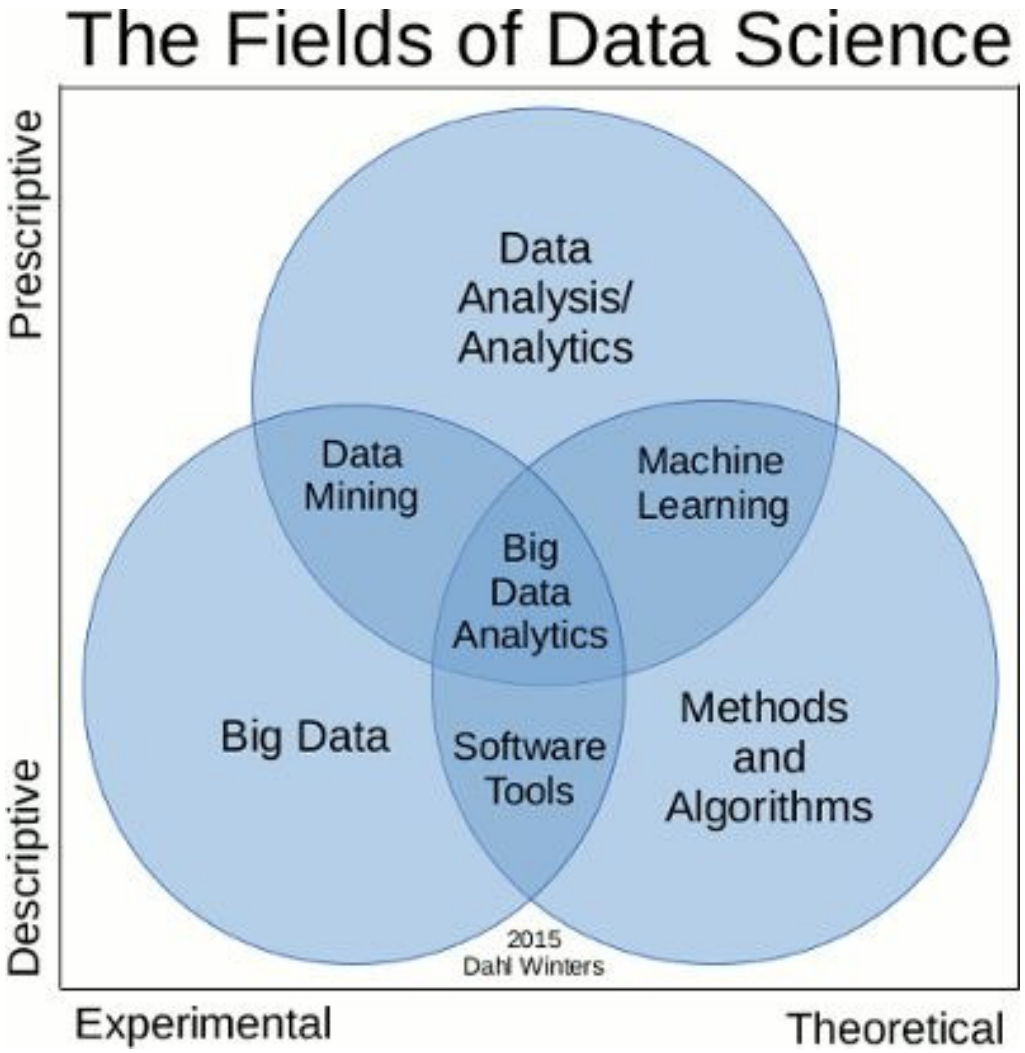
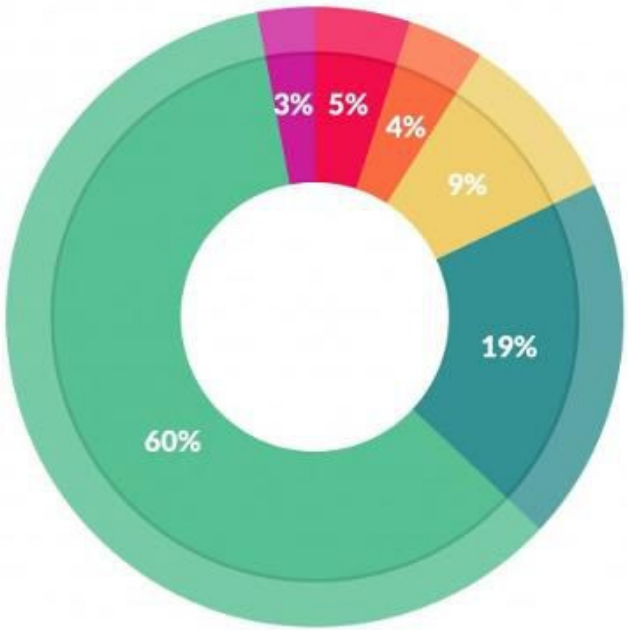
Wprowadzenie



