Linear ranking classifier

Jan Bielecki

June 21, 2018

Przykład - Learning data

Table: Learning data - D_L

	wzrost	wiek	nr buta	płeć
1	190	25	48	M
2	171	31	41	K
3	174	40	40	M
4	179	30	45	M
5	160	20	36	K
6	180	18	44	M
7	165	18	41	M
8	172	50	40	K
9	180	32	43	K
10	155	22	32	K

t = 10 (ilość rekordów), l = 3 (atrybutów), g = 2 (klas)

Przykład - Nowa obserwacja

Table: Learning data - D_L

	wzrost	wiek	nr buta	płeć
1	190	25	48	M
2	171	31	41	K
3	174	40	40	M
4	179	30	45	M
5	160	20	36	K
6	180	18	43	M
7	165	18	41	M
8	172	50	40	K
9	175	32	43	K
10	155	28	32	K

Nowa osoba: $x_n = \{\text{wzrost: } 180, \text{ wiek: } 22, \text{ nr buta: } 44\}$

Przebieg klasyfikacji

Do kazdej z klas liczymy odleglość zdefiniowana nastepujaco:
$d_j(x_n) = \sum_{i \neq j}^{g} \sum_{l=1}^{k} \frac{1}{\pi_j} \rho(D_L^l(j,i), x_n^l) * \frac{|r(D_L(j,i)^l, D_L(j,i)^{class})|}{\sqrt{\sum_{z=1}^{k} r^2(D_L(j,i)^z, D_L(j,i)^{class})}}$,
$D_L(j,i)$ - rekordy o klasie j lub i,
π_j - prawdopodobieństwo a priori,
$r(D_L(j,i)^l, D_L(j,i)^{class})$ - współczynnik pearsona pomiedzy atrubutem l z rekordów o klasie j lub i, a klasami rekordów o klasie j lub i (przy czym zamieniamy z symboli na j = 0, i = 1).

Przebieg klasyfikacji

$$d_j(x_n) = \sum_{i \neq j}^{g} \sum_{l=1}^{k} \frac{1}{\pi_j} \rho(D_L^l(j,i), x_n^l) * \frac{|r(D_L(j,i)^l, D_L(j,i)^{class})|}{\sqrt{\sum_{z=1}^{k} r^2(D_L(j,i)^z, D_L(j,i)^{class})}},$$

$$\rho(D_L^l(j,i), x_n^l) = \frac{rank_l(D_L^l(j,i), x_n^l)}{t},$$

dla $r(D_L(j,i)^l, D_L(j,i)^{class}) > 0$,

$$\rho(D_L^l(j,i), x_n^l) = 1 - \frac{rank_l(D_L^l(j,i), x_n^l)}{t},$$

dla $r(D_L(j,i)^l, D_L(j,i)^{class}) < 0$,

$rank_l(D_L^l(j,i), x_n^l)$ - ranking atrybutu l od x_n wśród zbioru $\{D_L^l(j,i), x_n^l\}$,

Przykład - ranking, $rank_l(D_L^l(j,i), x_n^l)$

Nowa osoba: $x_n = \{\text{wzrost: } 180, \text{ wiek: } 22, \text{ nr buta: } 44\}$

$l = 1(wzrost)$:

$\{155, 160, 165, 171, 172, 175, 174, 179, \mathbf{180}, \mathbf{180}, 190\}$

$rank_{wzrost}(D_L^{wzrost}(K, M), x_n^{wzrost}) = 8.5$

$l = 2(wiek)$:

