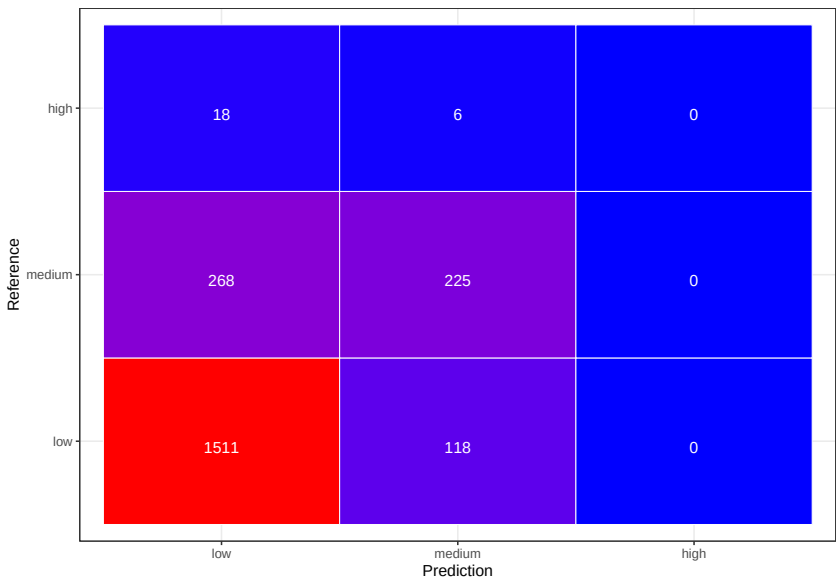
Metody analizy danych

Wprowadzenie do uczenia maszynowego i sztucznej inteligencji

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

Wersja 1
18 listopada 2021

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

	Obserwacja P	Obserwacja N
Predykcja P	T(rue)P(ositive)	F(alse)P(ositive)
Predykcja 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

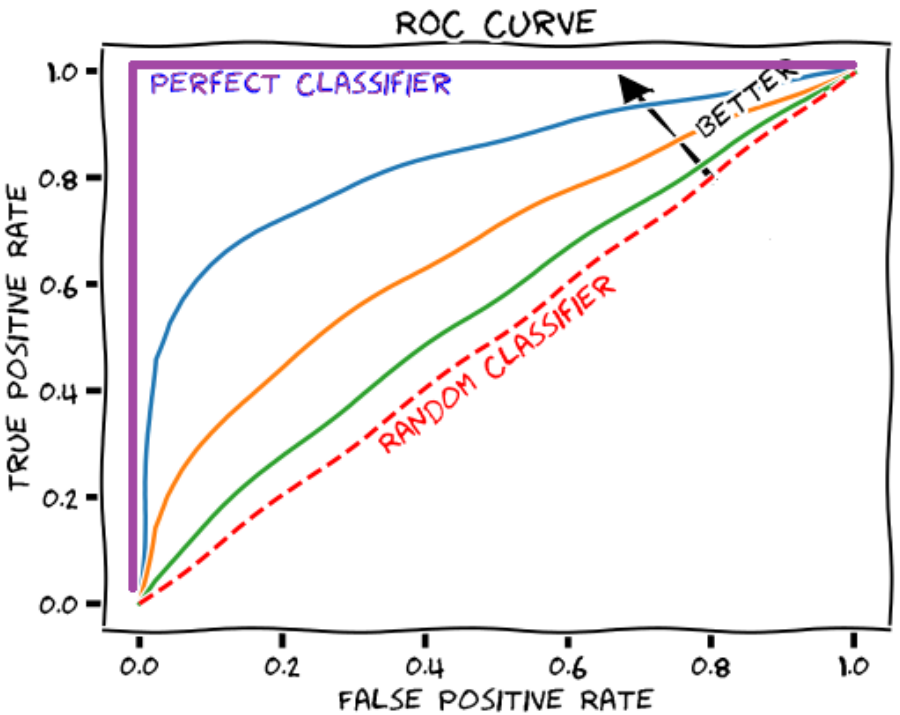
$$F1 = 2 * \frac{Sensitivity * Precision}{Sensitivity + Precision} = \frac{2TP}{2TP + FP + FN}$$

- 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.

