Podstawy Przetwarzania Danych

Wykład 3: Redukcja zaszumienia danych

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

Wydział Matematyki i Nauk Informacyjnych

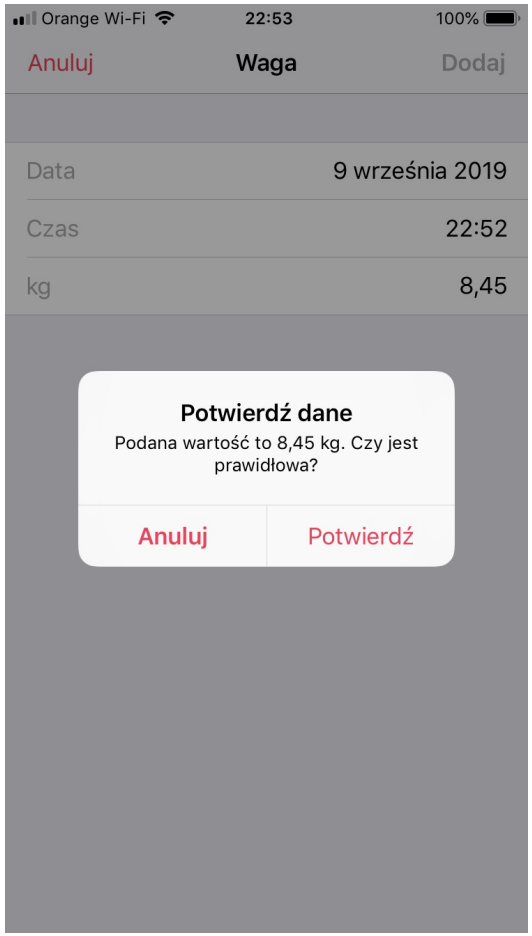
Wersja 1.3
18 października 2021

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.

Odrzucanie pomiarów



Rysunek 1: Prawdopodobnie błędny pomiar

- Jeżeli jeden z wyników serii pomiarowej wyraźnie odbiega od pozostałych to musimy zadecydować czy nie jest on wynikiem pomyłki i nie powinien zostać odrzuconym.
- Może on być jednak wynikiem wiarygodnym i powinien być wykorzystany z innymi pomiarami.

Analiza serii pomiarowej

- Rozważmy serię pomiarową podaną w [Taylor, 1995]

$$3.8; 3.5; 3.9; 3.9; 3.4; 1.8 \ [\mathrm{s}]$$

- Średnia wartość to $\bar{x} = 3.4$ s, a odchylenie standardowe $\sigma_x = 0.8$ s.
- Pomiar 1.8 s wydaje się podejrzany, bo odbiega od średniej o dwa odchylenia standardowe.
- Z rozkładu Gaussa wynika, że prawdopodobieństwo uzyskania takiego wyniku wynosi $1 - 0.95 = 0.05$.
- Oznacza to, że możemy oczekiwać jednego takiego wyniku na 20 pomiarów, a wykonaliśmy ich tylko 6.
- Czy możemy więc uznać, że wynik 1.8 jest podejrzany?

Kryterium Chauveneta

- Jeżeli wyniki pomiarów podlegają rozkładowi normalnemu o wartości oczekiwanej $\bar{x}$ i odchyleniu standardowym σ_x wyliczamy dla podejrzanej obserwacji x_{pod} wielokrotność odchyleń standardowych dzielącą ją od wartości średniej:

$$t_{pod} = \frac{|x_{pod} - \bar{x}|}{\sigma_x}.$$

- Następnie wyliczamy, z prawdopodobieństwa P, ile z N obserwacji może dawać taki odczyt:

$$n_{pod} = N * P.$$

- Jeżeli uzyskana liczba jest mniejsza niż $\frac{1}{2}$ to możemy uznać, że ani jedna obserwacja nie powinna przyjmować takiej wartości i powinniśmy odrzucić obserwację.

Zastosowanie kryterium Chauveneta

- Zastosujmy kryterium do szeregu

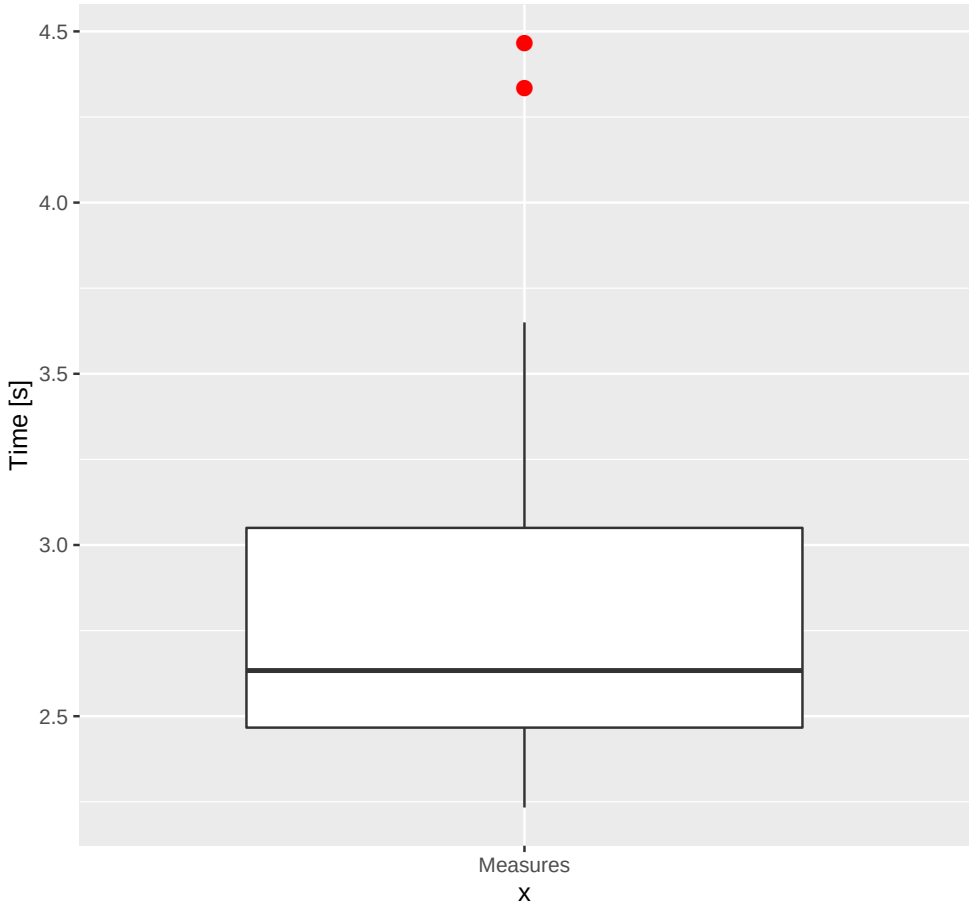
$$3.8; 3.5; 3.9; 3.9; 3.4; 1.8 \ [\mathrm{s}]$$

- Prawdopodobieństwo uzyskania wyniku 1.8 wynosi 0.05, dla sześciu pomiarów powinniśmy mieć $6 * 0.05 \approx \frac{1}{3} < \frac{1}{2}$.
- Zatem należy odrzucić obserwację.

Ograniczenia kryterium Chauveneta

- Odrzucenie obserwacji powoduje zmianę wartości oczekiwanej $\bar{x}$ i odchylenia standardowego σ_x.
- Moglibyśmy próbować zastosować ponownie kryterium dla nowych wartości, aby odrzucić kolejne obserwacje.
- Jednakże osiągnięty konsensus mówi, że ponowne stosowanie kryterium jest niewłaściwe.
- Z tego powodu nie można stosować kryterium w sytuacji, gdy mamy do czynienia z więcej niż jedną obserwacją odstającą.

