Podstawy Przetwarzania Danych

Wykład 9: Miary jakości

dr inż. Marcin Luckner
mluckner@mini.pw.edu.pl

Wydział Matematyki i Nauk Informacyjnych

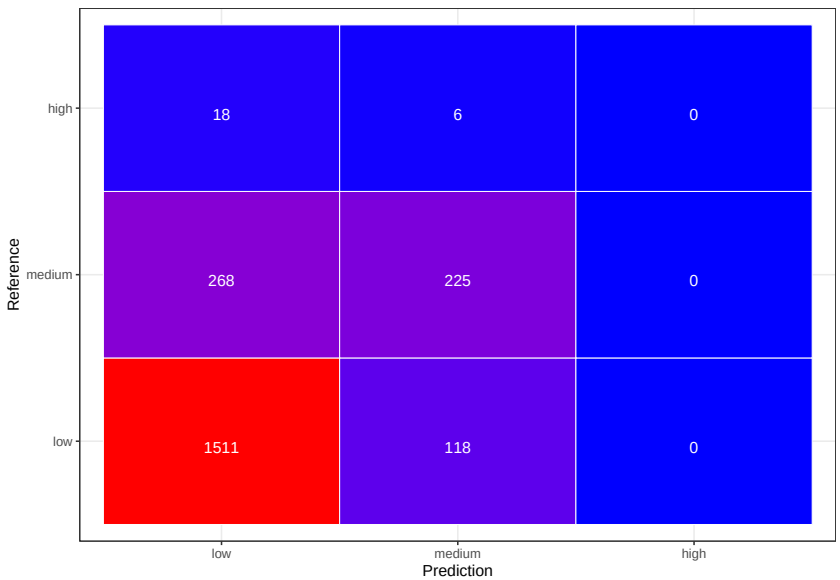
Wersja 1.1
5 marca 2021

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
Politechnika Warszawska
Unia Europejska Europejski Fundusz Społeczny

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie klasyfikacji



Rysunek 1: Wyniki klasyfikacji

- Dany jest zbiór trenujący $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$.
- Zbiór składa się z par $(\mathbf{x_i}, y_i)$ wektora cech opisujących $\mathbf{x}_i$, i cechy opisywanej y_i.
- W przypadku klasyfikacji $y_i \in Y$ jest cechą dyskretną z ograniczonego zbioru klas.
- Zadanie klasyfikacji polega na znalezieniu klasyfikatora $h : \mathbf{X} \rightarrow Y$ który przydziela obiektowi $\mathbf{x} \in \mathbf{X}$ klasę $y \in Y$

Metody klasyfikacji

- Drzewa decyzyjne
- Klasyfikatory Bayesowskie
- Sieci Neuronowe
- Analiza statystyczna
- Metaheurystyki (np. algorytmy genetyczne)
- Zbiory przybliżone
- k-NN – k-najbliższe sąsiedztwo

Macierz pomyłek

- Macierz pomyłek zawiera liczbę elementów z każdej klasy, przypisanej do każdej z klas.
- Jest wyliczana na podstawie predykcji i docelowych wartości.
- W przypadku zadania binarnego macierz pomyłek przybiera formę

	Condition P	Condition N
Predicted P	T(rue)P(ositive)	F(alse)P(ositive)
Predicted N	F(alse)N(egative)	T(rue)N(egative)

Statystyki

- Pola macierzy pomyłek służą do zdefiniowania miar statystycznych.
- Skuteczność - procent poprawnie rozpoznanych elementów

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- Czułość - zdolność rozpoznawania pozytywnych przypadków

$$Sensitivity = \frac{TP}{TP + FN}$$

- Specyficzność - zdolność niepopełniania błędów

$$Specificity = \frac{TN}{TN + FP}$$

- Precyzja - procent poprawnych rozpoznań

$$Precision = \frac{TP}{TP + FP}$$

Cechy statystyk

- Skuteczność może być mylącą miarą jeżeli liczność klas jest silnie zróżnicowana.
- Zazwyczaj zwiększanie Czułości powoduje spadek Specyficzności i na odwrót.
- Powstały miary pozwalające na balansowanie tych wskaźników F-measure i AUC.