Klasyfikacja zbiorów rozmytych

- Możemy rozważać przynależność klasyfikowanej obserwacji do wielu klas.
- W takim wypadku obliczamy prawdopodobieństwo przynależności obiektu do każdej z n rozpoznawanych klas $[p_1, p_2, \ldots, p_n]$.
- Klasyfikację rozmytą daje się łatwo sprowadzić do do klasyfikacji zero-jedynkowej przypisując 1 klasie osiągającej najwyższą wartość p_i:

$$y_j : j = argmax_{i \in \{1,2,\ldots,n\}}(p_i).$$

- Można także wykorzystać wyliczone prawdopodobieństwa do oceny jakości klasyfikacji.

Miara Log-loss

- Miara Log-loss określa jak blisko właściwej klasyfikacji jest wyznaczone prawdopodobieństwo.
- Zakładamy, że zadanie dotyczy klasyfikacji binarnej, dwóch klas oznaczonych jako 0 i 1.
- Im bardziej przewidywane prawdopodobieństwo różni się od właściwej wartości, tym wyższa jest wartość log-loss.
- Wartość dla i-tej obserwacji opisujemy wzorem:

$$Logloss_i = -[y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

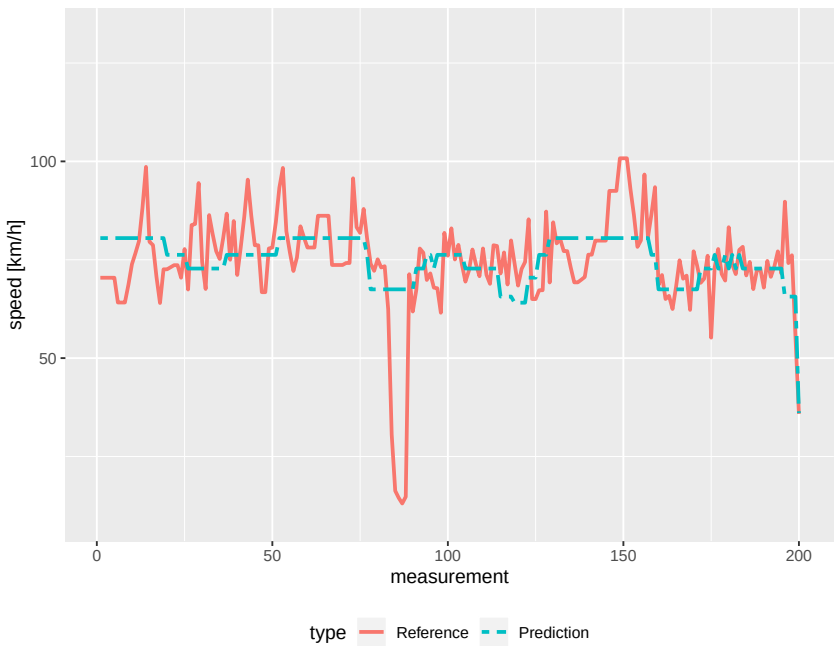
Ocena modelu miarą Log-loss

- Ocena modelu klasyfikacji przeprowadzona na N przykładach jest wyliczana jako uśrednienie wartości $Logloss_i$ dla wszystkich przykładów:

$$Logloss = \frac{1}{N} \sum_{i=1}^{N} Logloss_i.$$

- Im niższa wartość *Logloss* uzyskana dla modelu, tym działa on lepiej. Najlepszy model uzyska wartość 0.
- Miara *Logloss* jest o tyle lepsza od miary *Accuracy*, że pozwala różnicować wyniki modeli, które poprawnie rozpoznały taką samą liczbę przypadków. Nie pozwala jednak rozróżnić jakości klasyfikacji poszczególnych klas.