Wykres pudełkowy



Rysunek 2: Wykres pudełkowy ilustrujący czasu biegu na 400m

- Wykres pudełkowy (*box-plot*) prezentuje rozkład wartości jednej zmiennej.
- Przedstawia pięć statystyk
  - mediana,
  - 25 i 75 percentyl,
  - minimum i maksimum lub -1.5 i 1.5 IQR.
- Obserwacja przekraczająca ± 1.5 IQR (rozstępu międzykwartylowego) jest obserwacją odstającą.

Kontrowersje związane z odrzucaniem obserwacji

- Istnieje szkoła mówiąca, że odrzucanie jakichkolwiek obserwacji jest niewłaściwe, bo anomalny odczyt może być przejawem jakiegoś ważnego efektu.
- Jedynym właściwym podejściem jest powtarzanie pomiaru tak długo, aż ujawni się schemat/przyczyna powstawania anomalii lub pozostałe odczyty zniwelują jej oddziaływanie.
- Z przyczyn praktycznych to podejście nie może być zawsze stosowane.
- Zawsze dopuszczalne, a nawet wskazane, jest usuwanie pomiarów o jawnie zewnętrznym charakterze (uszkodzony miernik, błędnie wprowadzone dane, itp.)

Wykrywanie anomalii

- Przedstawione metody pozwalają na eliminację pojedynczej obserwacji lub na wykrycie odczytów odstających dla kilku zmiennych.
- Jednakże, jeżeli mamy do czynienia z obiektem opisanym w wielowymiarowej przestrzeni danych to ustalenie jego odstającego charakteru staje się bardziej skomplikowane.
- Służą do tego metody wykrywania anomalii.

Typy detektorów anomalii

- Oparte na gęstości rozkładu
  - Metoda zakłada, że normalne obserwacje umieszczone są gęsto, a anomalie leżą w oddaleniu
  - K-Nearest Neighborhood, Local outlier factor
- Oparte na grupowaniu
  - Metoda zakłada, że obserwacje odstające będą znajdować się poza klastrami tworzonymi przez normalne obserwacje
  - K-Mean clustering
- Oparte o funkcje jądrowe
  - Metoda zakłada, że da się, w przestrzeni jądrowej, wyznaczyć prostą oddzielającą normalne obserwacje od anomalii.
  - One-class SVM
- Oparte o autodekoder
  - Oparte na sieciach neuronowych, które kodują i dekodują dane podejście pozwalające na detekcję szumu.
  - Używane w uczeniu głębokim (deep learning)
  - Variational Autoencoder

Metoda wektorów nośnych

- Support Vector Machine (SVM) w literaturze polskiej:
  - metoda wektorów nośnych [Krzyśko et al., 2008],
  - metoda wektorów podpierających [Koronacki and Ćwik, 2006],
  - metoda wektorów wspierających [Jankowski, 2003],
  - metoda wektorów podtrzymujących [Osowski, 2006].

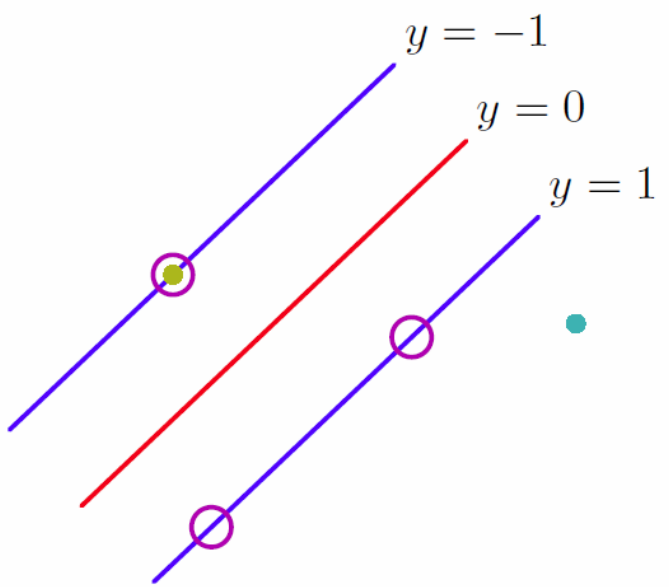
Założenia

- SVM jest binarną metodą dyskryminacji uczenia maszynowego.
- Zbiór uczący składa się z par obserwacja, etykieta

$$\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$$

  y_i jest etykietą klasy przypisaną do wektora cech $\boldsymbol{x}_i$.
- Etykiety należą do zbioru $\{-1, 1\}$

Zadanie



Rysunek 3: Optymalna hiperpłaszczyzna [Bishop, 2006]

- Należy stworzyć hiperpłaszczyznę dzielącą w sposób optymalny, czyli zachowując maksymalny margines między przedstawicielami odmiennych klas, przestrzeń danych.

Model liniowy

- W klasyfikacji wykorzystuje się funkcję f:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}^T \boldsymbol{x}_i + b,$$

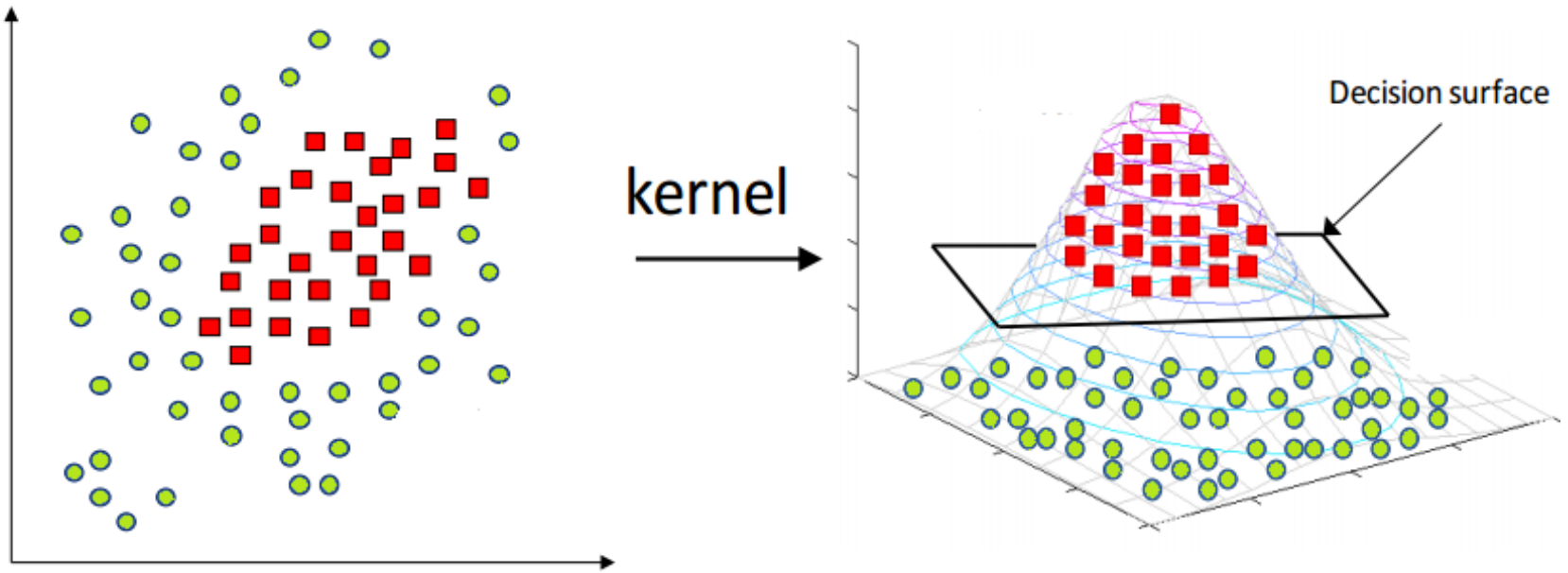
  N jest licznością zbioru uczącego, a parametry α_i i b są wyliczane w procesie uczenia.
- Funkcja f zwraca odległość punktu x od granicy podziału i decyzję o zaklasyfikowaniu przykładu podejmuje się na podstawie jej znaku $\text{sgn}(f(x))$.