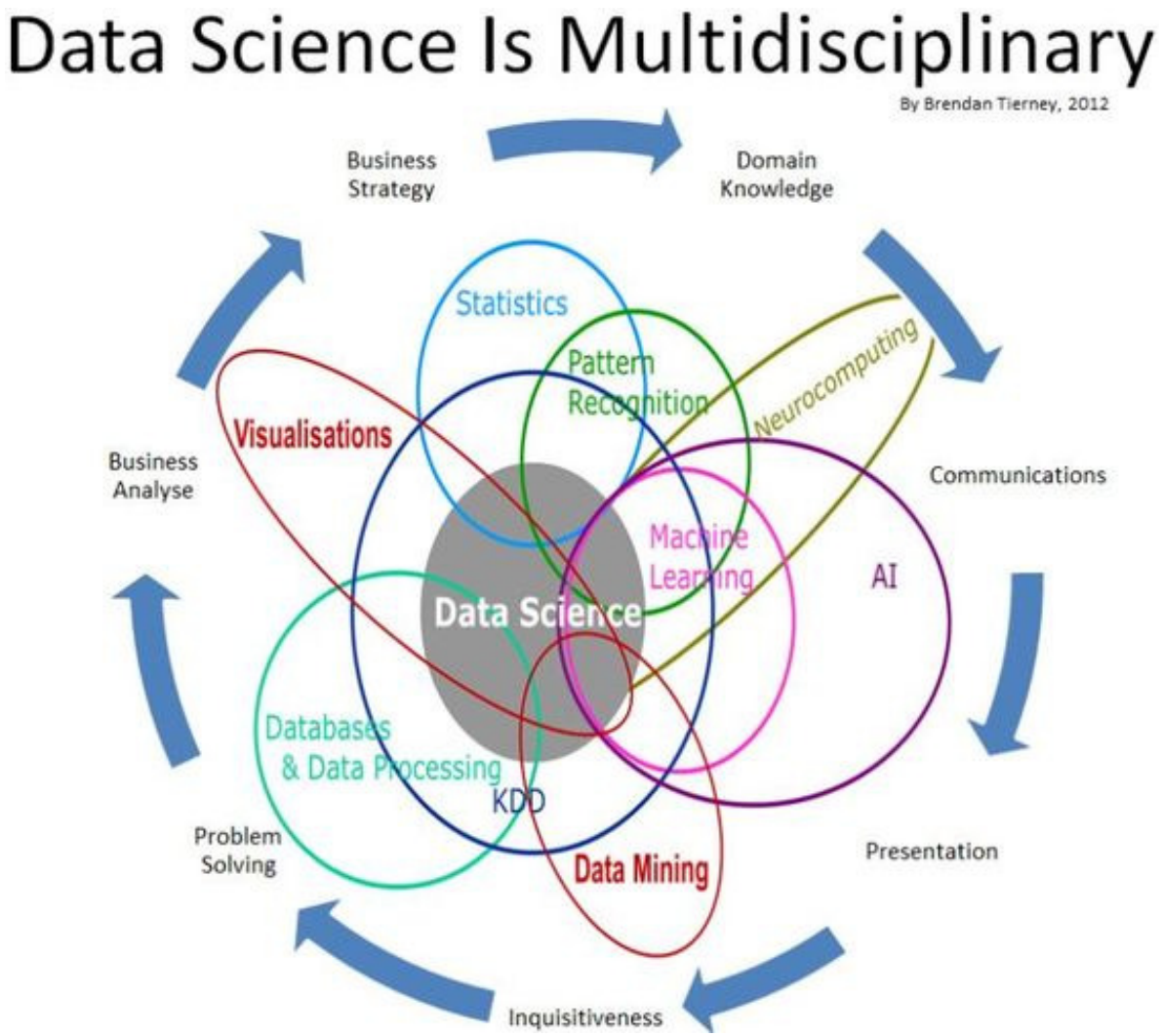
Wprowadzenie



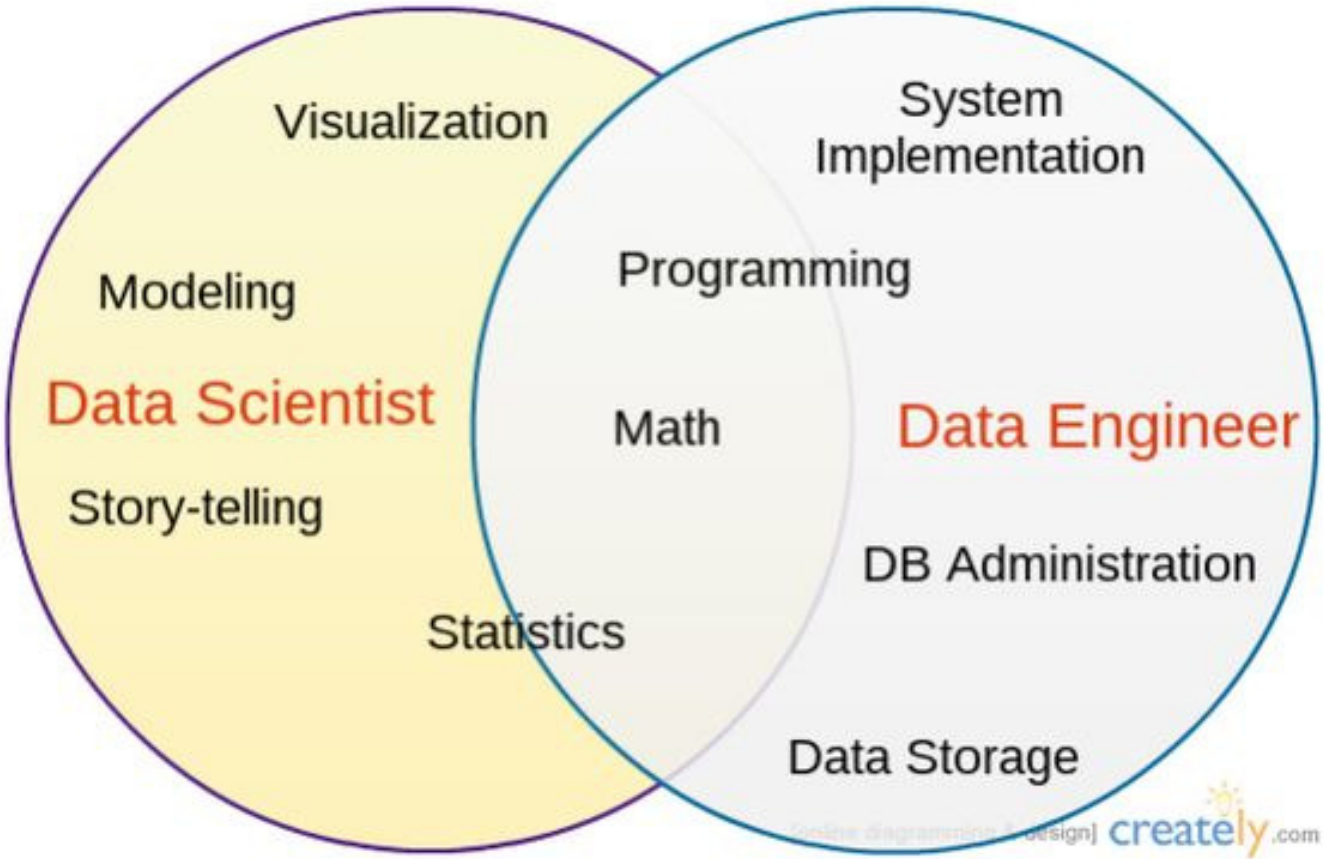
What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

