

Uczenie maszynowe przy użyciu danych meteorologicznych

Praca dyplomowa licencjacka

Jan Solarz,
Promotor: dr Radosław Michalski

Kierunek studiów: Matematyka i Statystyka
Specjalność: Statystyka i analiza danych
Wydział Matematyki
Politechnika Wrocławska

14 września 2021

1 Wprowadzenie do tematyki pracy

- Problem z dostępnością wody i zmieniającego się klimatu

2 Cele, przyjęte założenia i przypuszczenia

3 Modele wielokrotnej regresji liniowej.

- Konstrukcja modelu i jego dopasowanie
- Odlegość Coock'a

4 Drzewa decyzje CRAT

- Budowa drzew i Indeks Giniego

- Demonstracja przykładu

5 Naiwny klasyfikator Bayesowski

- Konstrukcja algorytmu

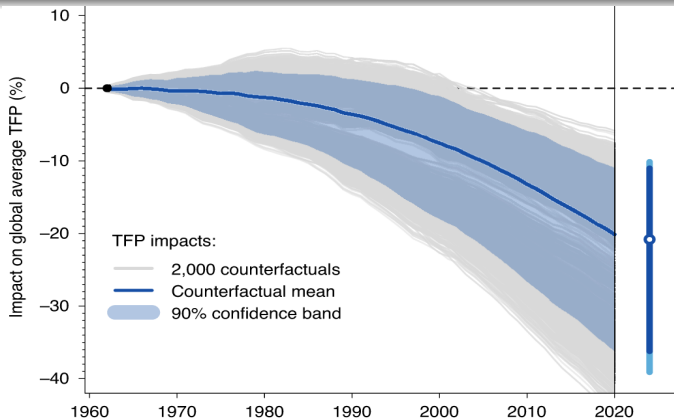
6 Część eksperymentalna

- Prezentacja wyników modeli regresji
- Porównanie modeli regresji
- Porównanie metod klasyfikacji
- Obserwacje, podsumowanie i wnioski

7 Bibliografia

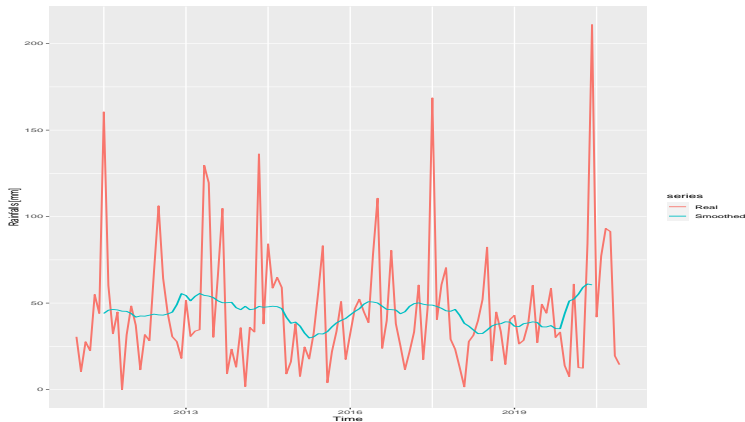
Meteorologia i predykcja w rolnictwie

- W porównaniu z danymi z 1961 r. rzeczywista wydajność, jeśli chodzi o uprawę roślin spadła o 21%.
- Ze względu na postęp technologiczny wydajność pracy rolników wzrasta, dynamicznie postępująca erozja gleby w wyniku prowadzonych przez człowieka procesów powoduje że zbliżamy się do punktu krytycznego.
- Szacuje się że od roku 1961 wzrost użycia nawozów azotowych wzrósł o 800%, a nawadniania roślin o 100% na całym świecie.
- ONZ przewiduje, że do 2050 r. liczba ludzi żyjących na ziemi wzrośnie do ponad 9 miliardów, a zatem produkcja żywności będzie musiała wzrosnąć o około 70%.