Model jądrowy

- Zadanie można przeformułować tak, aby podział dotyczył innej przestrzeni, dla której odnalezienie podziału liniowego będzie łatwiejsze.
- W tym celu należy zastąpić iloczyn skalarny dotyczący przestrzeni podstawowej, iloczynem skalarnym funkcji bazowych nowej przestrzeni, czyli funkcją jądrową K:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) + b.$$

Sztuczka jądrowa



Rysunek 4: Sztuczka jądrowa [Sharma, 2019]

- Nieseparowane dane stają się separowane gdy zwiększymy wymiar przestrzeni danych

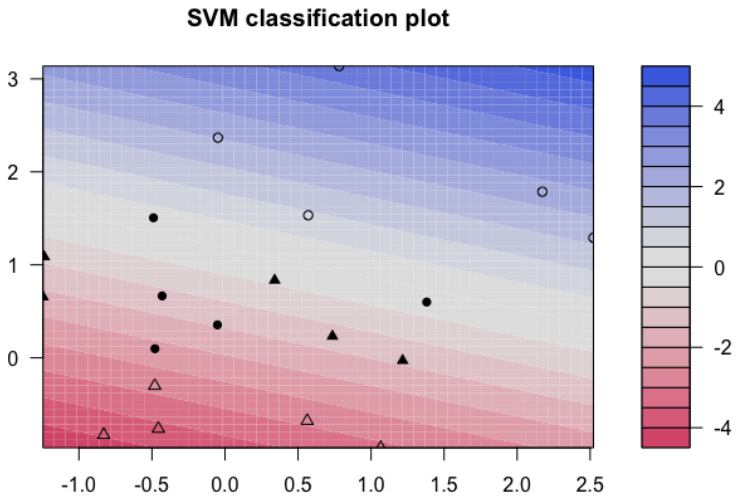
Funkcje jądrowe

- Popularnymi funkcjami jądrowymi dla których $K(\boldsymbol{x_i}, \boldsymbol{x}) = K(\boldsymbol{x_i})K(\boldsymbol{x})$ są:

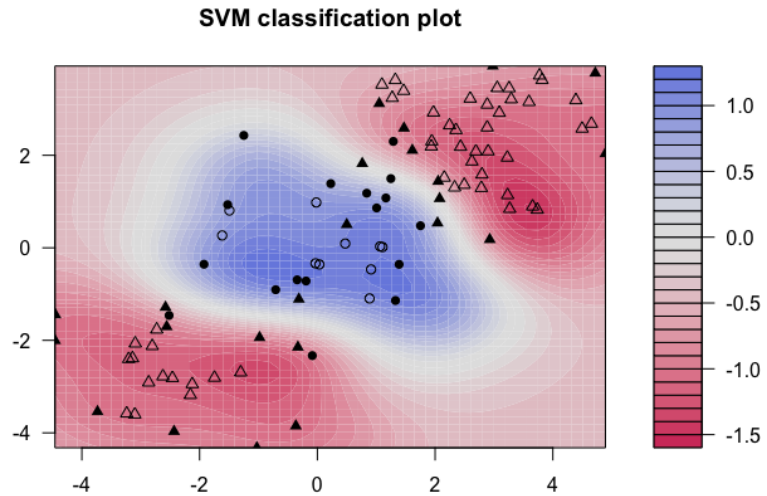
$$\begin{aligned}
\text{Liniowa:} &\quad \boldsymbol{x}^T \boldsymbol{x}_i, \\
\text{Wielomianowa:} &\quad \left(\gamma \boldsymbol{x}^T \boldsymbol{x}_i + c\right)^n, \\
\text{RBF:} &\quad \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{x}_i\|^2\right), \\
\text{Sigmoidalna:} &\quad \text{tgh}\left(\gamma \boldsymbol{x}^T \boldsymbol{x}_i + c\right).
\end{aligned}$$

- Należy zauważyć, że oprócz liniowej wszystkie funkcje jądra wprowadzają dodatkowe parametry.

Wpływ jądra na podział



Rysunek 5: Jądro liniowe



Rysunek 6: Jądro radialne

[UC Business Analytics R Programming Guide, 2000]

Wyliczanie rozdzielającej hiperpłaszczyzny

- Hiperpłaszczyznę charakteryzują wektor współczynników zmiennych ψ i wyraz wolny b.
- Dla ich wyznaczenia należy zmaksymalizować wyrażenie:

$$L(\boldsymbol{\psi}, \boldsymbol{b}, \alpha) = \frac{1}{2}(\boldsymbol{\psi} * \boldsymbol{\psi}) - \sum_{i=1}^{n} \alpha_i (y_i((\boldsymbol{x}_i * \boldsymbol{\psi}) + \boldsymbol{b}) - 1),$$

  gdzie $\alpha_i \geqslant 0$ są mnożnikami Lagrange'a.
- W tym celu należy rozwiązać zadanie optymalizacji programowania kwadratowego.

Warunki punktu siodłowego

- Warunki Karusha–Kuhna–Tuckera, które muszą być spełnione dla rozwiązania optymalnego, dają zależności:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \text{ i } \boldsymbol{\psi}^0 = \sum_{i=1}^{n} y_i \alpha_i \boldsymbol{x}_i$$

- Po podstawieniu uzyskujemy funkcję:

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i^T \boldsymbol{x}_j)$$

która minimalizujemy przy warunkach:

$$(\forall i) \alpha_i \geqslant 0$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Interpretacja wektorów nośnych

- Dla rozwiązania optymalnego:

$$\boldsymbol{\psi}^0 = \sum_{i=1}^{n} y_i \alpha_i^0 \boldsymbol{x}_i.$$

- Jednak dla wektorów innych niż nośne $\alpha^0 = 0$ stąd optymalna hiperpłaszczyzna to:

$$\sum_{\text{wektory nośne}} y_i \alpha_i^0 \boldsymbol{x}_i \boldsymbol{x} + \boldsymbol{b}^0 = 0.$$

Wyznaczanie $\boldsymbol{b}^0$

- Współczynnik $\boldsymbol{b}$, przesunięcie hiperpłaszczyzny, łatwo wyliczyć na podstawie płaszczyzn wyznaczanych przez wektory nośne:

$$\boldsymbol{b}^0 = \frac{1}{2}[((\boldsymbol{\psi}^0)^T \boldsymbol{x}^1) + ((\boldsymbol{\psi}^0)^T \boldsymbol{x}^{-1})],$$

gdzie $\boldsymbol{x}^{-1}$ i $\boldsymbol{x}^1$ są dowolnymi wektorami nośnymi klas -1 i 1.

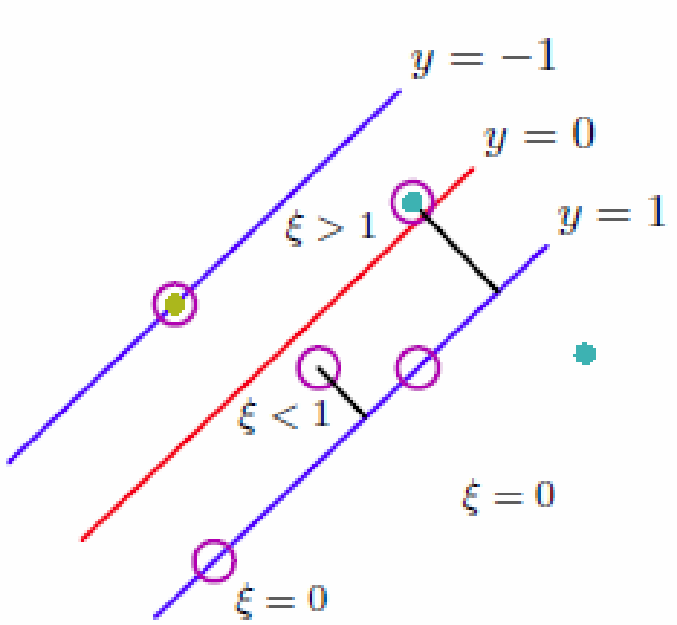
Rozluźnienie warunków

- Nawet po wprowadzeniu funkcji jądrowych rozwiązanie zadania może nie istnieć, ze względu na brak możliwości liniowego odseparowania klas.
- Przekształcona postać zadania umożliwia znalezienie rozwiązania przybliżonego:

$$L(\boldsymbol{\psi}, \boldsymbol{b}, \alpha, \mu) = \frac{1}{2}(\boldsymbol{\psi} * \boldsymbol{\psi}) + C \sum_{i=1}^{N} \xi_i + \\ - \sum_{i=1}^{N} \alpha_i (y_i((\boldsymbol{x}_i * \boldsymbol{\psi}) + b) - 1 + \xi_i) - \sum_{i=1}^{N} \mu_i \xi_i.$$

- Współczynniki ξ_i rozluźniają nierówności ograniczające dla wybranych wektorów x_i, a μ_i są mnożnikami Lagrange.