Zadanie regresji



Rysunek 3: 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)

Miary jakości regresji

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

$$R^2 = 1 - \frac{\sum_1^n (o_i - p_i)^2}{\sum_1^n (o_i - \bar{o}_i)^2}$$

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

$$RMSE = \frac{1}{n}\sqrt{\sum_1^n (o_i - p_i)^2}$$

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

$$MAE = \frac{1}{n}\sum_1^n |o_i - p_i|$$

Miara procentowa

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

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

Miara jest bardzo podatna na zmiany wielkości obserwowanej.

Zbiór testowy

- Niech $\hat{d}(\mathbf{x}; \mathcal{L}_n)$ oznacza klasyfikator skonstruowany za pomocą próby uczącej $\mathcal{L}_n$.
- Załóżmy, że błąd klasyfikacji będziemy określać jako prawdopodobieństwo zaistnienia błędnej klasyfikacji.

$$\hat{e} \equiv e(\hat{d}) = \Pr\left(e(\mathbf{X}) \neq Y | \mathcal{L}_n\right).$$

- Jeżeli istnieje *niezależna* próba testowa $\mathcal{T}_m = \{(\mathbf{X}_1^t, Y_1^t), \ldots, (\mathbf{X}_m^t, Y_m^t)\}$ możemy estymować błąd klasyfikatora jako

$$\hat{e}_T = \frac{1}{m} \sum_{j=1}^{m} I\left(\hat{d}\left(\mathbf{X}_j^t; \mathcal{L}_n\right) \neq Y_j^t\right).$$

Alternatywne metody testowania

- Estymacja błędu na podstawie zbioru testowego $\mathcal{T}_m$ niezależnego od próbki uczącej $\mathcal{L}_n$ jest najlepszym możliwym podejściem.
- Jednakże czasami, ze względu na rozmiar danych lub ich specyficzny charakter, nie dysponujemy niezależnym zbiorem testowym.
- Alternatywne metody testowania [Krzyśko et al., 2008]
  - Metoda ponownego podstawienia.
  - Metoda sprawdzenia krzyżowego.
  - Metoda jacknife.
  - Metoda prób bootstrapowych.

Metoda ponownego podstawiania

- Naturalną oceną poziomu błędu jest wartość estymatora ponownego podstawienia (redystrybucji) *resubstitution estimator*.
- Wyliczamy błąd dla zbioru uczącego, który służy też za zbiór testowy

$$\hat{e}_R = \frac{1}{n} \sum_{j=1}^{n} I\left(\hat{d}\left(\mathbf{X}_j; \mathcal{L}_n\right) \neq Y_j\right).$$

- Ponieważ próba ucząca jest też próbą testową estymator ten jest obciążony i zaniża rzeczywistą wartość błędu.
- Metoda może też prowadzić do przeuczenia klasyfikatora, czyli do utraty zdolności generalizacji reguł klasyfikacyjnych.

Podział próby

- Możemy podzielić próbę $\mathcal{L}_n$ na dwa podzbiory uczący i testowy.
- Metoda podziału próby (*holdout method*) powoduje jednak, że klasyfikator $\hat{d}$ będzie uczony tylko na części danych i może skutkować zaniżeniem estymacji błędu.
- Rozwiązaniem jest zastosowanie sprawdzenia krzyżowego (*cross-validation*), który stosuje wielokrotny podział próby do estymacji błędu.

