

Analiza statystyczna trudności tekstu

Łukasz Dębowski
ldebowsk@ipipan.waw.pl



Problem badawczy

Chcielibyśmy mieć wzór matematyczny, ...

- ▶ ... który dla dowolnego tekstu ...
- ▶ ... na podstawie pewnych statystyk ...
- ▶ ... obliczanych za pomocą programu komputerowego ...
- ▶ ... przewidywałby dostatecznie dobrze, ...
- ▶ ... jak trudny jest on do zrozumienia ...
- ▶ ... dla przeciętnego czytelnika.

Zastosowanie: aplikacja JASNOPIS.

Intencja: Chcemy zaproponować coś lepszego niż wzór Pisarka:

$$T = \frac{\sqrt{T_s^2 + T_w^2}}{2}. \quad (1)$$

Dane

- ▶ Przygotowaliśmy 35 tekstów o zróżnicowanej trudności (7 klas trudności a priori po 5 tekstów).
- ▶ Przeprowadziliśmy badania psychologiczne, jak dobrze teksty te są rozumiane przez potencjalnych czytelników (próba ok. 1000 osób, test cloze + test pytań otwartych)
- ▶ Wyodrębniliśmy kilkadziesiąt zmiennych lingwistycznych przypuszczalnie skorelowanych z trudnością tekstu (długość zdania, długość słowa, procent rzeczowników itp.)
- ▶ Za pomocą programu komputerowego wyznaczyliśmy wartości liczbowe tych zmiennych dla danych 35 tekstów.

Mając te dane chcielibyśmy wyznaczyć wzór matematyczny przewidujący trudność tekstu.

Tabela z danymi (35 wierszy, 69 kolumn)

plik	punkty cloze	punkty otwarte	liczba akapitów	liczba zdań	liczba słów	...
1/1-1.txt	25,48	4,40	13	31	285	...
1/1-2.txt	30,80	4,42	11	29	283	...
1/1-3.txt	28,25	4,33	16	39	300	...
1/1-4.txt	29,90	4,00	5	27	284	...
1/1-5.txt	25,31	4,40	9	33	319	...
2/2-1.txt	24,70	4,44	4	24	294	...
2/2-2.txt	27,96	4,35	8	24	318	...
2/2-3.txt	26,94	4,45	7	28	283	...
2/2-4.txt	23,67	4,54	9	23	289	...
2/2-5.txt	24,73	4,67	17	29	283	...
...

Metoda najmniejszych kwadratów

Oznaczenia:

- ▶ Y_i — punkty cloze/otwarte dla i -tego tekstu
- ▶ X_{ij} — j -ta zmienna objaśniająca dla i -tego tekstu
- ▶ N — liczba tekstów
- ▶ M — liczba zmiennych objaśniających

Szukamy wzoru postaci:

$$Y_i = \sum_{j=1}^M X_{ij} A_j + \text{szum losowy} \quad (2)$$

minimalizując sumę kwadratów błędów:

$$\sum_{i=1}^N \left(Y_i - \sum_{j=1}^M X_{ij} A_j \right)^2 = \min . \quad (3)$$

Ograniczenia metody najmniejszych kwadratów

- ▶ Metodę najmniejszych kwadratów można zastosować wyłącznie, gdy liczba tekstów M jest znacznie większa niż liczba zmiennych objaśniających N ...
- ▶ ... w przeciwnym wypadku zachodzi przeuczenie, czyli dopasowujemy się do szumu losowego w danych i wzór (2) nie przewiduje trudności tekstów spoza próby uczącej.
- ▶ W naszym przypadku liczba tekstów M jest mniejsza niż liczba zmiennych objaśniających N .

Rozwiązania:

- ▶ Zmniejszyć liczbę zmiennych objaśniających (jak we wzorze Pisarka).
- ▶ Zastosować regresję liniową z regularyzacją.

Dwa rodzaje regularyzacji

- ▶ Regresja lasso:

Minimalizujemy sumę kwadratów błędów z karą liniową:

$$\sum_{i=1}^N \left(Y_i - \sum_{j=1}^M X_{ij} A_j \right)^2 + \alpha \sum_{j=1}^M |A_j| = \min . \quad (4)$$

- ▶ Regresja ridge (grzbietowa):

Minimalizujemy sumę kwadratów błędów z karą kwadratową:

$$\sum_{i=1}^N \left(Y_i - \sum_{j=1}^M X_{ij} A_j \right)^2 + \alpha \sum_{j=1}^M A_j^2 = \min . \quad (5)$$

Trzy inne modele

- ▶ Regresja liniowa metodą najmniejszych kwadratów ze dwiema zmiennymi jak we wzorze Pisarka ($M = 2$) (średnia długość zdania, procent słów dłuższych niż 3 sylaby).
- ▶ Średnia ważona (komitet) z regresji liniowych metodą najmniejszych kwadratów dla trzech zmiennych ($M = 3$) (średnia długość zdania, procent słów dłuższych niż 3 sylaby oraz dowolna inna zmienna).
- ▶ Baseline (model odniesienia): Trudność dowolnego tekstu jest stała (szacowana jako średnia z próby uczącej).