F-measure

- F-measure (inaczej F1) jest miarą bilansującą Czułość i Precyzję.
- Jest to ich średnia harmoniczna

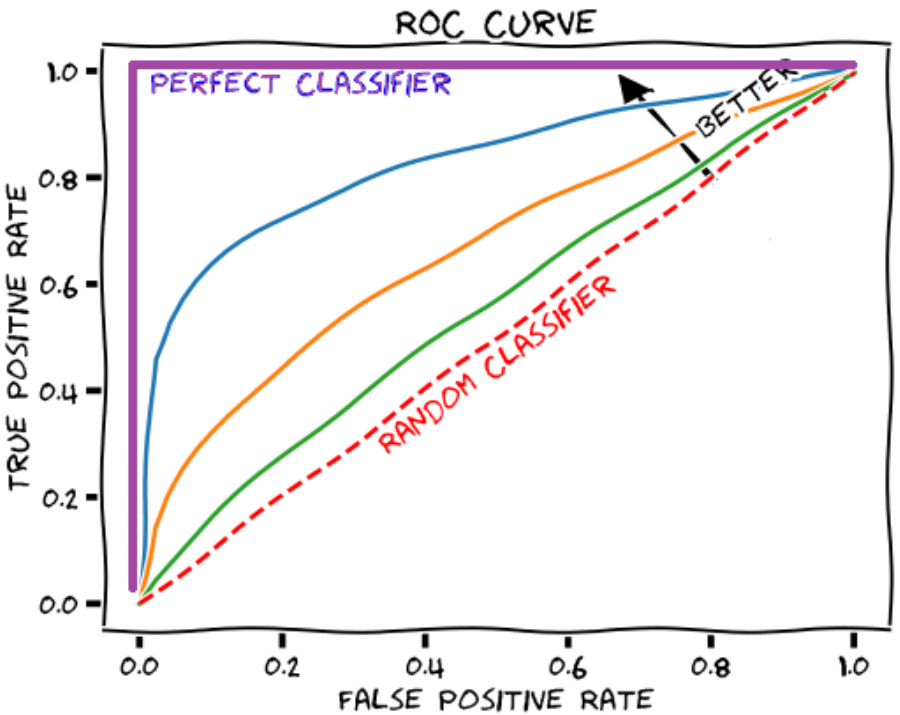
$$F1 = 2 * \frac{Sensitivity * Precision}{Sensitivity + Precision} = \frac{2TP}{2TP + FP + FN}$$

- Miara premiuje zrównoważone wartości obu cech.

Krzywa ROC i AUC

- Krzywa ROC (Receiver Operating Characteristic) jest krzywą opartą na kilku parametryzowanych testach klasyfikatora binarnego.
- Rzędne punktów krzywej określa Czułość a odcięte określa 1-Specyficzność danego testu.
- Im lepszy klasyfikator tym bardziej wykładniczy charakter krzywej.
- Do porównywania klasyfikatorów używa się pola pod krzywą AUC (Area Under Curve).

Interpretacja krzywej ROC



Rysunek 2: Krzywa ROC [Draelos, 2019]

- Krzywe ROC powinny się zawierać pomiędzy idealnym klasyfikatorem, a losową klasyfikacją, choć mogą przekraczać linię tej ostatniej.
- Im wyżej położona linia, tym lepszy klasyfikator, ale linie mogą się przecinać.
- Pole pod wykresem (miara AUC) rozstrzyga jednoznacznie który klasyfikator jest lepszy.