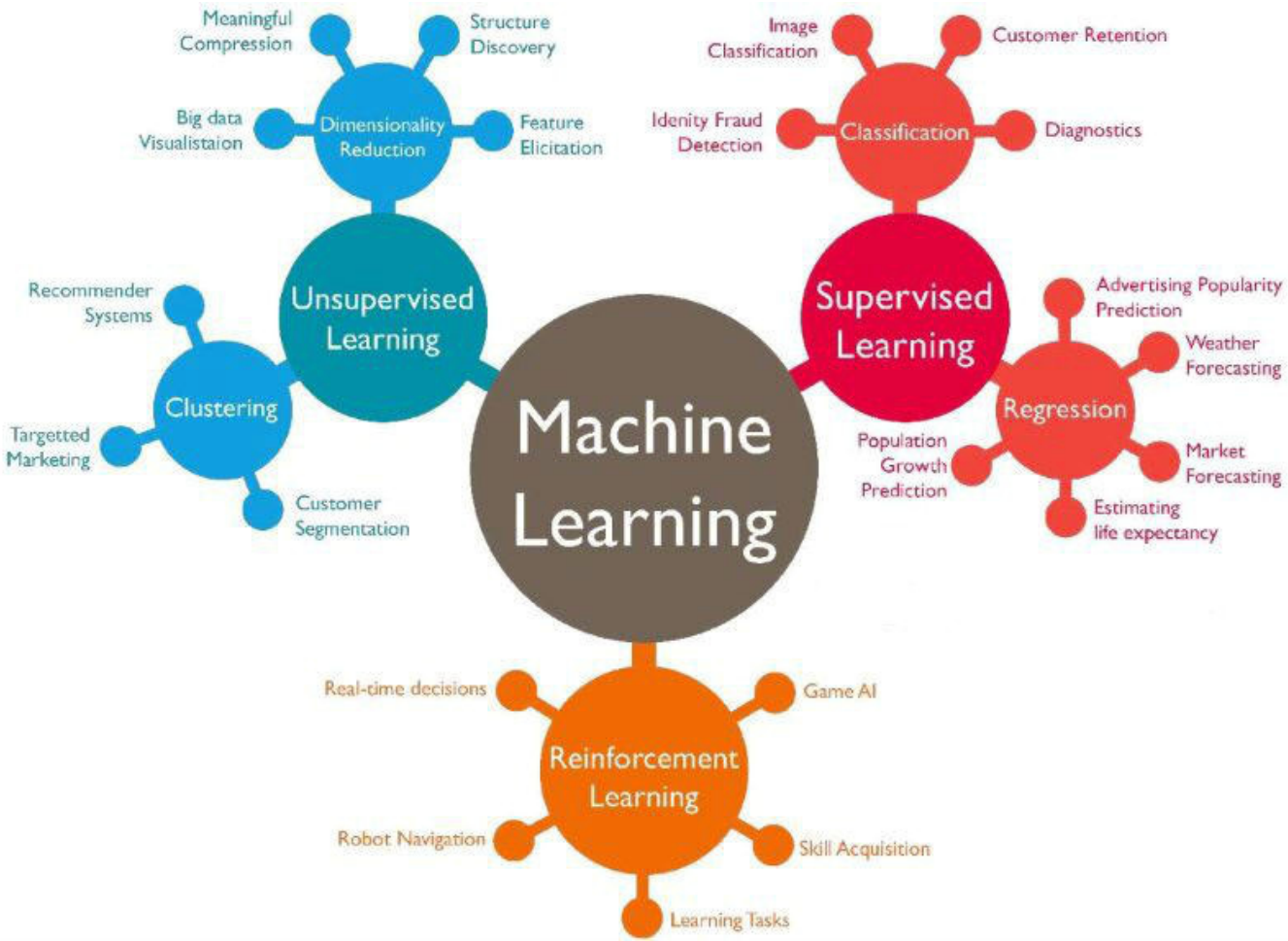
Wprowadzenie



Wprowadzenie

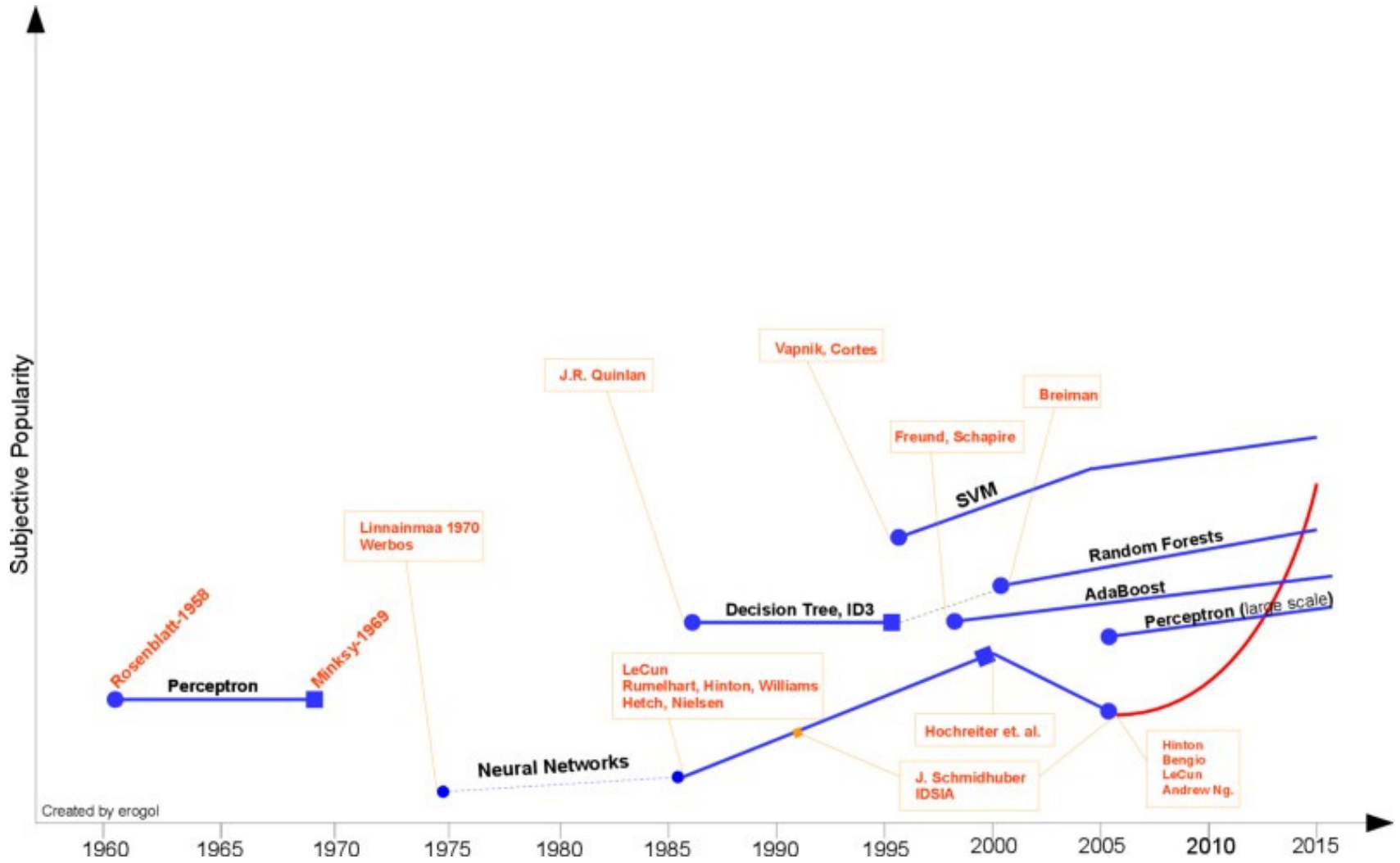


Wprowadzenie



Machine Learning Mind Map

Historia


Subjective Popularity
Created by erogol
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
Rosenblatt-1958
Perceptron
Minksy-1969
Linnainmaa 1970
Werbos
Neural Networks
J.R. Quinlan
Decision Tree, ID3
Vapnik, Cortes
SVM
Freund, Schapire
Breiman
Random Forests
AdaBoost
Perceptron (large scale)
LeCun
Rumelhart, Hinton, Williams
Hetch, Nielsen
Hochreiter et. al.
J. Schmidhuber
IDSIA
Hinton
Bengio
LeCun
Andrew Ng.

Wstęp

Uczenie się pod nadzorem lub uczenie się z przykładów jest procesem budowy, na bazie dostępnych danych wejściowych $\boldsymbol{X}_i$ oraz wyjściowych Y_i, $i = 1, 2, \ldots, n$, reguły klasyfikacyjnej zwanej inaczej klasyfikatorem, służącej do predykcji etykiety Y grupy, do której należy obserwacja $\boldsymbol{X}$.

Wstęp

Załóżmy, że dysponujemy K niezależnymi, prostymi próbami losowymi o liczebnościach, odpowiednio, $n_1, n_2, \ldots, n_K$, pobranymi z K różnych populacji (klas, grup):

$$\begin{aligned} \boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1} &\quad - \quad \text{z populacji 1} \\ \boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2} &\quad - \quad \text{z populacji 2} \\ &\ldots \\ \boldsymbol{X}_{K1}, \boldsymbol{X}_{K2}, \ldots, \boldsymbol{X}_{Kn_K} &\quad - \quad \text{z populacji } K, \end{aligned}$$