Metoda sprawdzania krzyżowego

- Oznaczmy przez $\mathcal{L}_n^{(-j)}$ próbę uczącą $\mathcal{L}_n$ z której usunięto obserwację $\mathbf{Z}_j = (\mathbf{X}_j; Y_j)$.
- Klasyfikator tworzymy na zbiorze $\mathcal{L}_n^{(-j)}$ a następnie testuje na jednej obserwacji $\mathbf{Z}_j$.
  - Z powyższego powodu metodę nazywa się też *leave-one-out* LOO.
- Operację powtarza się n razy, dla każdej obserwacji $\mathbf{Z}_j$ z osobna.
- Estymator przybiera postać

$$\hat{e}_{CV} = \frac{1}{n} \sum_{j=1}^{n} I\left(\hat{d}\left(\mathbf{X}_j; \mathcal{L}_n^{(-j)}\right) \neq Y_j\right).$$

Ograniczenia sprawdzania krzyżowego

- Ze względu na testowanie budowanych klasyfikatorów na pojedynczej obserwacji, powstały estymator ma dużą wariancję.
- Ponadto wymaga konstrukcji n klasyfikatorów co może być zbyt kosztownym podejściem przy analizie większych zbiorów danych lub przy stosowaniu długotrwałych metod uczących.
- Rozwiązaniem problemu jest v-krokowa metoda sprawdzania krzyżowego *v-fold cross-validation method*.

Metoda v-krokowego sprawdzania krzyżowego

- Próba ucząca $\mathcal{L}_n$ jest dzielona na v losowych próbek.
- $v-1$ próbek tworzy próbę uczącą, a jedna próbę testową.
- Operację budowy klasyfikatora powtarza się v razy.
- Estymator przybiera postać

$$\hat{e}_{vCV} = \frac{1}{n}\sum_{i=1}^{v}\sum_{j=1}^{n} I\left(\mathcal{Z}_j \in \tilde{\mathcal{L}}_n^{(i)}\right) I\left(\hat{d}\left(\mathbf{X}_j; \tilde{\mathcal{L}}_n^{(-i)}\right) \neq Y_j\right).$$

gdzie

- $\tilde{\mathcal{L}}_n^{(1)}, \ldots, \tilde{\mathcal{L}}_n^{(v)}$ jest losowym podzbiorem próby $\mathcal{L}_n$ na równoliczne zbiory.
- $\tilde{\mathcal{L}}_n^{(-i)} = \mathcal{L}_n \setminus \tilde{\mathcal{L}}_n^{(i)}$ dla $i = 1, 2, \ldots, v$

Omówienie v-krokowego sprawdzania krzyżowego

- Metoda v-krokowego sprawdzania krzyżowego daje mniejsze obciążenie błędu niż metoda podziału na próby i wymaga mniejszej liczby klasyfikatorów niż metoda sprawdzania krzyżowego o ile $v < n$.
- W zagadnieniu estymacji aktualnego poziomu błędu zalecane jest dobranie wartości $v = 10$ [Webb, 2003].
- W celu praktycznego wykorzystania oceny modeli wybieramy spośród stworzonych klasyfikatorów ten o najmniejszym błędzie.

Metoda prób bootstrapowych

- Próbą bootstrapową nazywamy próbę n elementową, pobraną z n elementowej próbki, za pomocą n krotnego losowania ze zwracaniem.item W tak zbudowanej próbie znajduje się około 63 procent obserwacji.
- Pozostałe dane mogą posłużyć jako zbiór testowy.
- Dodatkowo istnieje szereg estymatorów błędu dla metody prób bootstrapowych.

Estymator błędu prób bootstrapowych

- Dla ciągu B prób bootstrapowych $\mathcal{L}_n^{*1}, \mathcal{L}_n^{*2}, \ldots, \mathcal{L}_n^{*B}$ możemy wyliczyć błąd jako błąd ponownego podstawienia plus obciążenie elementów nieujętych w próbce

$$\hat{e}_{B_1} = \hat{e}_R + \frac{1}{Bn} \sum_{b=1}^{B} \sum_{j=1}^{n} I\left(\hat{d}\left(\mathbf{X}_j; \tilde{\mathcal{L}}_n^{*b} \neq Y_j\right)\right) - \frac{1}{Bn} \sum_{b=1}^{B} \sum_{j=1, \mathbf{Z}_j \in \mathcal{L}_n^{*b}}^{n} I\left(\hat{d}\left(\mathbf{X}_j; \tilde{\mathcal{L}}_n^{*b} \neq Y_j\right)\right)$$