Statystyki dla klasyfikacji wieloklasowej

- Podane statystyki mają zastosowanie tylko w przypadku klasyfikacji binarnej.
- W przypadku klasyfikacji wieloklasowej zazwyczaj wylicza się statystyki osobno dla każdej klasy, traktując wszystkie pozostałe klasy jako meta-klasę *inne*.
- Tak wyliczone statystyki można uśrednić dla wszystkich klas stosując mikro lub makro uśrednianie.

Makro uśrednianie

- W wypadku makro uśredniania, jeżeli mamy wyliczone miary m_i dla klas $i \in 1 \ldots n$, takie jak Precyzja, Czułość itp. możemy wyliczyć ich średnią wartość jako

$$\bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i.$$

- Przykładowo średnią Precyzję wyliczymy jako

$$\bar{Precision}_M = \frac{1}{n} \sum_{i=1}^{n} Precision_i.$$

Mikro uśrednianie

- W wypadku mikro uśredniania bazujemy na wartościach z macierzy pomyłek, które służą do wyliczenia miary m_i czyli TP_i, FP_i, FN_i i TN_i.
- Przykładowo średnią Precyzję wyliczymy jako

$$\bar{Precision}_{\mu} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i}.$$

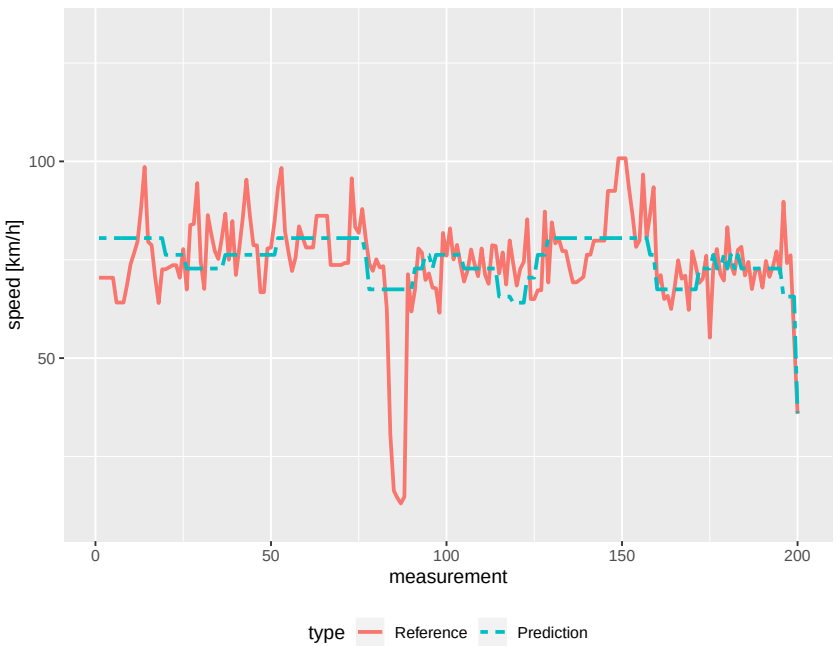
Porównanie mikro i makro uśredniania

Micro-averaged scores			
Micro	Precision	Recall	F1
k-NN	0.7787	0.7787	0.7787
CBC	0.7271	0.7271	0.7271
Micro-averaged scores			
Macro	Precision	Recall	F1
k-NN	0.6806	0.5479	0.5715
CBC	0.6274	0.7436	0.6482

Rysunek 3: Porównanie mikro i makro uśredniania [Oliveira and Filho, 2017]

- Porównano mikro i makro uśrednianie w zadaniu klasyfikacji 21 typów dokumentów [Oliveira and Filho, 2017].
- Różnice wynikają z nierównomiernego rozkładu dokumentów między klasami.
- Makro uśrednianie traktuje wszystkie klasy tak samo, a mikro uśrednianie faworyzuje większe klasy [Sokolova and Lapalme, 2009].