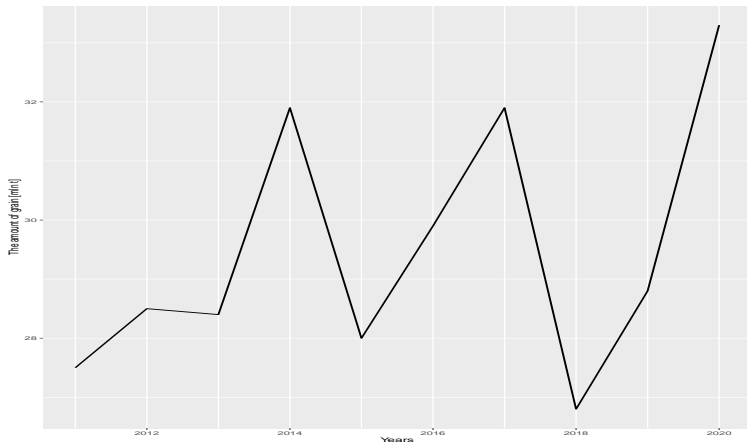


Rysunek 1: Wykres przedstawiający wpływ antropogenicznych zmian klimatu (ACC) na całkowitą produktywność czynników (TFP) w latach 1961-2020 na świecie.

Problem z dostępnością wody i zmieniającego się klimatu



Rysunek 2: Wykres opadów w latach 2011-2020.



Rysunek 3: Wykres zebranych zbóż w latach 2011-2020.

- W latach 2017-2019 występowały susze w Polsce, przełomowy rok 2020 przyniósł o 16% więcej plonów w porównaniu do roku poprzedniego.
- Dla porównania wielkość plonu z niektórych gatunków roślin uprawianych w Polsce, które były nawadniane o około 20-30% więcej w porównaniu z roślinami bez nawadniania, wzrosła o ponad 50%.
- W Polsce ograniczenia plonów potencjalnych roślin wywołane niedoborami wody sięgają nawet 60%.
- Średnie zużycie wody na rolnictwo w Europie wynosi 24%, są jednak regiony na poziomie 60%-80%.

Cele, przyjęte założenia i przypuszczenia

- Konstruujemy zestaw danych składający się z parametrów meteorologicznych z okresu historycznego we Wrocławiu z interwałem miesięcznym.
- Budujemy modele i badamy ich skuteczność przeprowadzając predykcje opadów deszczu przyjmując różne podejścia.
- Staramy się omówić trzy metody uczenia maszynowego, zwracając uwagę na najbardziej istotne elementy ich działania.
- Istnieje obawa braku znaczących parametrów danych oraz użycia zbyt podstawowych technik uczenia maszynowego podejścia do tej konkretnej problematyki. Problemem może być zbyt wysoka losowość danych i brak istotnych zależności.

Konstrukcja modelu regresji i jego dopasowanie

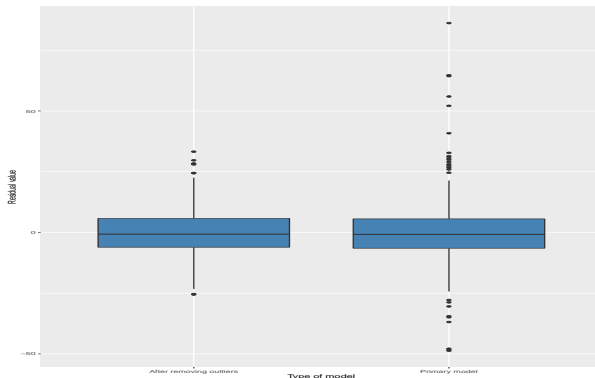
Funkcja regresji wielokrotnej

W modelu zakładamy, że zależność między zmienną objaśnianą Y , a zmiennymi objaśniającymi X_1, \dots, X_{p-1} ma postać:

$$r(x) := E(Y|X=x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_{p-1} + \varepsilon.$$

- Wektor $\beta = (\beta_0, \dots, \beta_{p-1})^T$ jest nieznany i chcemy go estymować za pomocą metody najmniejszych kwadratów.
- Błędy losowe $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ nie są obserwowane i zakłada się $\mathbb{E}(\varepsilon) = 0$.
- Do porównywania ze sobą modeli zagnieżdżonych, przeprowadza się testowanie hipotez. Licząc wartości statystyk F korzysta się m. in. z rozkładu Snedecora.

Odlegość Cook'a



Rysunek 4: Wykres przedstawiający rozrzut residuów przed i po usunięciu wartości odstających.