$\{18, 18, 20, \mathbf{22}, 25, 28, 30, 31, 32, 40, 50\}$

$rank_{wiek}(D_L^{wiek}(K, M), x_n^{wiek}) = 3$

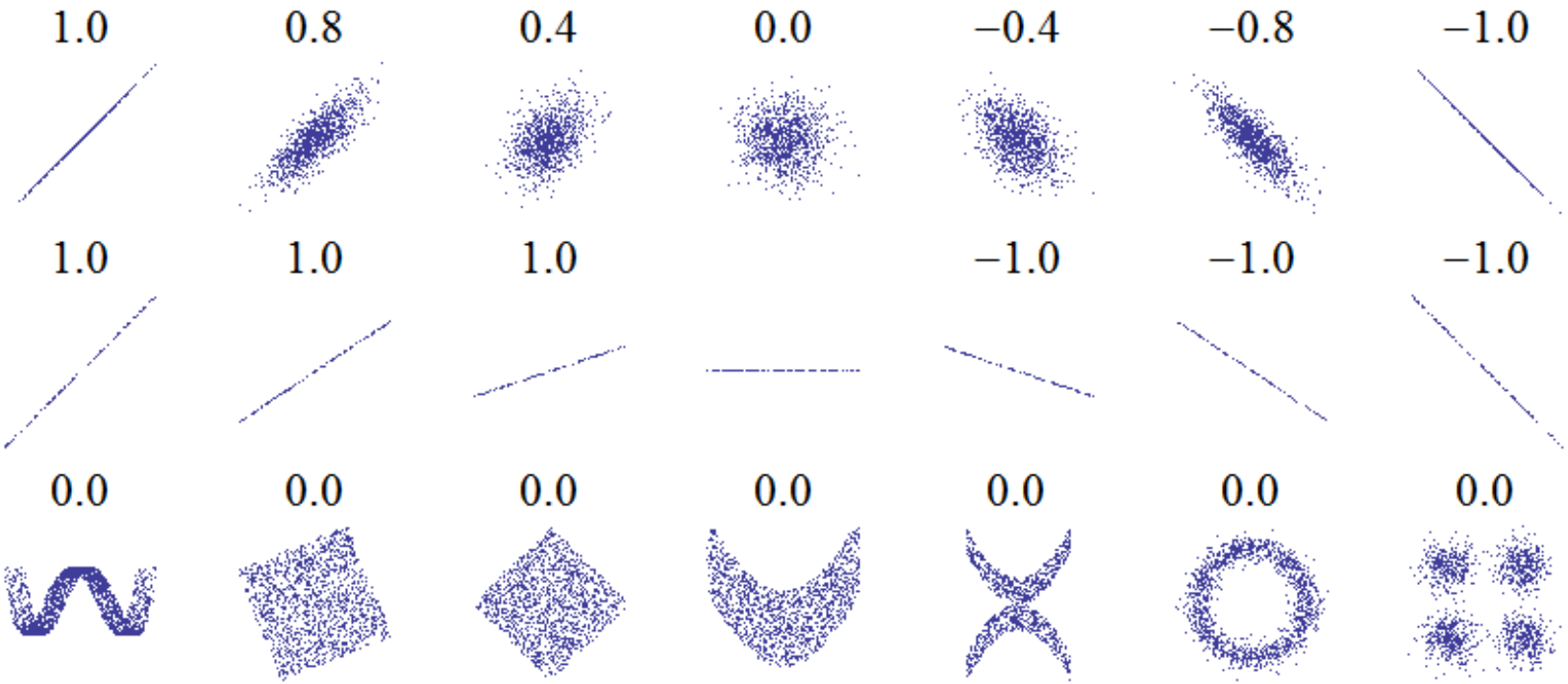
$l = 3(nrbuta)$:

$\{32, 36, 40, 40, 41, 41, 43, 43, \mathbf{44}, 45, 48\}$

$rank_{nrbuta}(D_L^{nrbuta}(K, M), x_n^{nrbuta}) = 8$

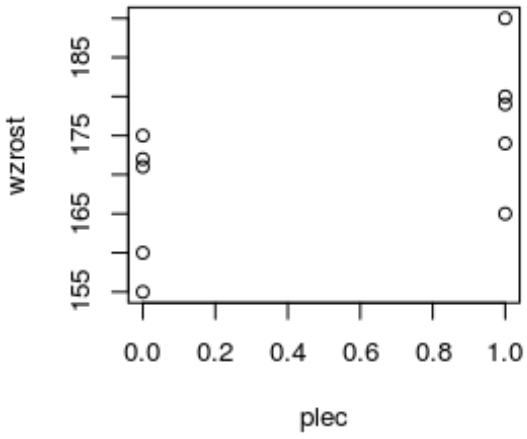
Współczynnik pearsona

$$r(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, r(X, Y) \in [-1, 1]$$

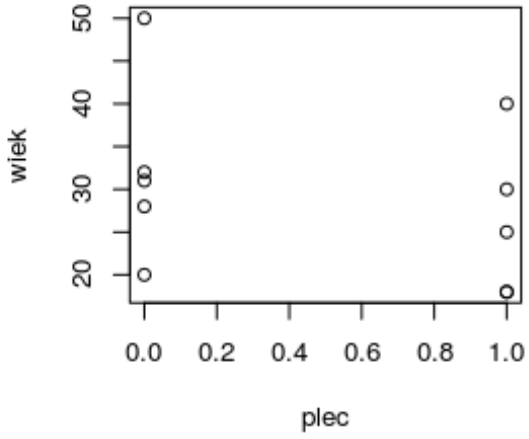


Przykład - współczynnik pearsona

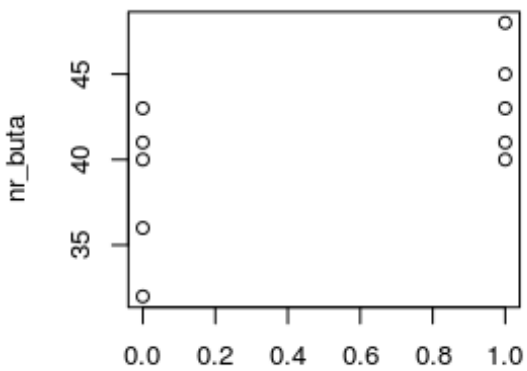
wsp. pearsona = 0.56943629237242



wsp. pearsona = -0.31352224676030



wsp. pearsona = 0.58778801321549



Przykład - Odległości od klas

$$d_j(x_n) = \sum_{i \neq j}^{g} \sum_{l=1}^{k} \frac{1}{\pi_j} \rho(D_L^l(j,i), x_n^l) * \frac{|r(D_L(j,i)^l, D_L(j,i)^{class})|}{\sqrt{\sum_{z=1}^{k} r^2(D_L(j,i)^z, D_L(j,i)^{class})}},$$

$$d_K(x_n) = \frac{1}{0.5} * \left(\frac{8.5}{10} * \frac{0.57}{1.47} + \left(1 - \frac{3}{10}\right) * \frac{0.31}{1.47} + \frac{8}{10} * \frac{0.59}{1.47}\right)$$

$$d_M(x_n) = \frac{1}{0.5} * \left(\left(1 - \frac{8.5}{10}\right) * \frac{0.57}{1.47} + \frac{3}{10} * \frac{0.31}{1.47} + \left(1 - \frac{8}{10}\right) * \frac{0.59}{1.47}\right)$$

$$d_K(x_n) = 1.60$$

$$d_M(x_n) = 1.45$$

Przykład - Klasyfikacja

Nowa osoba: $x_n = \{\text{wzrost: } 180, \text{ wiek: } 22, \text{ nr buta: } 44\}$
$d_K(x_n) = 1.60$
$d_M(x_n) = 1.45$
$d_{min}(x_n) = d_M(x_n)$
Nowa osobe klasyfikujemy jako M.

Ocena klasyfikatora - Breast Cancer Wisconsin (Diagnostic) Data Set

`https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`
k = 32 (atrybutów), g = 2 (klasy),
Learning data = 450 rekordów,
Test data = 119 rekordów,
Skuteczność na podstawie 1000 losowych prób (119 000 test rekordów).
Sprawdzono skuteczność LRC (Linear ranking classificator) oraz k_{nn}(maksymalizacja skuteczności przez znormalizowanie danych i dla najlepszego k).
$eff_{LRC} = \mathbf{0.9314}$
$eff_{knn} = \mathbf{0.9615}$

Ocena klasyfikatora - IRIS Data Set

k = 7 (atrybutów), g = 4 (klasy),
Learning data = 80 rekordów,
Test data = 20 rekordów,
Skuteczność na podstawie 1000 losowych prób (20 000 test rekordów).
Sprawdzono skuteczność LRC (Linear ranking classificator) oraz k_{nn}(dla najlepszego k).
$eff_{LRC} = \mathbf{0.8690}$
$eff_{knn} = \mathbf{0.9337}$

Klasyfikator LRC - zalety

- Addytywność atrybutów (atrybuty traktowane sa jako niezależne, im wiecej tym teoretycznie lepiej)
- Odpornosc na duże warjancje atrybutów próby uczacej (ranking niweluje wplyw warjancji)

Klasyfikator LRC - wady

- Wspolczynnik pearsona nie jest najlepsza waga dla odleglości od klasy na podstawie rankingu (lepszej nie znaleziono)
- Wyraznie slabsza skuteczność niz prosty k_{nn}