Zadanie regresji



Rysunek 4: Wyniki regresji

- Zadanie regresji polega na modelowaniu ciągłej zmiennej opisywanej Y poprzez cechy opisujące $\mathbf{X}$
- W regresji parametrycznej zakładamy, że istnieje pewien model, którego parametry mamy odnaleźć.
- W regresji nieparametrycznej nie zakładamy określonego modelu i estymujemy funkcję na podstawie serii obserwacji.

Regresja parametryczna

- Ogólna postać modelu

$$Y = f(\mathbf{X}, \beta) + \epsilon$$

gdzie

- X wektor zmiennych objaśniających,
- Y zmienna objaśniana,
- β wektor współczynników regresji
- ϵ błąd losowy

Metody regresji parametrycznej

- Regresja liniowa
- Regresja nieliniowa
- Uogólnione modele liniowe (GLM)
- Regresja logistyczna

Regresja nieparametryczna

- Postać modelu nie jest jednoznacznie określona
  - nie znamy postaci analitycznej funkcji składowych modelu,
  - liczba funkcji składowych modelu nie jest ustalona,
  - na etapie budowy modelu nie jest jednoznacznie określony zestaw zmiennych w modelu końcowym.
- Wymogi wobec zmiennych objaśniających stawiane modelom nieparametrycznym są niższe. Nie muszą mieć one rozkładu normalnego i być niewspółliniowe.
- Ogólnie modele nieparametryczne są elastyczniejsze i mają szersze zastosowania.

Metody regresji nieparametrycznej

- metody rekurencyjnego podziału (Rpart)
- metody zestawu drzew regresyjnych (Bagging, Random Forest)
- metody wektorów nośnych (SVM)
- sieci neuronowe (Nnet)