gdzie $\boldsymbol{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})'$ jest j-tą obserwacją z i-tej populacji zawierającą p obserwowanych cech, $i = 1, 2, \ldots, K$, $j = 1, 2, \ldots, n_i$.

Wstęp

Powyższe dane można wygodniej zapisać w innej postaci, a mianowicie w postaci jednego ciągu n uporządkowanych par losowych

$$(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n),$$

gdzie $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})' \in \mathcal{X} \subset \mathbb{R}^p$ jest i-tą obserwacją, natomiast Y_i jest etykietą populacji, do której ta obserwacja należy, przyjmującą wartości w pewnym skończonym zbiorze $\mathcal{Y}$, $i = 1, 2, \ldots, n$, $n = n_1 + n_2 + \cdots + n_K$.
Składowe wektora $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ nazywać będziemy cechami, zmiennymi lub atrybutami. Próbę $\mathcal{L}_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ nazywać będziemy próbą uczącą.

Wstęp

Interesuje nas problem predykcji etykiety Y na podstawie wektora cech $\boldsymbol{X}$. Problem ten nazywany jest klasyfikacją, dyskryminacją, uczeniem się pod nadzorem lub rozpoznawaniem wzorców.
Reguła klasyfikacyjna, zwana krótko klasyfikatorem, jest funkcją $d : \mathcal{X} \to \mathcal{Y}$. Gdy obserwujemy nowy wektor $\boldsymbol{X}$, to prognozą etykiety Y jest $d(\boldsymbol{X})$.