Przykład

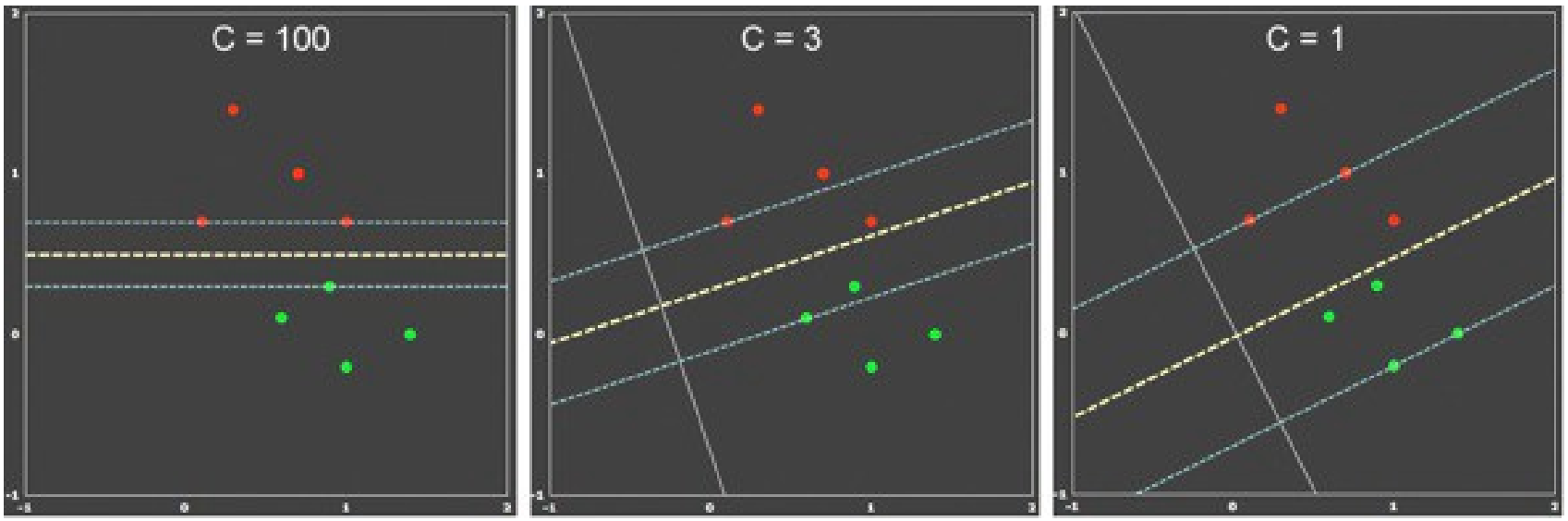


Rysunek 7: Przykłady różnych ξ_i [Bishop, 2006]

Współczynnik C

- Współczynnik C określa wymiar kary za rozluźnienie warunków rozwiązania optymalnego.
- Pozwala sterować stosunkiem szerokości marginesu między klasami, czyli zdolnością do generalizacji, a liczbą błędów w zbiorze uczącym.
- Ponieważ jego dobór nie jest intuicyjny pojawiły się rozwiązania które próbują zastąpić go innymi parametrami lub całkowicie wyeliminować.

Wpływ współczynnika C



Rysunek 8: Wpływ współczynnika C na margines [Mandot, 2017]

Jednoklasowa maszyna wektorów nośnych

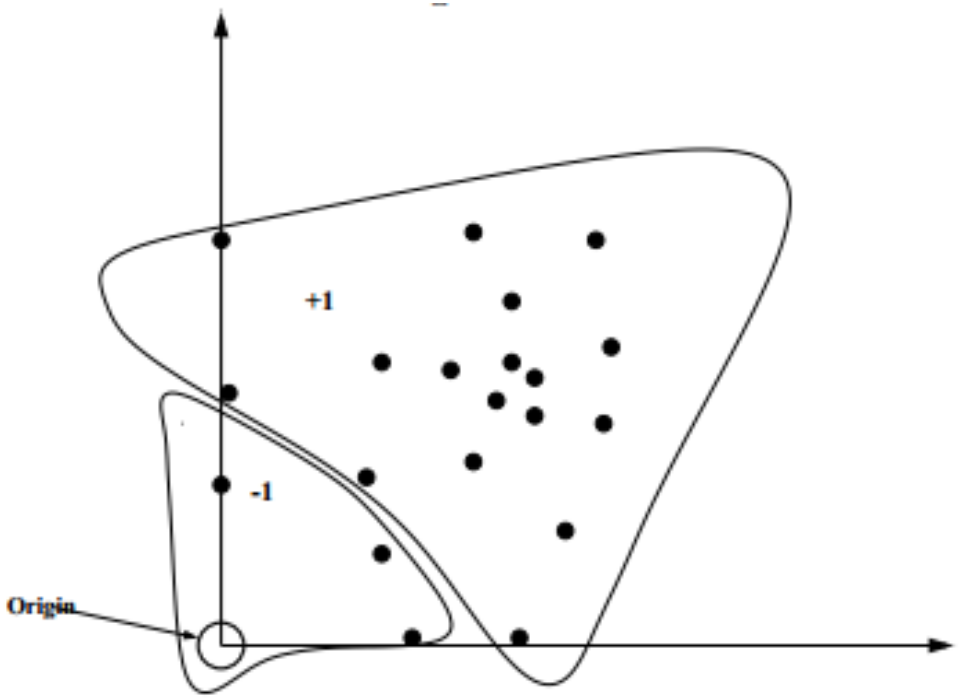
- Maszyna wektorów nośnych jest klasyfikatorem binarnym, który maksymalizuje odległość między dwoma klasami.
- W celu wytrenowania SVM musimy znać przykłady uczące reprezentujące obie klasy.
- Jednoklasowa maszyna wektorów nośnych (*One-class SVM*) oparta jest na jednej klasie, a maksymalizuje odległość między klasą, a początkiem układu odniesienia.

Zasada działania

- Traktujemy zbiór uczący jako jedną klasę.
- Po transformacji danych do przestrzeni jądra traktujemy początek układu odniesienia jako drugą klasę, reprezentowaną przez pojedynczą obserwację.
- W ten sposób oddzielamy przestrzeń S zajmowaną przez zbiór uczący od całej reszty przestrzeni danych $\bar{S}$.
- Funkcja klasyfikacyjna przyjmuje postać:

$$f(x) = \begin{cases} +1 & x \in S \\ -1 & x \in \bar{S} \end{cases}$$

Ilustracja



Rysunek 9: Powstały podział może spowodować, że część punktów uczących nie będzie poprawnie rozpoznawana, ale także odrzuca wszystkie punkty leżące poza zbiorem S [Manevitz and Yousef, 2002]

Uczenie klasyfikatora

- Znalezienie rozwiązania wymaga minimalizacji funkcji:

$$\frac{1}{2}(\boldsymbol{\psi} * \boldsymbol{\psi}) + \frac{1}{nv}\sum_{i=1}^{n}\xi_i - \boldsymbol{b},$$

  przy założeniach $(\forall i)\,(\boldsymbol{\psi} * \Phi(\boldsymbol{x}_i)) \geqslant \boldsymbol{b} - \xi_i$ i $\xi_i \geqslant 0$,
  gdzie Φ jest funkcją jądra, a v mnożnikiem kary.
- W praktyce minimalizuje się funkcję Lagrange'a analogicznie jak dla zadania binarnego.