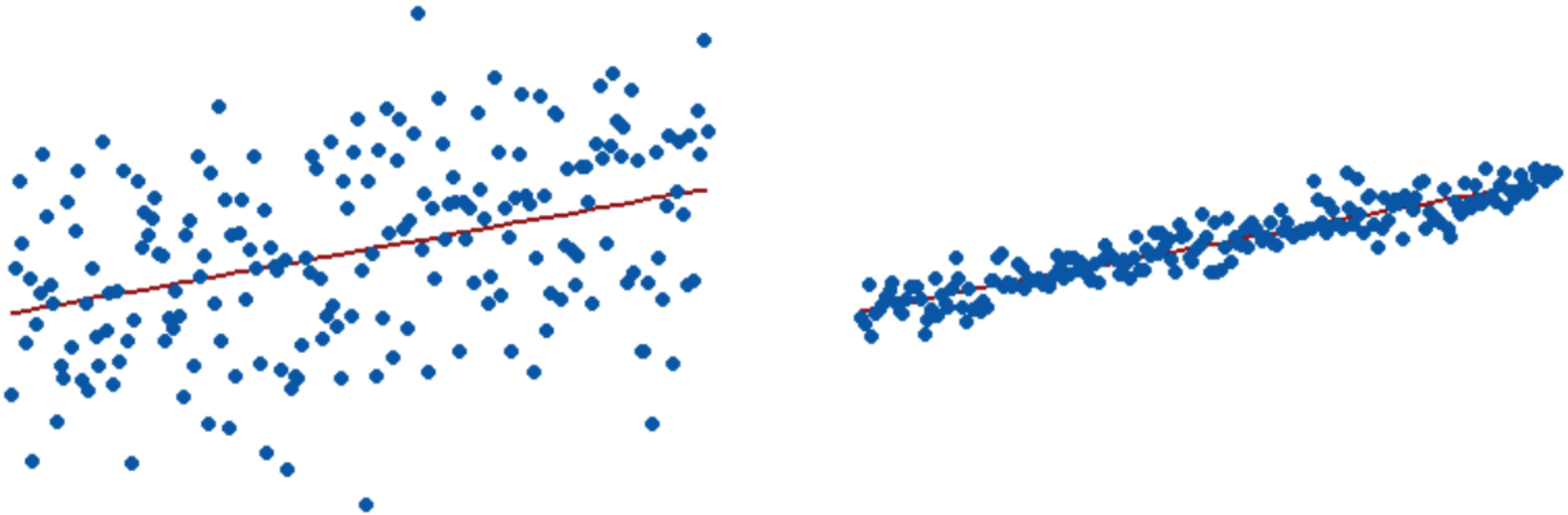
Miara R^2

R-squared R^2 reprezentuje kwadrat korelacji między predykcjami, a oczekiwanymi wynikami.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o}_i)^2}$$

- Miara R^2 prezentuje procent wytłumaczenia obserwowanej wariancji przez model.
- Ogólnie im wyższa, tym lepsza, ale nie jest to miara intuicyjna.

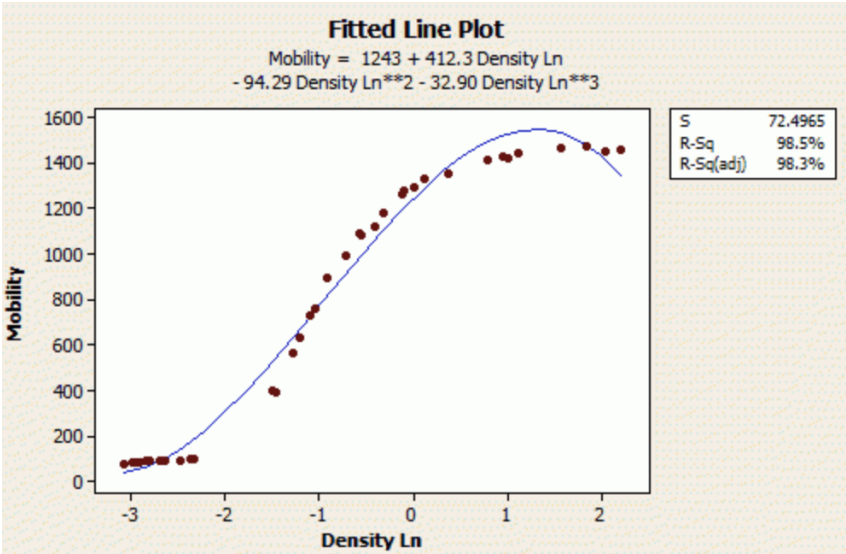
Interpretacja R^2



Rysunek 5: Zestawienie dwóch modeli regresyjnych [Frost, 2019]

- Wartość R^2 dla pierwszego modelu wynosi 0.15.
- Wartość R^2 dla drugiego modelu wynosi 0.85.
- W tym przypadku wartość R^2 jest dobrze interpretowalna.

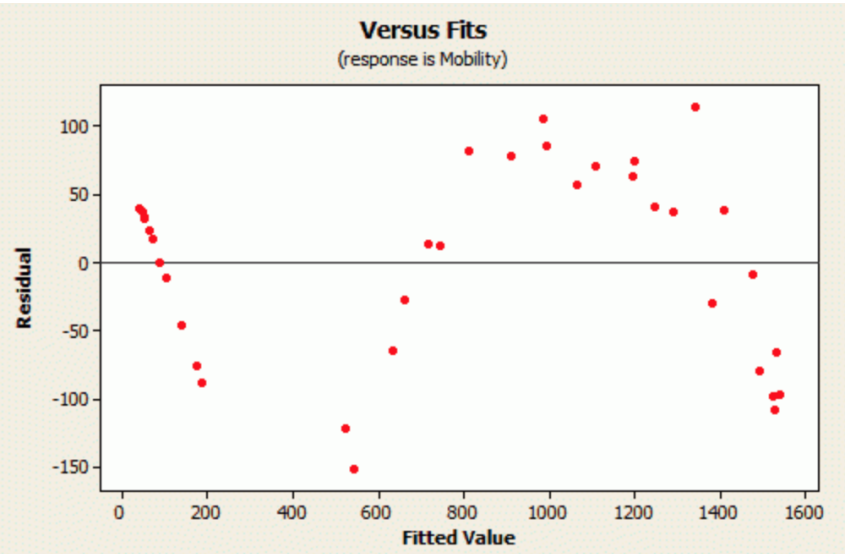
Ocena modelu z wysokim R^2



Rysunek 6: Model zależności między mobilnością, a gęstością elektronu [Frost, 2019]