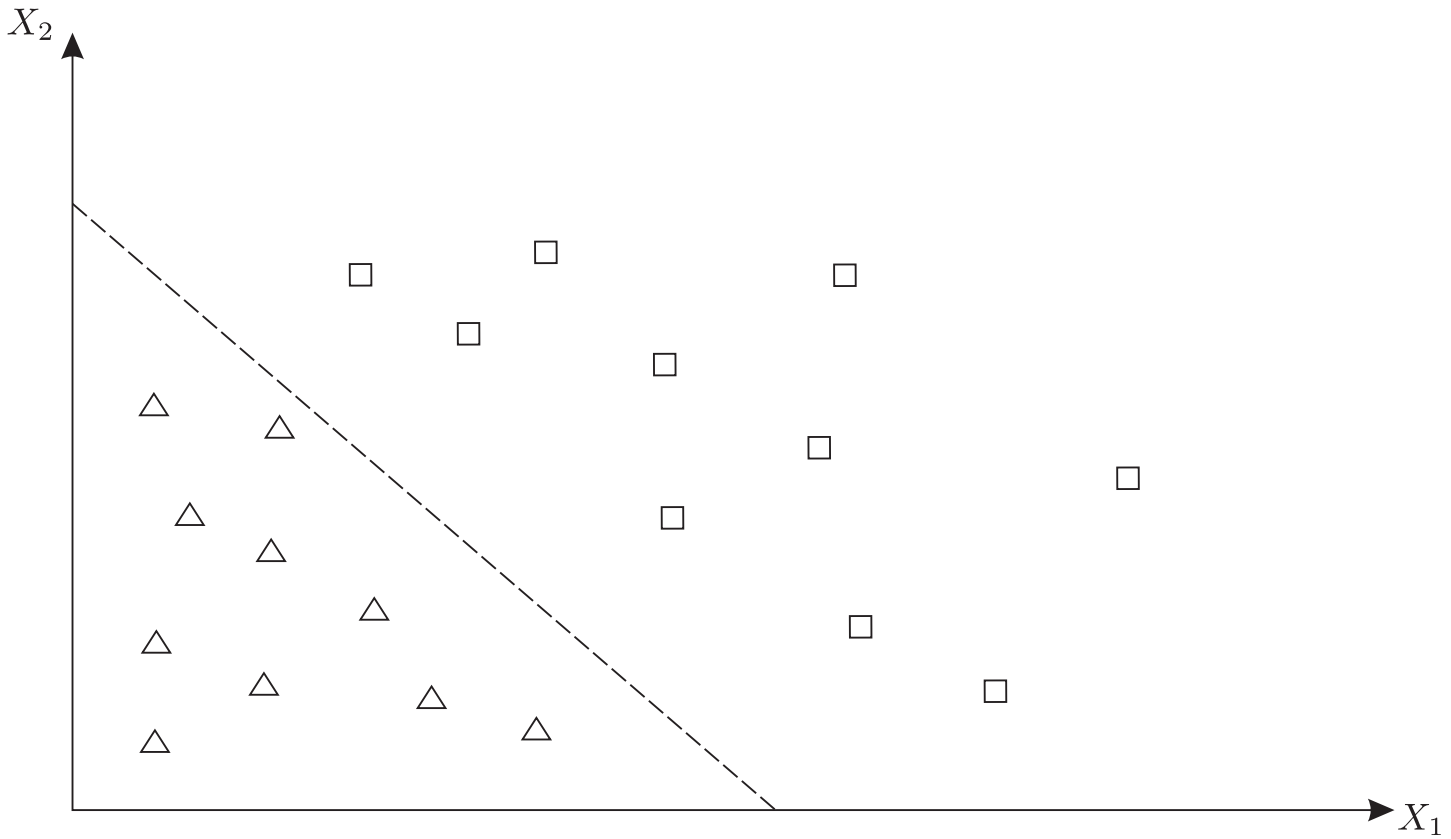
Wstęp

Na poniższym rysunku pokazanych jest 20 punktów. Wektor cech $\boldsymbol{X} = (X_1, X_2)'$ jest dwuwymiarowy a etykieta $Y \in \mathcal{Y} = \{1, 0\}$. Wartości cechy $\boldsymbol{X}$ dla $Y = 0$ reprezentowane są przez trójkąty, a dla $Y = 1$ przez kwadraty. Linia przerywana reprezentuje liniową regułę klasyfikacyjną postaci

$$d(\boldsymbol{x}) = \begin{cases} 1, & \text{jeżeli } a + b_1 x_1 + b_2 x_2 > 0, \\ 0, & \text{poza tym.} \end{cases}$$

Każdy punkt leżący poniżej tej linii klasyfikowany jest do grupy o etykiecie 0 oraz każdy punkt leżący powyżej tej linii klasyfikowany jest do grupy o etykiecie 1.

Wstęp



Rzeczywisty poziom błędu

Klasyczny problem klasyfikacji polega na predykcji nieznanej etykiety $Y \in \mathcal{Y}$ na podstawie wektora cech $\boldsymbol{X} \in \mathcal{X} \subseteq \mathrm{R}^p$ klasyfikowanego obiektu. Naszym celem jest znalezienie takiego klasyfikatora $d : \mathcal{X} \to \mathcal{Y}$, który daje dokładną predykcję. Miarą jakości klasyfikatora jest jego rzeczywisty poziom błędu.

Definicja

Rzeczywisty poziom błędu klasyfikatora d jest równy

$$e(d) = P(d(\boldsymbol{X}) \neq Y). \tag{1}$$

Klasyfikator bayesowski

Weźmy wpierw pod uwagę przypadek dwóch klas, tj. gdy $\mathcal{Y} = \{1, 0\}$. W celu stworzenia modelu klasyfikacji załóżmy, że $(\boldsymbol{X}, Y)$ jest parą losową o wartościach w $\mathrm{R}^p \times \{1, 0\}$ oraz, że jej rozkład prawdopodobieństwa opisuje para (μ, r), gdzie μ jest miarą probabilistyczną wektora $\boldsymbol{X}$ oraz r jest regresją Y względem $\boldsymbol{X}$. Bardziej precyzyjnie, dla zbioru $A \subseteq \mathrm{R}^p$,

$$\mu(A) = \mathrm{P}(\boldsymbol{X} \in A)$$

oraz dla każdego $\boldsymbol{x} \in \mathrm{R}^p$,

$$\begin{aligned} r(\boldsymbol{x}) &= \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = 1 \cdot \mathrm{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) + 0 \cdot \mathrm{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x}) \\ &= \mathrm{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) \end{aligned}$$

Zatem $r(\boldsymbol{x})$ jest prawdopodobieństwem warunkowym, że $Y = 1$, gdy $\boldsymbol{X} = \boldsymbol{x}$. Rozkład prawdopodobieństwa pary $(\boldsymbol{X}, Y)$ wyznacza para (μ, r). Funkcja $r(\boldsymbol{x})$ nazywa się prawdopodobieństwem a posteriori.

Klasyfikator bayesowski

Z twierdzenia Bayes'a mamy

$$\begin{aligned} r(\boldsymbol{x}) &= \mathrm{P}(Y = 1 | \boldsymbol{X} = \boldsymbol{x}) \\ &= \frac{f(\boldsymbol{x}|Y = 1)\mathrm{P}(Y = 1)}{f(\boldsymbol{x}|Y = 1)\mathrm{P}(Y = 1) + f(\boldsymbol{x}|Y = 0)\mathrm{P}(Y = 0)} \\ &= \frac{\pi_1 f_1(\boldsymbol{x})}{\pi_1 f_1(\boldsymbol{x}) + \pi_0 f_0(\boldsymbol{x})}, \end{aligned}$$

gdzie

$$\pi_1 = \mathrm{P}(Y = 1), \quad \pi_0 = \mathrm{P}(Y = 0), \quad \pi_1 + \pi_0 = 1$$

są prawdopodobieństwami a priori.

Klasyfikator bayesowski

Definicja

Klasyfikator postaci

$$d_B(\boldsymbol{x}) = \begin{cases} 1, & \text{jeżeli } r(\boldsymbol{x}) > \frac{1}{2}, \\ 0, & \text{poza tym.} \end{cases}$$

nazywać będziemy klasyfikatorem bayesowskim

Klasyfikator bayesowski

Klasyfikator bayesowski zapisać można w innych równoważnych postaciach:

$$d_B(\boldsymbol{x}) = \begin{cases} 1, & \text{jeżeli } \mathrm{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > \mathrm{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x}), \\ 0, & \text{poza tym.} \end{cases}$$

lub

$$d_B(\boldsymbol{x}) = \begin{cases} 1, & \text{jeżeli } \pi_1 f_1(\boldsymbol{x}) > \pi_0 f_0(\boldsymbol{x}), \\ 0, & \text{poza tym.} \end{cases}$$

Klasyfikator bayesowski

Twierdzenie

Klasyfikator bayesowski jest optymalny, tj. jeżeli d jest jakimkolwiek innym klasyfikatorem, to $e(d_B) \leq e(d)$, gdzie $e(d)$ jest rzeczywistym poziomem błędu klasyfikatora d.