Rozluźnienie warunków dla maszyny jednoklasowej

- Dla maszyny jednoklasowej modyfikujemy zadanie w sposób zbliżony do modelu binarnego

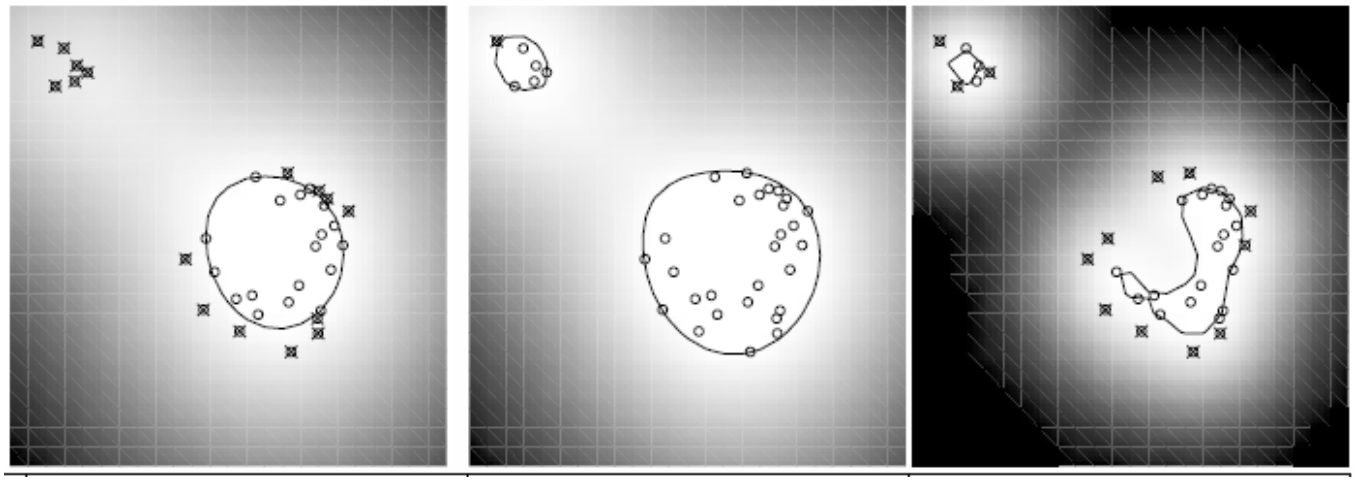
$$L(\boldsymbol{\psi}, \boldsymbol{b}, \alpha, \mu) = \frac{1}{2}(\boldsymbol{\psi} * \boldsymbol{\psi}) + \frac{1}{nv}\sum_{i=1}^{n}\xi_i + \\ - \sum_{i=1}^{n}\alpha_i(y_i((\boldsymbol{x}_i * \boldsymbol{\psi}) + b) - 1 + \xi_i) - \sum_{i=1}^{n}\mu_i\xi_i.$$

- Współczynniki ξ_i rozluźniają nierówności ograniczające dla wybranych wektorów x_i, a μ_i są mnożnikami Lagrange.

Współczynnik v

- Współczynnik v pozwala modelować kształt obszaru S.
- Duża wartość zmniejsza obszar wyznaczany przez klasę.
- Mała wartość pozwala uwzględnić podczas modelowania obserwacje odstające.

Znaczenie parametrów



Rysunek 10: Przykłady działania dla $\gamma = 0.5$, $v = 0.5$; $\gamma = 0.5$, $v = 0.1$; $\gamma = 0.1$, $v = 0.5$ [Schölkopf et al., 2001]

- W pierwszym przypadku v jest zbyt duże, aby uwzględnić wpływ odległych obserwacji.
- W ostatnim przypadku γ jest zbyt małe, aby uwzględnić obserwacje na krawędzi.

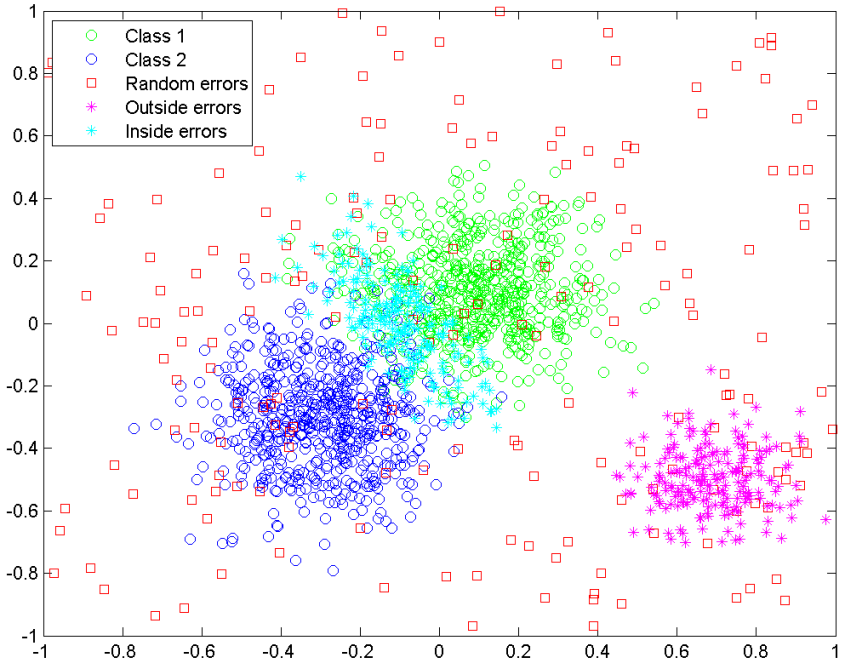
Wykrywanie anomalii

- One–class SVM pozwala wykrywać anomalie rozumiane jako obiekty nie należące do modelowanej klasy.
- Anomaliami są wszystkie obserwacje, dla których funkcja klasyfikująca zwraca -1.

$$f(x) = \text{sgn}((\boldsymbol{\psi} \cdot \phi(\boldsymbol{x}_i)) - \rho) = \text{sgn}(\sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) - \boldsymbol{b})$$

- W przypadku zadań wieloklasowych należy zastosować odrzucanie dla każdej z klasy osobno.