- Przedstawiony model zależności między mobilnością, a gęstością elektronu ma bardzo wysoki współczynnik $R^2 = 0.985$ [Frost, 2019].
- Wydaje się, że model jest niemal idealny, ale można zaobserwować systematyczność w niedoszacowaniu i przeszacowaniu obserwacji.

Analiza rezyduł



Rysunek 7: Analiza rezyduł [Frost, 2019]

- Analizując rezydua (różnice między wartością obserwowaną, a estymowaną) modelu, widzimy, że błędy mają charakter systematyczny.
- Oznacza to, że model jest tendencyjny.
- Analiza rezyduł, w celu potwierdzenia ich losowego rozkładu, jest obowiązkowym działaniem przy interpretacji R^2.

Miary RMSE i MAE

Root Mean Squared Error błąd średnio-kwadratowy między predykcjami, a oczekiwanymi wynikami.

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(o_i - p_i)^2}$$

Mean Absolute Error średni błąd bezwzględny między predykcjami, a oczekiwanymi wynikami.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|o_i - p_i|$$

Zależności między RMSE i MAE

- Wynik *RMSE* będzie zawsze większy lub równy *MAE*.
- Wynik *RMSE* jest zawsze mniejszy lub równy $MAE \times \sqrt{n}$, gdzie n jest liczbą obserwacji.
- *RMSE* uwypukla błędy grube.
- *MAE* jest łatwiejsze w interpretacji.

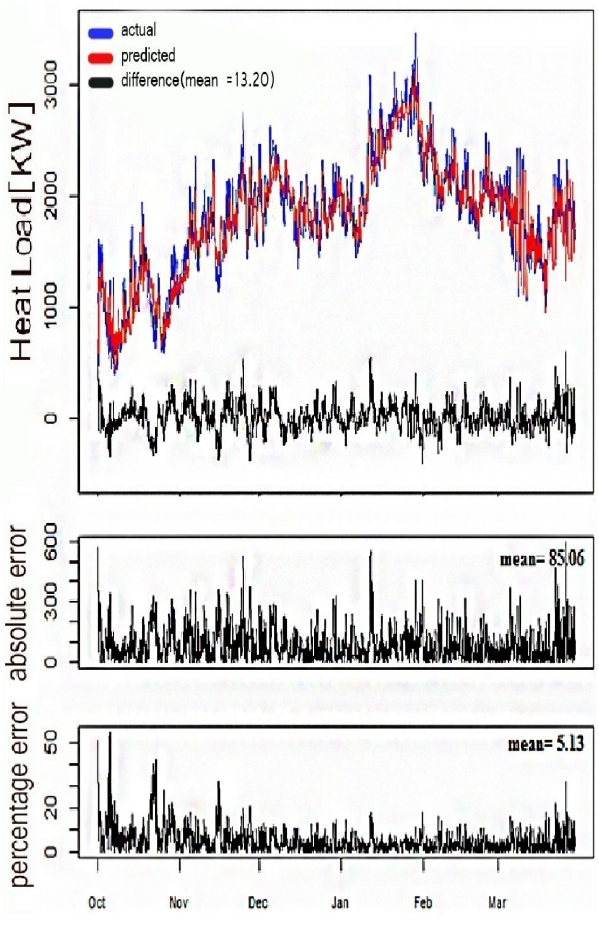
Miara procentowa

Mean absolute percentage error średni procentowy błąd bezwzględny wyliczany jako procentowa odchyłka od wartości oczekiwanej

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{|o_i - p_i|}{|o_i|}.$$

Miara jest bardzo podatna na zmiany wielkości obserwowanej.