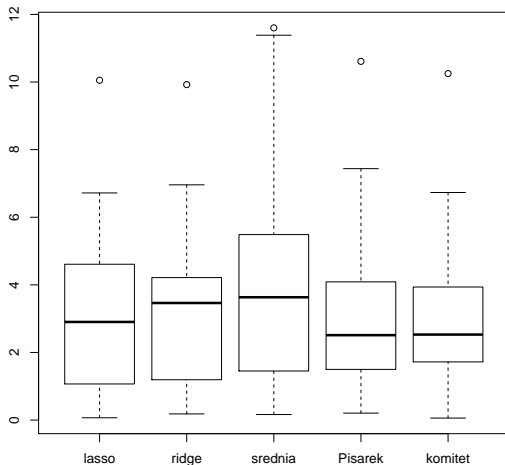
Który z tych modeli jest najlepszy?

Pewnym sposobem sprawdzenia tego, jest krosvalidacja:

- ▶ Wyjmujemy z próby uczącej jeden tekst, dopasowujemy model do pozostałych tekstów i sprawdzamy, jak dobrze ów model przewiduje zmienną objaśnianą dla wyjątku tekstu.
- ▶ Błąd (odchylenie modelu od wartości przewidywanej) mierzymy dla każdego tekstu w próbie uczącej i sporządzamy wykres pudełkowy obrazujący przeciętną wartość i rozrzut tego błędu.
- ▶ Szukamy metody, dla której błąd jest najmniejszy.

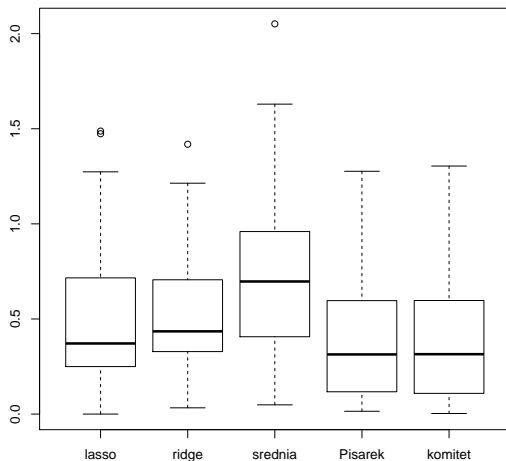
Wykres pudełkowy błędu

Zmienna przewidywana: punkty cloze



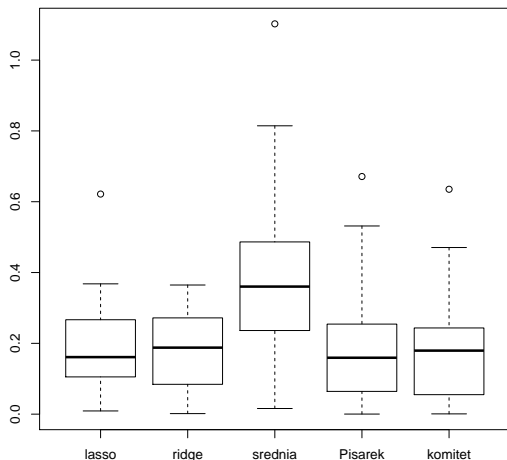
Wykres pudełkowy błędu

Zmienna przewidywana: punkty otwarte

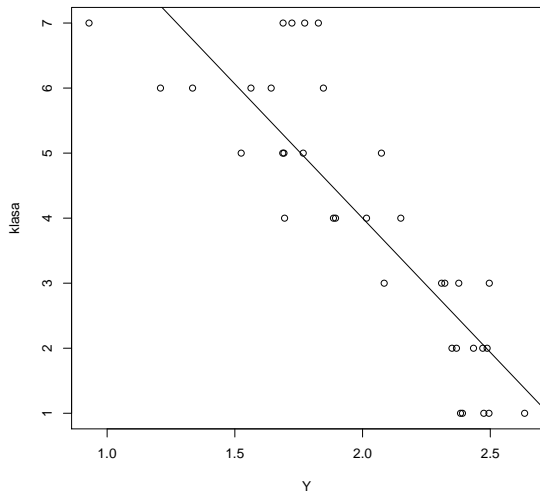


Wykres pudełkowy błędu

Przewidywane: średnia ważona punktów cloze i punktów otwartych



Jak zmienna przewidywana koreluje z klasą tekstu?



Wzór na trudność tekstu

$$\text{Klasa} = 12.25 - 4.12 * \text{Ridge}$$

$$\begin{aligned} \text{Ridge} = & 2.7 - 0.0034 * (\text{średnia długość zdania}) \\ & - 0.0027 * (\text{procent słów trudnych}) \\ & + 0.0026 * (\text{procent rzeczowników}) \\ & - 0.0044 * (\text{procent rzeczowników trudnych}) \\ & + 0.0037 * (\text{procent czasowników}) \\ & + 0.0053 * (\text{procent czasowników trudnych}) \\ & - 0.00043 * (\text{średnia długość akapitu}) \\ & - 0.013 * (\text{średnia długość łańcucha dopełniaczowego}) \\ & - 0.0033 * (\text{procent dopełniaczy}) \\ & - 0.0019 * (\text{procent rzeczowników na '-ość'}) \\ & + \dots \end{aligned}$$

Podsumowanie

- ▶ Skonstruowaliśmy wzór na trudność tekstu, którego maksymalny błąd predykcji jest trzy razy mniejszy niż modelu odniesienia (w którym trudność tekstu nie zależy od tekstu), a dwa razy mniejszy niż wzoru Pisarka.
- ▶ Wzór jest zbyt skomplikowany by liczyć go ręcznie, ale jest prosty do zaimplementowania w programie komputerowym.