Budowa drzew i Indeks Giniego

Reguła klasyfikacyjna w drzewach decyzyjnych:

$$\hat{p}_{m,c} = \frac{1}{n_m} \sum_{x_i \in M} \mathbb{1}_{\{y_i=c\}} = \frac{n_{m,c}}{n_m}$$

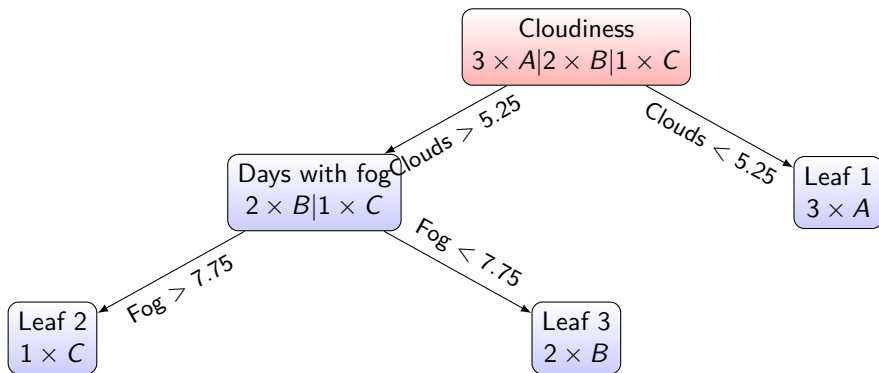
$$\hat{k}(m) = \operatorname{argmax}_{c \in C} \hat{p}_{m,c}$$

Miara nieczystości- Index Giniego:

$$Gini(Leaf_k) = 1 - \sum_{class=1}^C (\text{probability}(\text{class} | Leaf_k))^2 = 1 - \sum_{c=1}^C (p(c | L_k))^2$$

$$Gini(Node_j) = \sum_K \left(\frac{\text{count}(L_k)}{\text{count}(L_1, \dots, L_k)} \right) Gini(L_k)$$

Demonstracja przykładu



Rysunek 5: Finalny efekt skonstruowanego drzewa z *Fog* jako korzeniem.

Konstrukcja algorytmu NB

Niech $y_k \in 1, \dots, K$ oznacza klasę. Definiujemy Klasyfikator Bayesowski :

$$d_B(x) = \operatorname{argmax}_{k \in \{1, \dots, K\}} P(y_k | X_i)$$

- $\hat{f}_k(X_i) = \hat{f}_k(x_1, \dots, x_n) = \prod_{i=1}^n \hat{f}_{k,i}(x_i)$ oznacza gęstość zmiennej x_i dla k klasy.
- $\hat{\pi}_k = \frac{n_k}{n}$ odnosi się do prawdopodobieństwa *a priori* klasy k .

Naiwny Klasyfikator Bayesowski: $\hat{d}_{NB}(x) = \operatorname{argmax}_{k \in \{1, \dots, K\}} \hat{\pi}_k \hat{f}_k(x)$

Prezentacja wyników modeli regresji

Tablica 1: Tabela opisująca zmienne Modelu 1

Variable	Coefficient	Std. Error	Test value	p-value
Intercept	212.002	190.367	1.114	0.26602
Maximum daily rainfall	2.11061	0.07702	27.403	< 2e-16
Days with rain	2.55217	0.20410	12.504	< 2e-16
Days with snow	1.60260	0.27053	5.924	6.24e-09
Days with storm	1.15066	0.38132	3.018	0.00269
Days with haze	0.28065	0.12002	2.338	0.01981
Month	-0.58612	0.22496	-2.605	0.00948
Average water vapor pressure [hPa]	0.74970	0.37618	1.993	0.04687
Average relative humidity [%]	0.41369	0.17060	2.425	0.01570
Days with fog	-0.38440	0.24202	-1.588	0.11292
Average pressure at the station level [hPa]	-0.26770	0.18936	-1.414	0.15813

Porównanie modeli regresji

Tablica 2: Tabela opisująca zmienne Modelu 4

Variable	Coefficient	Std. Error	Test value	p-value
Intercept	124.58988	51.41567	2.423	0.015779
Month	-1.81796	0.59679	-3.046	0.002454
Average relative humidity [%]	-2.21836	0.67504	-3.286	0.001095
Days with haze	1.45387	0.33933	4.285	2.24e-05
Average cloud cover	12.93757	3.65072	3.544	0.000436
Days with low snow blizzard	-4.42766	1.70331	-2.599	0.009645
Average water vapor pressure [hPa]	-4.86565	3.13191	-1.554	0.120991
Sum of monthly rainfalls	0.17027	0.05886	2.893	0.004005
Max temperature	11.11933	3.82510	2.907	0.003831
Average temperature	-10.85719	4.98134	-2.180	0.029808

Results of the simulation were obtained via R.