- Możemy też ograniczyć się do elementów spoza próbki

$$\hat{e}_{B_2} = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{j=1}^{n} I\left(\mathbf{Z}_j \notin \mathcal{L}_n^{*b}\right) I\left(\hat{d}\left(\mathbf{X}_j; \tilde{\mathcal{L}}_n^{*b} \neq Y_j\right)\right)}{\sum_{j=1}^{n} I\left(\mathbf{Z}_j \notin \mathcal{L}_n^{*b}\right)}$$

Modyfikacje estymatorów

- Na podstawie poprzednich estymatorów, które stosowały metodę sprawdzania krzyżowego dla prób bootstrapowych, możemy wyprowadzić nowe estymatory.
- Pamiętając, że z prawdopodobieństwem około 37% nie wylosujemy danej obserwacji do próbki stosujemy

$$\hat{e}_{.632} = .368\hat{e}_R + .632\hat{e}_{B_2}$$

- Alternatywnie

$$\hat{e}_{.632+} = (1 - \omega)\hat{e}_R + \omega\hat{e}_{B_2}$$

Parametry estymatora $\hat{e}_{.632+}$

- Estymator $\hat{e}_{.632+}$ jest parametryzowany wyrażeniem ω

$$\omega = \frac{0.632}{1 - 0.368R},$$

- gdzie

$$R = \begin{cases} 1, \text{ jeżeli } \gamma \leqslant \hat{e}_{B_2}, \\ \frac{\hat{e}_{B_2} - \hat{e}_R}{\gamma - \hat{e}_R}, \text{ jeżeli } \gamma, \hat{e}_{B_2} > \hat{e}_R, \\ 0, \text{ w p.p.} \end{cases}$$

