generates. How could you produce this as a linear combination of other column vectors? Explain.

(c) Suppose we want to rank these individual seasons. (This might help us determine an MVP for the season, or compare individual performances from this season with well-regarded performances of the past.) One way to do this is by determining a single "score" for each player and then ranking them by score. A score can be computed by selecting a "weight vector" w and then taking an inner product $w^T v$ with a player vector v. Come up with such a weight vector and write it down explicitly. Explain your reasoning for using this weight vector. Use R to compute the scores for these 11 players. Report on the resulting ranking.

(d) Come up with another way of ranking these individual seasons. Explain your methodology, produce a weight vector w, compute scores, and rank the players.

(e) Wins Above Replacement (WAR) is a currently used metric to rate players seasonal contribution to their teams. A WAR of 3.5 would mean, for instance, that having that player on your team would yield an expected 3.5 additional wins over the season compared to if your team would have fielded a generic "replacement" player. How do your rankings systems compare to the WAR metric for these players?