Prezentacja wyników modeli regresji

Tablica 3: Zestawienie parametrów skontruowanych modeli regresji

Type of model	R^2	Adj R^2	F-statistic and DT	Pearson Correlation	MSE
Model 1	0.8399	0.8363	236.6 on 10 and 451	0.9376	202.47
Model 1 after Cook	0.87	0.8675	319 on 9 and 428	0.9378	220.45
Model 2	0.4903	0.4778	39.35 on 11 and 450	0.768	725.89
Model 3	0.2506	0.2356	16.76 on 9 and 451	0.5132	934.42
Model 4	0.2568	0.2419	17.24 on 9 and 449	0.4302	733.14

Porównanie metod klasyfikacji

Tablica 4: Zestawienie Błędów Klasyfikacji badanych metod

Proper model and sample	NV	Shift NV	DT	Shift DT	Shift Regr
Control model, training sample	0.5195	0.5390	0.2359	0.2239	-
Control model, test sample	0.6111	0.5657	0.3081	0.3706	-
Regression model, training sample	0.3593	0.3788	0.2121	0.2283	-
Regression model, test sample	0.4293	0.4040	0.3081	0.3706	0.6231






NV-Naive Bayes Classifier prediction for present time;
Shift NV-Naive Bayes Classifier prediction for 3 months;
DT-Decision Tree prediction for present time;
Shift DT-Decision Tree prediction for 3 months;
Shift Reg-Regression Model 4 prediction for 3 months

Obserwacje, podsumowanie i wnioski

- Z *Tablicy 1* widoczny jest największy wpływ zmiennych *Maximum daily rainfall* i *Days with rain* na zmienną zależną *Sum of monthly rainfall*. Są za to odpowiedzialne wszystkie 3 parametry- Standard Error, Test Value i p-value. P-value na poziomie $<2e-16$.
- Predykcja opadów deszczu na jeden lub trzy miesiące w przód jest niemożliwa w przyjętej koncepcji pracy. *Model 3* i *Model 4* posiadają bardzo bliskie sobie parametry. Wniosek jest dość oczywisty- nie ma znaczenia jak daleko chcielibyśmy prognozować opady w interwale miesięcznym. Nie występują korelacje liniowe.

Obserwacje, podsumowanie i wnioski

- Zdecydowanie *Model regresyjny* jest bardziej skuteczny w NB, co świadczy o tym że nie zawsze większa ilość predyktorów oznacza wyższą dokładność.
- Błąd klasyfikacji w DT w *Modelu regresyjnym* również jest niższy od błędów w *Modelu kontrolnym*. Jedynie dla zbioru treningowego w Shift DT otrzymujemy nieco lepszą skuteczność w *Modelu regresyjnym*.
- Drzewa decyzyjne prezentują najlepszą skuteczność ze wszystkich metod (0.3081 and 0.3706).
- Modele Regresji liniowej uzyskały najwyższe błędy (0.6231).- może to być spowodowane inną koncepcją.

-  Agresti, A. Foundations of linear and generalized linear models. John Wiley Sons, 2015
-  Breiman, L., Friedman, J., Olshen, R., Stone, C. Classification and regression trees. wadsworth int. Group 37, 15 (1984)
-  Cure, J. D., Acock, B. Crop responses to carbon dioxide doubling. Agricultural and forest meteorology 38, 1-3 (1986)
-  Johnson, R. A., Wichern, D. W., et al. Applied multivariate statistical analysis, vol. 6. Pearson London, UK:, 2014.
-  Mbow, C., Rosenzweig, C. E., Barioni, L. G., Benton, T. G., Herrero, M., Krishnapillai, M., Ruane, A. C., Liwenga. Food security.