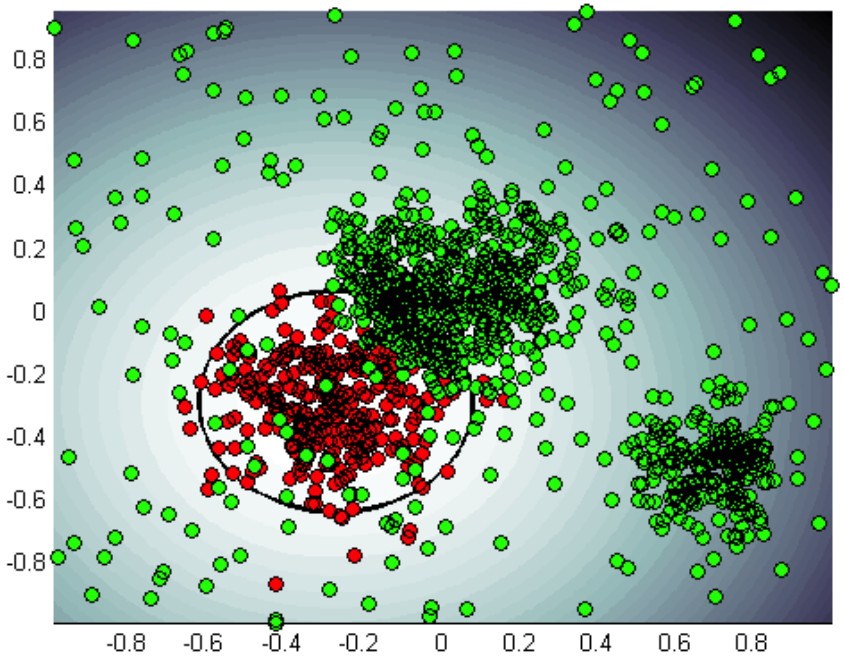
Zadanie klasyfikacji z zakłóceniami



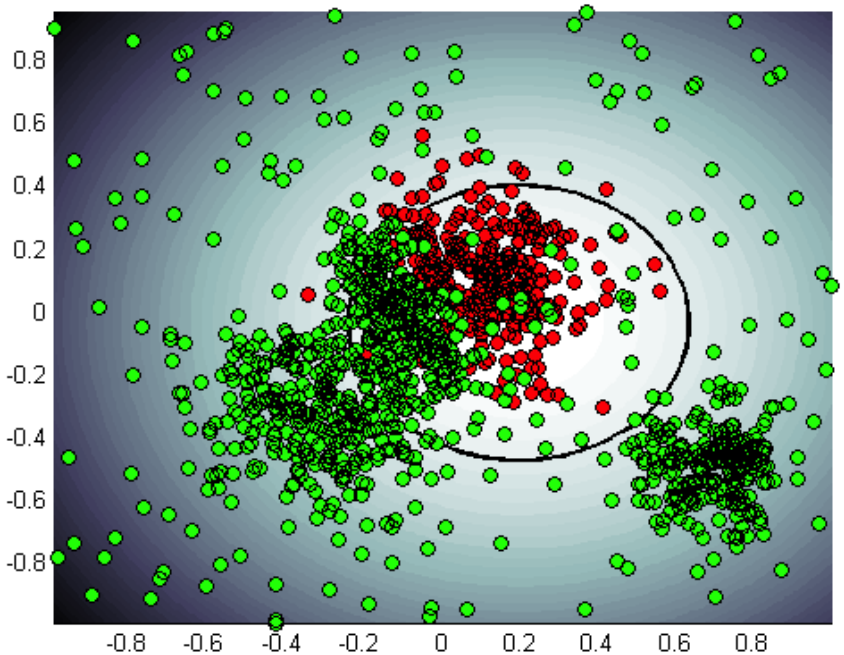
Rysunek 11: Środowisko testowe [Homenda et al., 2014]

- Do zadania rozróżniania dwóch klas dodano następujące zakłócenia
  - losowy szum,
  - punty leżące między klasami,
  - trzecią klasę.

Działanie klasyfikatorów jednoklasowych



Rysunek 12: Elementy zaakceptowane i odrzucone przez lewą klasę



Rysunek 13: Elementy zaakceptowane i odrzucone przez prawą klasę

Wyniki

Tabela 1: Wyniki działania klasyfikatorów one–class SVM

Error	TP	FP	TN	FN	Acc	Pre	Rec	F1
random	581	51	58	8	0.92	0.92	0.99	0.95
outside	560	9	100	31	0.94	0.98	0.95	0.96
inside	562	89	21	25	0.84	0.86	0.96	0.91
all	571	133	180	15	0.84	0.81	0.97	0.89

Lasy izolacyjne

- Las izolacyjny [Liu et al., 2012] to metoda wykrywania anomalii w oparciu o kolekcję drzew izolacyjnych.
- Drzewa izolacyjne to drzewa binarne, których liście zawierają pojedyncze obserwacje lub grupy obserwacji o tych samych wartościach.
- Zastosowana miara izolacji pozwala na ocenę, czy ten węzeł może być uznawany za zbiór obserwacji odstających.

Drzewo izolacyjne

- Niech T będzie wierzchołkiem drzewa izolacyjnego.
  - T jest albo wierzchołkiem zewnętrznym, bez potomków, albo wierzchołkiem wewnętrznym z dokładnie dwoma potomkami (T_l, T_r) i jednym testem.
- Test składa się z atrybutu q i progu p.
  - W ramach testu wszystkie obserwacje z wartościami atrybutu q poniżej progu p trafiają do wierzchołka T_l, pozostałe do wierzchołka T_r.

Tworzenie drzewa izolacyjnego

- Niech $\mathbf{X} = \{x_1, \ldots, x_n\}$ będzie danym zbiorem. Próbka φ instancji $X' \subset X$ jest wykorzystywana do budowy drzewa izolacyjnego (iDrzewo).
- Rekurencyjnie dzielimy próbkę X' poprzez losową selekcję q i p.
- Budowę drzewa kończy się w jednym z dwóch przypadków:
  - w wierzchołku znajduje się tylko jedna obserwacja.
  - wszystkie obserwacje w wierzchołki mają taką samą wartość cechy q.
- Zakładając pesymistyczny rozkład wartości cech (każda wartość jest inna), liczba wierzchołków iDrzewa, które jest drzewem właściwym binarnym wynosi:

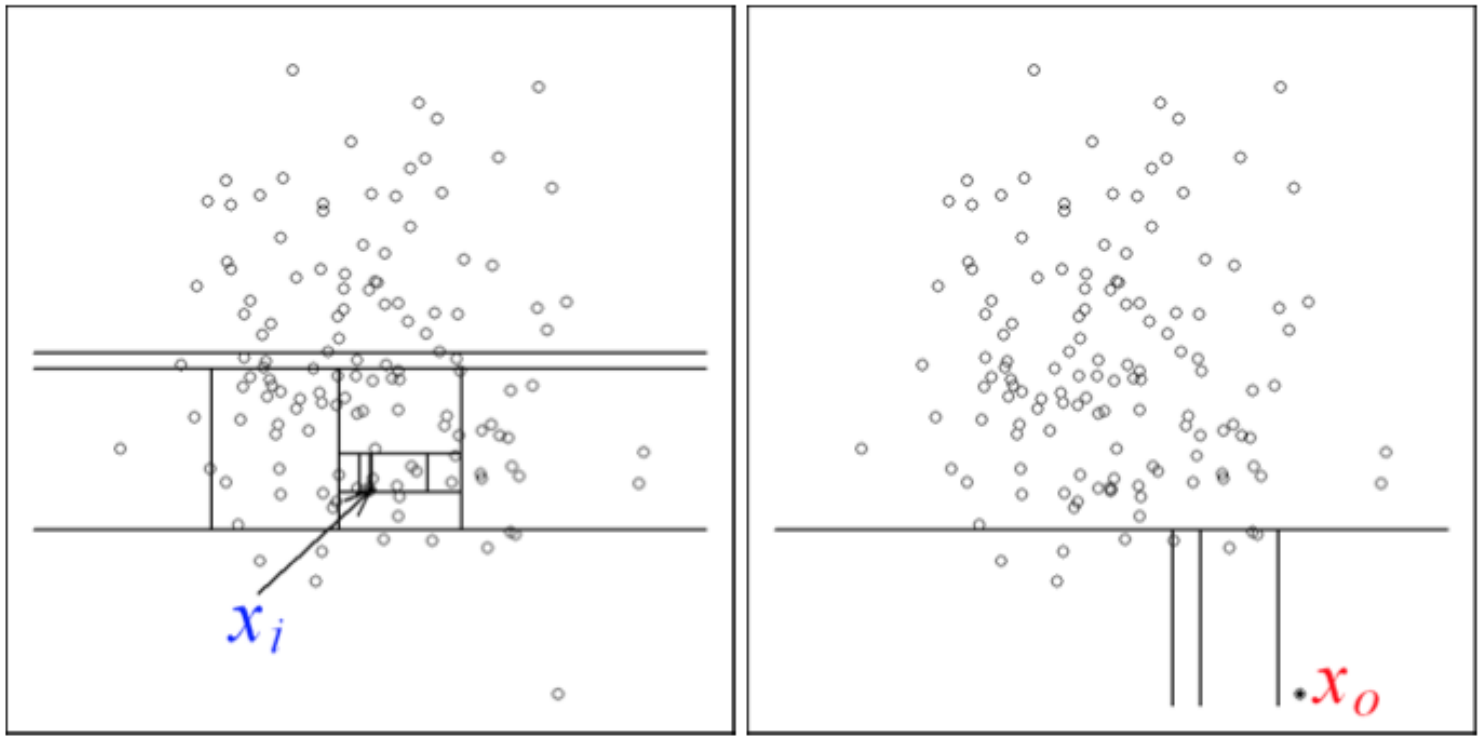
$$2\varphi - 1 = l_i + l_e,$$

gdzie $l_i = \varphi - 1$ a $l_e = \varphi$ więc przyrost pamięciowy jest liniowy.