MAPE w predykcji zapotrzebowania ciepła



Rysunek 8: Analiza zapotrzebowania ciepła [Provatas, 2014]

- Do oceny algorytmu estymacji zapotrzebowania na ciepło posłużono się miarą MAPE [Provatas, 2014].
- Jakie są zalety i wady takiego podejścia?
- Miara procentowa jest łatwa w analizie i wprost przekłada się na straty związane z błędną prognozą.
- Miara staje się niestabilna gdy zapotrzebowanie spadnie blisko zera.

Test Wilcoxona

- Test Wilcoxona dla par obserwacji pozwala porównać dwie równoliczne próbki dające się połączyć w pary.
- Używany do porównywania danych zebranych przed i po eksperymencie w celu zbadania, czy nastąpiła istotna statystycznie zmiana.
- Może być też użyty do porównania wyników uzyskanych różnymi metodami uczenia maszynowego lub dla różnych parametrów algorytmu.

Założenia testu

- Rozważamy sytuację tej samej populacji testowanej w różnych warunkach C_1 i C_2.
- Jeśli w C_2 wystąpi poprawa, wówczas większość wyników zarejestrowanych w C_2 będzie lepsza od tych zarejestrowanych w C_1, a te, które są gorsze, nie będą o wiele gorsze.
- Wyniki dla warunków C_1 i C_2 muszą być porównywane w odpowiadających sobie parach.
- Przykładowe zastosowania
  - błędy regresji uzyskane przez dwa różne algorytmy dla tych samych danych wejściowych.,
  - miary jakości dwóch różnych klasyfikatorów uzyskiwane na tych samych zbiorach danych.

Przygotowanie testu

- Załóżmy, że mamy dwa klasyfikatory f_1 i f_2 których jakość mierzymy miarą $m()$.
- Dla każdego przebiegu klasyfikacji $i \in \{1, 1, \ldots, n\}$ wyliczamy różnicę w zmierzonej jakości klasyfikacji $d_i = m_i(f_2) - m_i(f_1)$ gdzie m_i jest jakością uzyskaną w przebiegu i.
- Wartości $|d_i|$ porządkujemy rosnąco i przypisujemy im rangi odpowiadające kolejności.
  - Takim samym wartością $|d_i|$ przypisujemy średnią wagę.
- Wyliczamy sumy rang dla każdego klasyfikatora:

$$W_{s1} = \sum_{i=1}^{n} I(d_i > 0) \times rank(d_i),$$

$$W_{s2} = \sum_{i=1}^{n} I(d_i < 0) \times rank(d_i).$$

Rozstrzyganie remisów

- Przypadki $d_i = 0$ nie wpływają na W_{s1} i W_{s2}.
- Możemy je zignorować.
  - Jeżeli jest r takich wartości musimy wyliczyć nową liczbę par $n = n - r$.
- Możemy też rozdzielić przypadki po równo między wagi.
- Jeżeli liczba r jest nieparzysta odrzucamy jeden przypadek.
- Dla pozostałych wypadków wyliczamy zmodyfikowane sumy

$$W_{s1} = \sum_{i=1}^{n} I(d_i > 0) \times rank(d_i) + \frac{1}{2} \sum_{i=1}^{n} I(d_i = 0) \times rank(d_i),$$

$$W_{s2} = \sum_{i=1}^{n} I(d_i < 0) \times rank(d_i) + \frac{1}{2} \sum_{i=1}^{n} I(d_i = 0) \times rank(d_i).$$