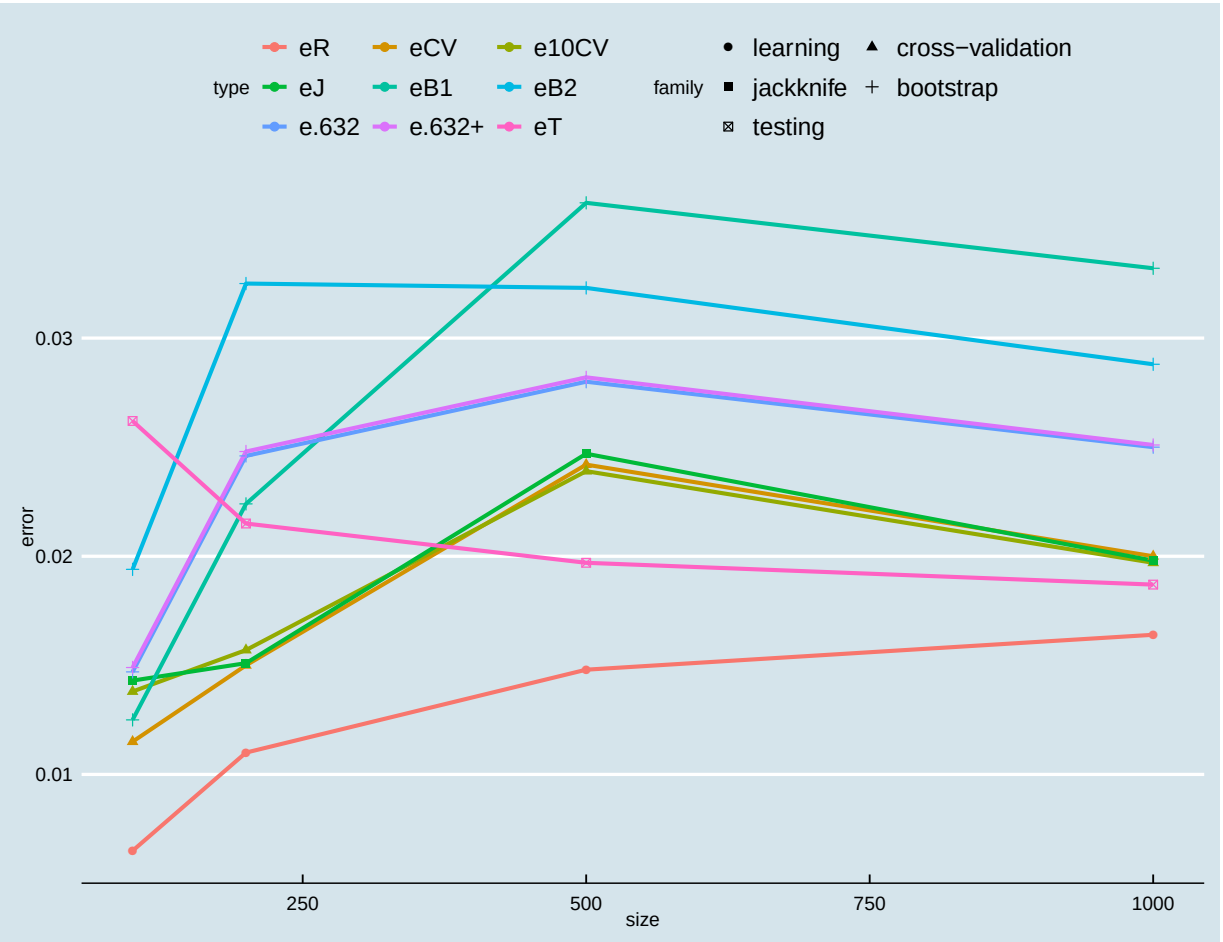
- dla parametru

$$\gamma = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} I\left(\hat{d}\left(\mathbf{x}_j; \tilde{\mathcal{L}}_n \neq Y_j\right)\right)$$

Porównanie estymatorów

- Przeprowadzono porównanie metod estymacji błędu na wygenerowanych danych [Krzyśko et al., 2008].
- Dla dwóch klas wygenerowano po 100, 200, 500 i 1000 przykładów.
- Wyniki działania estymatorów porównano z wynikiem na próbie testowej składającej się z 100 tysięcy elementów.
- Wyniki uzyskane na zbiorze testowym potraktujemy jako wyniki referencyjne.

Wyniki porównania estymatorów



- Dla próbki 100 elementowej wszystkie estymatory zaniżają wartość błędu (próbka jest zbyt mała).
- Estymator ponownego podstawienia zawsze powoduje niedoszacowanie.
- Metody bootstrapowe dają wyższe oszacowania niż metody sprawdzenia krzyżowego.

Rysunek 4: Porównanie estymatorów

Ograniczenia estymatorów

- Metody sprawdzania krzyżowego są szeroko stosowane do porównywania klasyfikatorów. Jednakże należy pamiętać, że opierają się one na doborze podzbiorów zbioru uczącego (podobnie jak metody bootstrapowe).
- Z tego też powodu nie należy ich używać w sytuacji, gdy klasyfikator będzie używany na wyraźnie innych danych niż dane uczące.
- Dotyczy to w szczególności analizy danych zmiennych w czasie. W takim wypadku powinniśmy wyodrębniać zbiór testowy.

Bibliografia I

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

Measuring performance: Auc (auroc).

[Krzyśko et al., 2008] Krzyśko, M., Wołyński, W., Górecki, T., and Skorzybut, M. (2008).

Systemy uczące się. Rozpoznawanie wzorców analiza skupień i redukcja wymiarowości.

WNT.

[Webb, 2003] Webb, A. R. (2003).

Statistical Pattern Recognition.

John Wiley & Sons.