Zastosowanie iDrzew w detekcji anomalii

- Zagadnienie detekcji anomalii jest rozumiane jako utworzenie rankingu względem stopnia anormalności obserwacji.
- W iDrzewach miarą ta jest oddawana poprzez długość ścieżki do obserwacji.
- Długość ścieżki $h(x)$ dla punktu x jest wyliczana jako liczba krawędzi, którą pokonuje obserwacja x od korzenia do wierzchołka zewnętrznego.
  - krótka ścieżka oznacza wysokie podejrzenie anomalii
  - długa ścieżka oznacza niskie podejrzenie anomalii
- Wynika to z powiązania między gęstością danych, a długością ścieżki.

Długość ścieżki a gęstość



(a) Isolating x_i (b) Isolating x_o

Rysunek 14: Przykład powiązania długiej i krótkiej ścieżki z gęstością danych [Liu et al., 2012]

iLas

- Struktura iLas pozwala na budowę detektora anomalii.
- Detekcja opiera się na dwóch fazach.
  - W fazie uczenia tworzymy kolekcję iDrzew stosując różne podzbiory obserwacji.
  - W fazie ewaluacji, na podstawie iDrzew, wyliczana jest miara anormalności dla każdej z obserwacji.

Faza uczenia

Algorithm 1 : $iForest(X, t, \psi)$

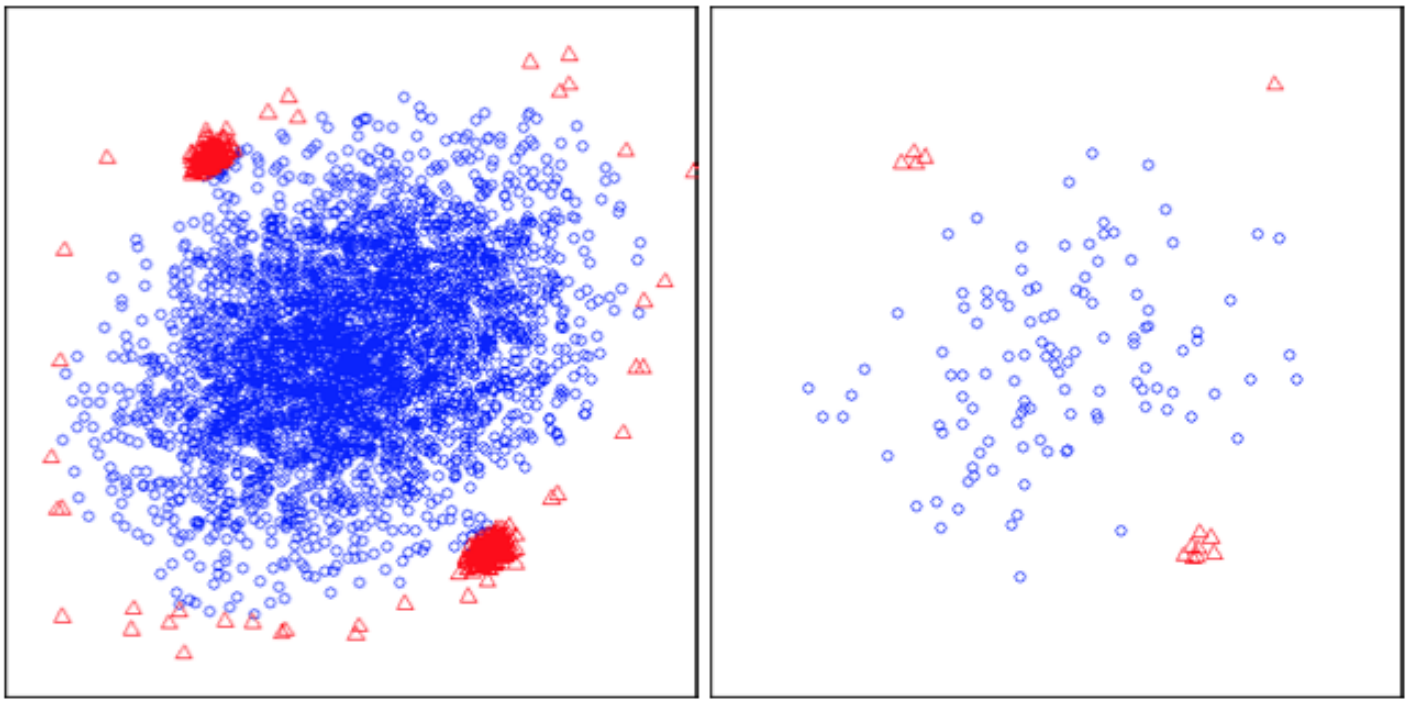
Inputs: X - input data, t - number of trees, ψ - subsampling size
Output: a set of t *iTrees*

1: **Initialize** $Forest$
2: **for** $i = 1$ to t **do**
3: $\quad X' \leftarrow sample(X, \psi)$
4: $\quad Forest \leftarrow Forest \cup iTree(X')$
5: **end for**
6: **return** $Forest$

Rysunek 15: Algorytm tworzenia iLasu [Liu et al., 2012]

- Tworzenie lasu zależy od dwóch parametrów:
  - liczność próbki $\psi = 256$
  - liczba drzew $t = 100$

Rozmiar próbki



(a) Original sample (4096 instances)

(b) Sub-sample (128 instances)

Rysunek 16: Przykład wykrywania anomalii dla wszystkich danych X i próbki X' [Liu et al., 2012]

Faza ewaluacji

Algorithm 3 : $PathLength(x, T, hlim, e)$

Inputs : x - an instance, T - an iTree, $hlim$ - height limit, e - current path length; to be initialized to zero when first called

Output: path length of x

1: **if** T is an external node or $e \geq hlim$ **then**
2: return $e + c(T.size)$ {$c(.)$ is defined in Equation 1}
3: **end if**
4: $a \leftarrow T.splitAtt$
5: **if** $x_a < T.splitValue$ **then**
6: return $PathLength(x, T.left, hlim, e + 1)$
7: **else** $\{x_a \geq T.splitValue\}$
8: return $PathLength(x, T.right, hlim, e + 1)$
9: **end if**

Rysunek 17: Algorytm ewaluacji obserwacji x [Liu et al., 2012]

- Ewaluacja zlicza długość ścieżki.
- Wyliczenia trwają do osiągnięcia węzła zewnętrznego lub głębokośc *hlim*.
- Zwracaną wartością jest długość ścieżki plus dopasowanie $c(T.size)$

Dopasowanie

- IDrzewa mają podobną strukturę do Binary Search Tree (BST).
- Dlatego można odnieść długość uzyskanej ścieżki do średniej długości ścieżki nieudanego wyszukiwania w BST:

$$c(\varphi) = \begin{cases} 2H(\psi - 1) - 2(\psi - 1)/n & \text{dla} \quad \psi > 2 \\ 1 & \text{dla} \quad \psi = 2 \\ 0 & \text{w p.p.} \end{cases}$$

,gdzie $H(x) \sim ln(x) + 0.5772156649$ (liczba harmoniczna)

- $n = |X|$ liczność zbioru.