Aktualny poziom błędu

Niestety, klasyfikator bayesowski zależy od rozkładu prawdopodobieństwa pary $(\boldsymbol{X}, Y)$. Najczęściej rozkład ten nie jest znany i stąd również nie jest znany klasyfikator bayesowski d_B. Pojawia się zatem problem skonstruowania klasyfikatora $\hat{d}(\boldsymbol{x}) = \hat{d}(\boldsymbol{x}; \mathcal{L}_n)$ opartego na próbie uczącej $\mathcal{L}_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ zaobserwowanej w przeszłości. Proces konstruowania klasyfikatora $\hat{d}$ nazywany jest uczeniem się, uczeniem pod nadzorem lub uczeniem się z nauczycielem. Chcemy znaleźć taki klasyfikator $\hat{d}$, dla którego $e(\hat{d})$ jest bliskie $e(d_B)$. Jednakże $e(\hat{d})$ jest zmienną losową, ponieważ zależy od losowej próby uczącej $\mathcal{L}_n$.

Aktualny poziom błędu

W naszym modelu klasyfikacyjnym zakładamy, że próba ucząca $\mathcal{L}_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ jest ciągiem niezależnych par losowych o identycznym rozkładzie prawdopodobieństwa takim, jak rozkład pary $(\boldsymbol{X}, Y)$. Jakość klasyfikatora $\hat{d}$ mierzona jest za pomocą warunkowego prawdopodobieństwa błędu

$$e(\hat{d}) = \mathrm{P}(\hat{d}(\boldsymbol{X}) \neq Y | \mathcal{L}_n),$$

gdzie para losowa $(\boldsymbol{X}, Y)$ jest niezależna od próby uczącej $\mathcal{L}_n$. Wielkość $e(\hat{d})$ nazywamy aktualnym poziomem błędu klasyfikatora. Chcemy znaleźć taki klasyfikator $\hat{d}$, dla którego $e(\hat{d})$ jest bliskie $e(d_B)$. Jednakże $e(\hat{d})$ jest zmienną losową, ponieważ zależy od losowej próby uczącej $\mathcal{L}_n$.

Estymacja aktualnego poziomu błędu

Niech $\hat{d}(\boldsymbol{x}) = \hat{d}(\boldsymbol{x}; \mathcal{L}_n)$ oznacza klasyfikator skonstruowany przy pomocy próby uczącej $\mathcal{L}_n$. Niech $\hat{e} \equiv \hat{e}(\hat{d})$ oznacza ocenę aktualnego poziomu błędu klasyfikatora $\hat{d}$. Ocenę tę nazywać będziemy błędem klasyfikacji. W sytuacjach, kiedy na populacje nie narzuca się żadnej konkretnej rodziny rozkładów, jedyną drogą oceny prawdopodobieństwa $e(\hat{d})$ jest użycie metod estymacji nieparamerycznej.

Estymacja aktualnego poziomu błędu

W najlepszej sytuacji jesteśmy wtedy, gdy dysponujemy m-elementową próbą testową $\mathcal{T}_m$ niezależną od próby uczącej $\mathcal{L}_n$. Niech zatem $\mathcal{T}_m = \{(\boldsymbol{X}_1^t, Y_1^t), \ldots, (\boldsymbol{X}_m^t, Y_m^t)\}$. Wtedy za estymator aktualnego poziomu błędu klasyfikatora $\hat{d}$ przyjmujemy:

$$\hat{e}_{\mathcal{T}} = \frac{1}{m} \sum_{j=1}^{m} I(\hat{d}(\boldsymbol{X}_j^t; \mathcal{L}_n) \neq Y_j^t).$$

W przypadku gdy nie dysponujemy niezależną próbą testową, do estymacji używamy jedynie próby uczącej.

Estymacja aktualnego poziomu błędu

Naturalną oceną aktualnego poziomu błędu jest wtedy wartość estymatora ponownego podstawienia (resubstytucji)

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

Wartość tego estymatora uzyskuje się poprzez klasyfikację regułą $\hat{d}$ tych samych obserwacji, które służyły do jej konstrukcji. Oznacza to, iż próba ucząca jest zarazem próbą testową. Estymator ten jest więc obciążonym estymatorem wielkości $e(\hat{d})$ i zaniża jej rzeczywistą wartość. Uwidacznia się to szczególnie w przypadku złożonych klasyfikatorów opartych na relatywnie małych próbach uczących. Redukcję obciążenia można uzyskać stosując poniższe metody estymacji.

Estymacja aktualnego poziomu błędu

Jednym ze sposobów redukcji obciążenia estymatora $\hat{e}_R$ jest tzw. metoda podziału próby na dwa podzbiory: próbę uczącą i próbę testową. Wówczas klasyfikator konstruuje się za pomocą pierwszego z nich, drugi natomiast służy do konstrukcji estymatora. Wykorzystanie tylko części informacji w celu uzyskania reguły klasyfikacyjnej prowadzi jednak często do zawyżenia wartości estymatora błędu. Rozwiązaniem tego problemu jest metoda sprawdzania krzyżowego.