Statystyka testu

- Test Wilcoxona przypomina t-test i sprawdza czy rozkłady mają tę samą medianę.
- Statystyka testu to $T_w = min(W_{s1}, W_{s2})$.
- Dla $n \leqslant 25$ wartość krytyczna testu jest odczytywana z tabeli.
- Dla większych n dystrybucja T_w jest wyliczana ze statystyki

$$Z_w = \frac{T_w - \mu_{T_w}}{\sigma_{T_w}}$$

gdzie

$$\mu_{T_w} = \frac{n(n+1)}{4}, \sigma_{T_w} = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

- Następnie sprawdzamy w tabeli rozkładu normalnego czy można odrzucić hipotezę zerową.
- W obu przypadkach H_0 odrzucamy jeżeli T_w jest mniejsze niż wartość dla danego n zapisana w tablicy.

Porównanie skuteczności dwóch klasyfikatorów

Domain no.	NB accuracy	SVM accuracy	NB–SVM	|NB–SVM|	Ranks (|NB–SVM|)	± Ranks (|NB–SVM|)
1	0.9643	0.9944	−0.0301	0.0301	3	−3
2	0.7342	0.8134	−0.0792	0.0792	6	−6
3	0.7230	0.9151	−0.1921	0.1921	8	−8
4	0.7170	0.6616	+0.0554	0.0554	5	+5
5	0.7167	0.7167	0	0	Remove	Remove
6	0.7436	0.7708	−0.0272	0.0272	2	−2
7	0.7063	0.6221	+0.0842	0.0842	7	+7
8	0.8321	0.8063	+0.0258	0.0258	1	+1
9	0.9822	0.9358	+0.0464	0.0464	4	+4
10	0.6962	0.9990	−0.3028	0.3028	9	−9

Rysunek 9: Porównanie jakości klasyfikacji dla SVM i NB [Japkowicz and Shah, 2011]

- Porównano wyniki klasyfikacji (skuteczność) dla dwóch klasyfikatorów Support Vector Machine (SVM) i Naive Bayes (NB) [Japkowicz and Shah, 2011].
- Wyliczona suma rang dla SVM wyniosła 28, a dla NB 17.
- Czy metoda SVM jest lepsza niż NB?

Analiza statystyki T_w

P	5	2.5	1	0.5	0.1
$n = 5$	0	–	–	–	–
6	2	0	–	–	–
7	3	2	0	–	–
8	5	3	1	0	–
9	8	5	3	1	–

Rysunek 10: Fragment tablicy testu Wilcoxona [Japkowicz and Shah, 2011]

- $T_w = \min(17, 28) = 17$,
- $n = 10 - 1 = 9$.
- Dla poziomu istotności $p = 0.05$ i $n = 9$ wartość z tablicy wynosi 8.
- Statystyka nie jest mniejsza niż wartość w tabeli, więc wygrane słabszego z klasyfikatorów nie są wystarczająco małe, aby uznać statystyczną przewagę drugiego.
- Zatem, nie stwierdzono, że SVM jest lepsze niż NB.

Bibliografia I

[Draelos, 2019] Draelos, R. (2019).

Measuring performance: Auc (auroc).

[Frost, 2019] Frost, J. (2019).

Regression Analysis, An intuitive guide for using and interpreting linear models.

Statistics By Jim Publishing.

[Japkowicz and Shah, 2011] Japkowicz, N. and Shah, M. (2011).

Evaluating Learning Algorithms: A Classification Perspective.

Cambridge University Press, New York, NY, USA.

[Oliveira and Filho, 2017] Oliveira, E. and Filho, D. B. (2017).

Automatic classification of journalistic documents on the Internet.

Transinformacao, 29(3):245–255.

Bibliografia II

[Provatas, 2014] Provatas, S. (2014).

An online machine learning algorithm for heat load forecasting in district heating systems.

[Sokolova and Lapalme, 2009] Sokolova, M. and Lapalme, G. (2009).

A systematic analysis of performance measures for classification tasks.

Information Processing and Management, 45(4):427–437.