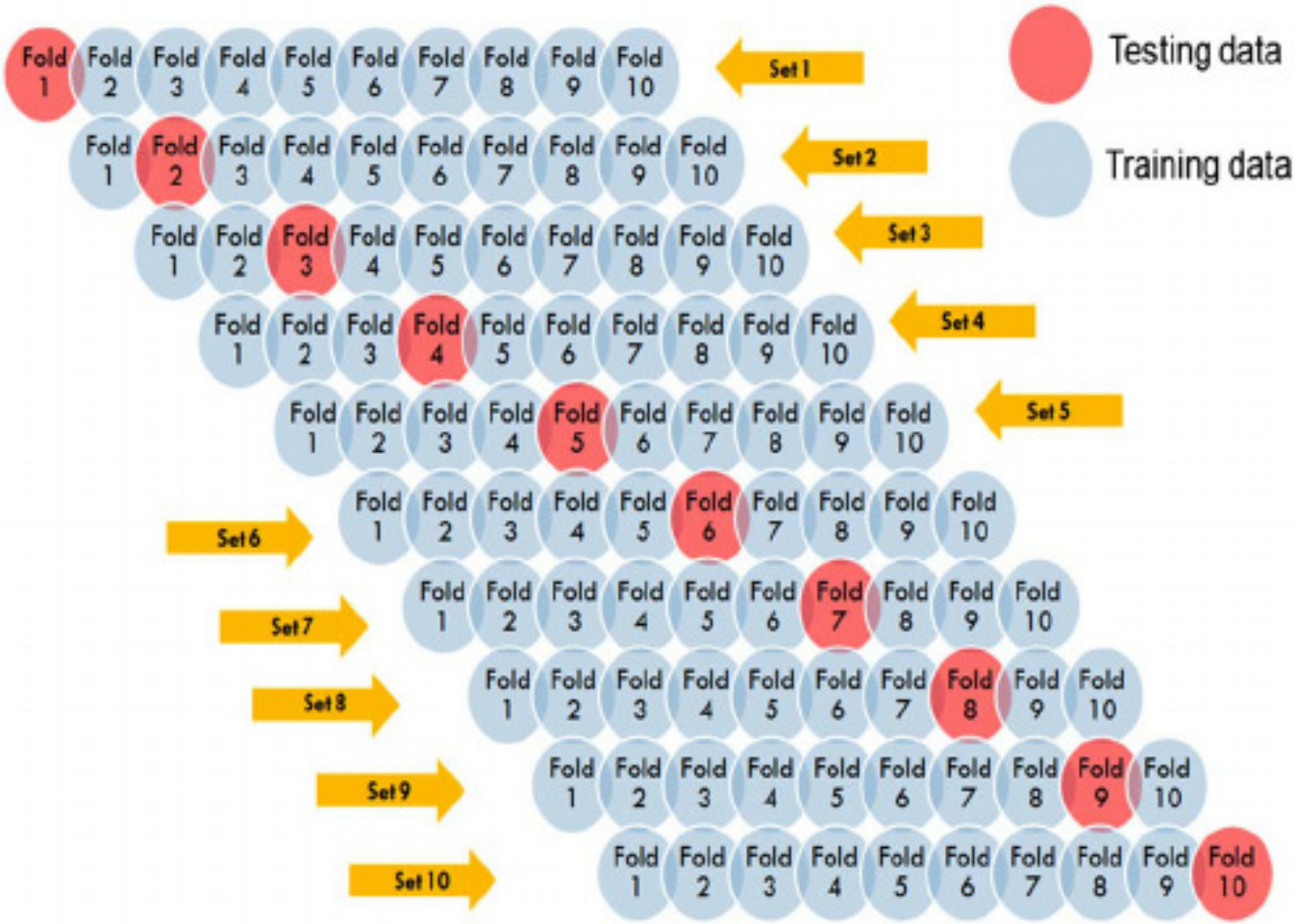
Oznaczmy przez $\mathcal{L}_n^{(-j)}$ próbę uczącą $\mathcal{L}_n$ z której usunięto obserwację $\boldsymbol{Z}_j = (\boldsymbol{X}_j, Y_j)$. Klasyfikator konstruuje się wykorzystując próbę $\mathcal{L}_n^{(-j)}$, a następnie testuje się go na pojedynczej obserwacji $\boldsymbol{Z}_j$. Czynność tę powtarza się n razy, dla każdej obserwacji $\boldsymbol{Z}_j$ z osobna. Odpowiedni estymator ma postać:

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

Estymacja aktualnego poziomu błędu

Procedura ta w każdym z n etapów jest w rzeczywistości metodą podziału próby dla przypadku jednoelementowego zbioru testowego. Każda obserwacja próby jest użyta do konstrukcji klasyfikatora $\hat{d}$. Każda z nich jest też (dokładnie jeden raz) elementem testowym.
Estymator ten, choć granicznie nieobciążony, ma większą wariancję. Ponadto wymaga on konstrukcji n klasyfikatorów, co dla dużych n oznacza znaczący wzrost obliczeń. Rozwiązaniem pośrednim jest metoda rotacyjna, zwana często v-krokową metodą sprawdzania krzyżowego. Polega ona na losowym podziale próby na v podzbiorów, przy czym $v-1$ z nich tworzy próbę uczącą, natomiast pozostały — próbę testową. Procedurę tę powtarza się v razy, dla każdego podzbioru rozpatrywanego kolejno jako zbiór testowy.

Estymacja aktualnego poziomu błędu



Estymacja aktualnego poziomu błędu

Odpowiedni estymator jest postaci:

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

gdzie $\tilde{\mathcal{L}}_n^{(1)}, \tilde{\mathcal{L}}_n^{(2)}, \ldots, \tilde{\mathcal{L}}_n^{(v)}$ jest losowym v-podziałem próby $\mathcal{L}_n$ na równoliczne podzbiory, a $\tilde{\mathcal{L}}_n^{(-i)} = \mathcal{L}_n \backslash \tilde{\mathcal{L}}_n^{(i)}$, $i = 1, 2, \ldots, v$.

Estymacja aktualnego poziomu błędu

Metoda ta daje mniejsze obciążenie błędu niż metoda podziału próby i wymaga mniejszej liczby obliczeń w porównaniu ze sprawdzaniem krzyżowym (jeśli tylko $v < n$). W zagadnieniu estymacji aktualnego poziomu błędu zalecane jest obranie wartości $v = 10$. Metoda sprawdzania krzyżowego jest powszechnie wykorzystywana w zagadnieniu wyboru modelu. Z rodziny klasyfikatorów opisanej parametrycznie wybieramy wtedy klasyfikator, dla którego błąd klasyfikacji ma wartość najmniejszą.

Estymacja aktualnego poziomu błędu

Definicja

Próbą bootstrapową nazywamy próbę n-elementową pobraną z n-elementowej próby uczącej w procesie n-krotnego losowania pojedynczych obserwacji ze zwracaniem.

Niech $\mathcal{L}_n^{*1}, \mathcal{L}_n^{*2}, \ldots, \mathcal{L}_n^{*B}$ będzie ciągiem kolejno pobranych B prób bootstrapowych. Bootstrapowa ocena aktualnego poziomu błędu ma postać

$$\hat{e}_B = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{j=1}^{n} I(\boldsymbol{Z}_j \notin \mathcal{L}_n^{*b}) I(\hat{d}(\boldsymbol{X}_j; \mathcal{L}_n^{*b}) \neq Y_j)}{\sum_{j=1}^{n} I(\boldsymbol{Z}_j \notin \mathcal{L}_n^{*b})}.$$

Widać, że powyższa ocena aktualnego poziomu błędu jest uzyskana metodą sprawdzania krzyżowego zastosowaną do prób bootstrapowych.

Estymacja aktualnego poziomu błędu

W celu dalszej redukcji obciążenia tego estymatora zaproponowano estymator postaci:

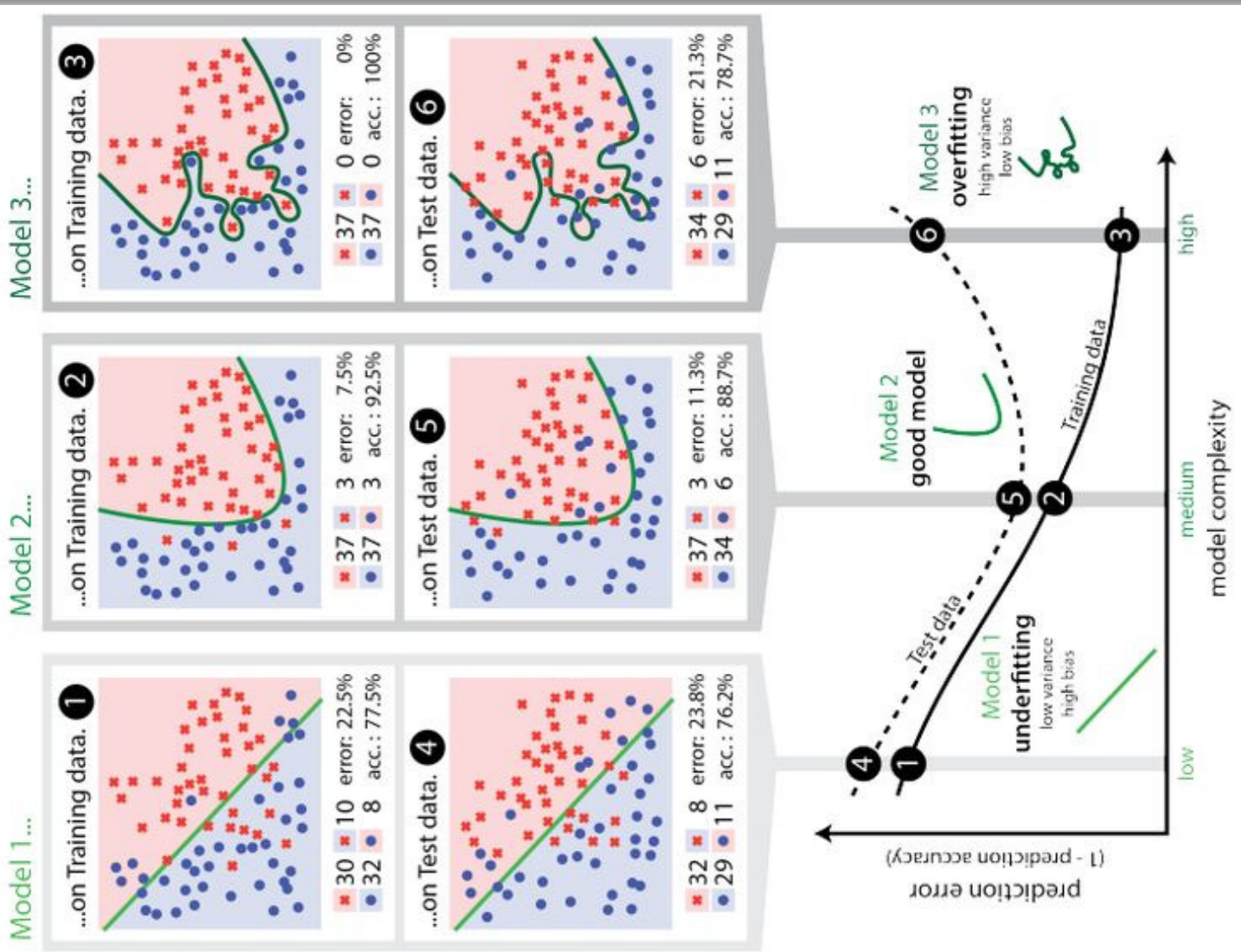
$$\hat{e}_{.632} = 0{,}368\,\hat{e}_R + 0{,}632\,\hat{e}_B.$$

Waga 0,368 jest przybliżoną wartością wielkości $e^{-1} = \lim_{n\to\infty}(1 - 1/n)^n$, i jest graniczną wartością prawdopodobieństwa nie wylosowania obserwacji z próby uczącej do próby bootstrapowej.

Macierz pomyłek

Sam błąd klasyfikacji nie wystarczy aby stwierdzić, czy klasyfikator działa dobrze. Po zakończeniu procedury oceny błędu klasyfikacji naszego modelu zostajemy z listą obserwacji testowych, gdzie dla każdej z nich znamy klasę obserwowaną oraz klasę przewidywaną przez nasz model. Zliczając liczbę przypadków dla każdej z kombinacji tych dwóch klas (obserwowanej i przewidywanej) możemy stworzyć tzw. macierz pomyłek (ang. confusion matrix). Możemy na jej podstawie ocenić błąd modelu: musimy zsumować wartości na głównej przekątnej (obserwacje poprawnie rozpoznane) i podzielić je przez liczbę wszystkich obserwacji testowych.

Przeuczenie i niedouczenie modelu



ZeroR – najprostszy klasyfikator

Algorytm ZeroR (Zero Rule) ignoruje wszystkie cechy, oprócz etykiet. Na ich podstawie, sprawdza najczęściej występującą klasę i tworzy klasyfikator który zawsze będzie ją zwracać. Zatem algorytm ZeroR zawsze przypisuje nowe obserwacje do tej samej klasy (klasy większościowej), niezależnie od wartości cech.

OneR

Algorytm OneR (One Rule) jest tylko nieco bardziej wyrafinowany niż ZeroR. Podczas gdy ZeroR ignoruje wszystkie cechy, OneR wybiera tylko jedną i ignoruje pozostałe. Wybierając najlepszą cechę sprawdza on błąd na danych uczących dla klasyfikatorów zbudowanych na każdej z cech osobno i wybiera tą z nich, która minimalizuje błąd. Dla każdej cechy dzieli on dane uczące na podzbiory ze względu na wartość tej cechy. Następnie, na każdym z nich używa algorytmu ZeroR. Pokazano, że pomimo prostoty jest to metoda jedynie nieznacznie ustępująca najlepszym klasyfikatorom.

R.C. Holte. (1993). *Very simple classification rules perform well on most commonly used datasets*. Machine Learning 11(1):63–90.

OneR

Day	Outlook	Temperature	Humidity	PlayTennis
D1	sunny	hot	high	NO
D2	sunny	hot	high	NO
D3	overcast	hot	high	YES
D4	overcast	hot	normal	YES
D5	rain	mild	high	NO

Skonstruujmy klasyfikatory dla różnych atrybutów:

Outlook	YES	NO	trafność
sunny	0	2	5/5 = 100%
overcast	2	0	
rain	0	1	

Temperature	YES	NO	trafność
hot	2	2	3/5 = 60%
mild	0	1	

Humidity	YES	NO	trafność
high	1	3	4/5 = 80%
normal	1	0	

Jak widzimy, najwyższą trafność osiągamy korzystając z atrybutu *Outlook*. Algorytm stworzy więc następujący klasyfikator:

IF $Outlook = sunny$ THEN $PlayTennis = NO$
IF $Outlook = overcast$ THEN $PlayTennis = YES$
IF $Outook = rain$ THEN $PlayTennis = NO$