Wyliczanie anormalności

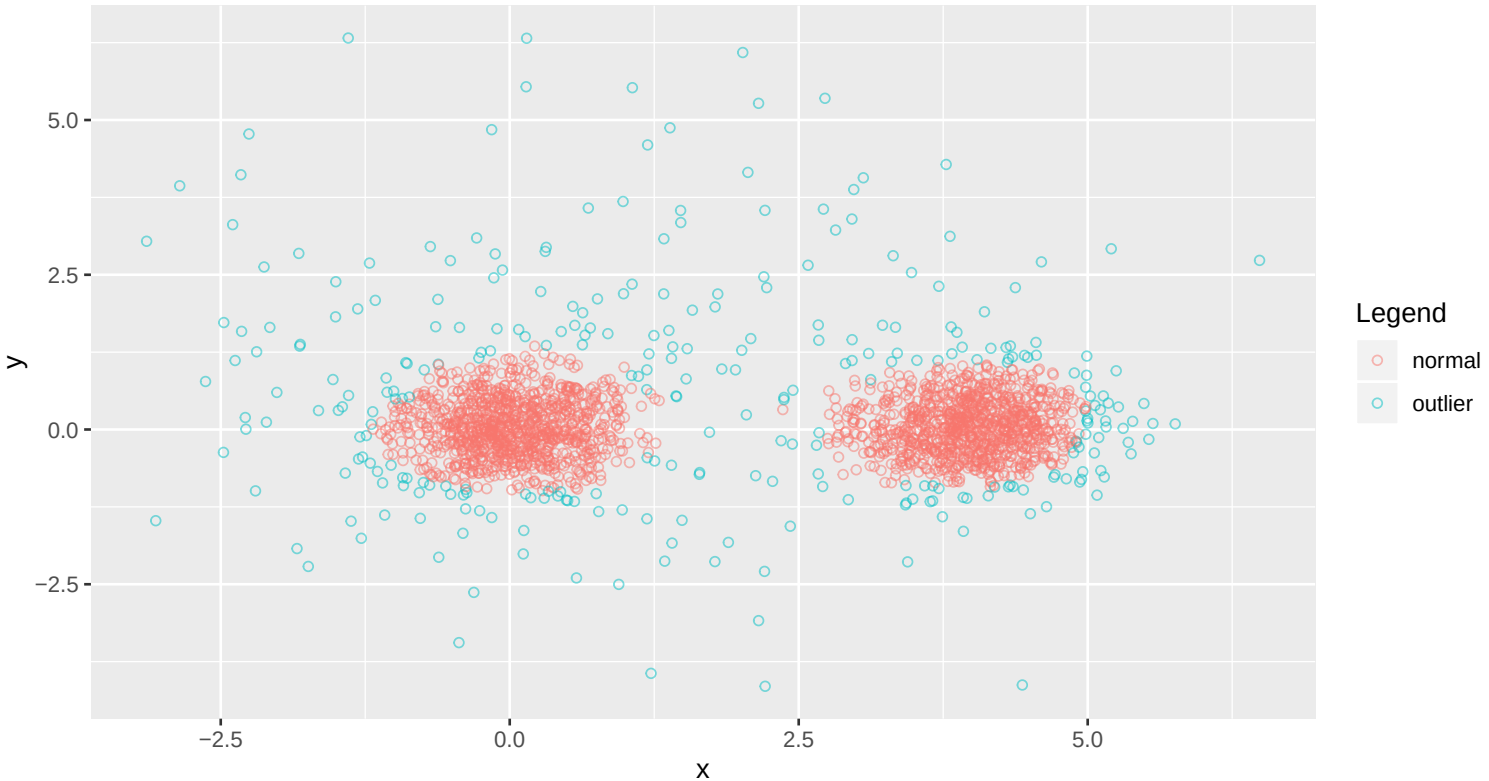
- Dopasowanie możemy wykorzystać do normalizacji średniej wartości $h(x)$.
- Wtedy miara anormalności wyniesie:

$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}},$$

gdzie $E(h(x))$ jest średnią wartością $h(x)$ wyliczoną dla iLasu.

- Gdy $E(h(x)) \to c(\psi)$ $s \to 0.5$, co pozwala dobrać próg dla detekcji anomalii.

Przykład



Rysunek 18: Przykład zastosowania iLasu $s > 0.65$

Wygładzanie danych

- Niektóre typy danych są szczególnie podatne na powstawanie drobnych odchyleń od lokalnej średniej wartości.
- Dotyczy to zwłaszcza szeregów czasowych i obrazów.
- Wygładzenie danych prowadzi do ich większej czytelności.

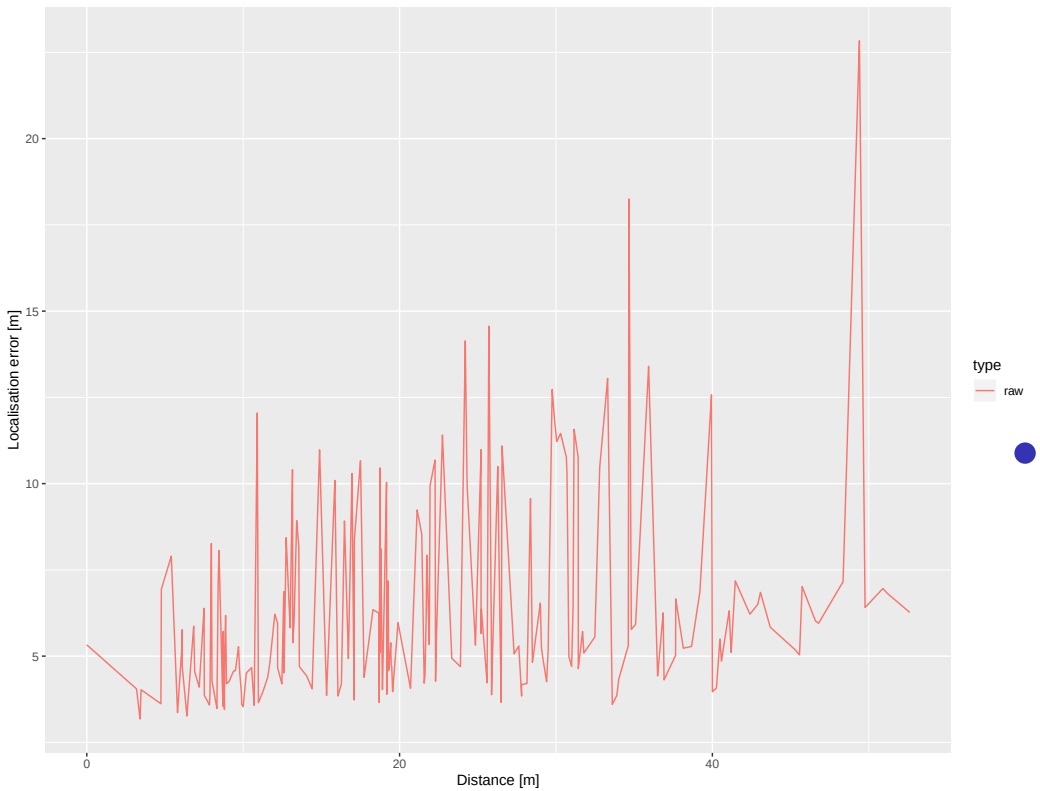
Średnia krocząca

- Średnia krocząca zastępuje odczytaną wartość lokalną wartością średnią.
- Wartość wylicza się jako średnią arytmetyczną z poprzednich n odczytów:

$$SMA = \frac{1}{n} \sum_{i=1}^{n} s_i.$$

- Pierwsze n odczytów nie podlega zmianie

Przykład szeregu czasowego



Rysunek 19: Zmiana błędu z odległością

- Dane pokazują zmianę błędu lokalizacji telefonu komórkowego, wraz z oddaleniem urządzenia namierzającego od początkowej pozycji.
- Ponieważ oddalenie było przeprowadzane skokowo i zachodziły też inne czynniki zakłócające, to dane są mało czytelne.
- Spróbujemy wygładzić wykres.

Wpływ szerokości okna

Rysunek 20: Wpływ wielkości okna na wygładzenie

- Zwiększanie szerokości okna powoduje coraz to wiekszą redukcję pików na wykresie.

Modyfikacje średniej kroczącej

- Standardowa średnia krocząca może być zbyt inwazyjna.
- Dlatego zaproponowano inne sposoby uśredniania, w których wpływ elementów maleje wraz z odległością.
- Ważona średnia ruchoma:

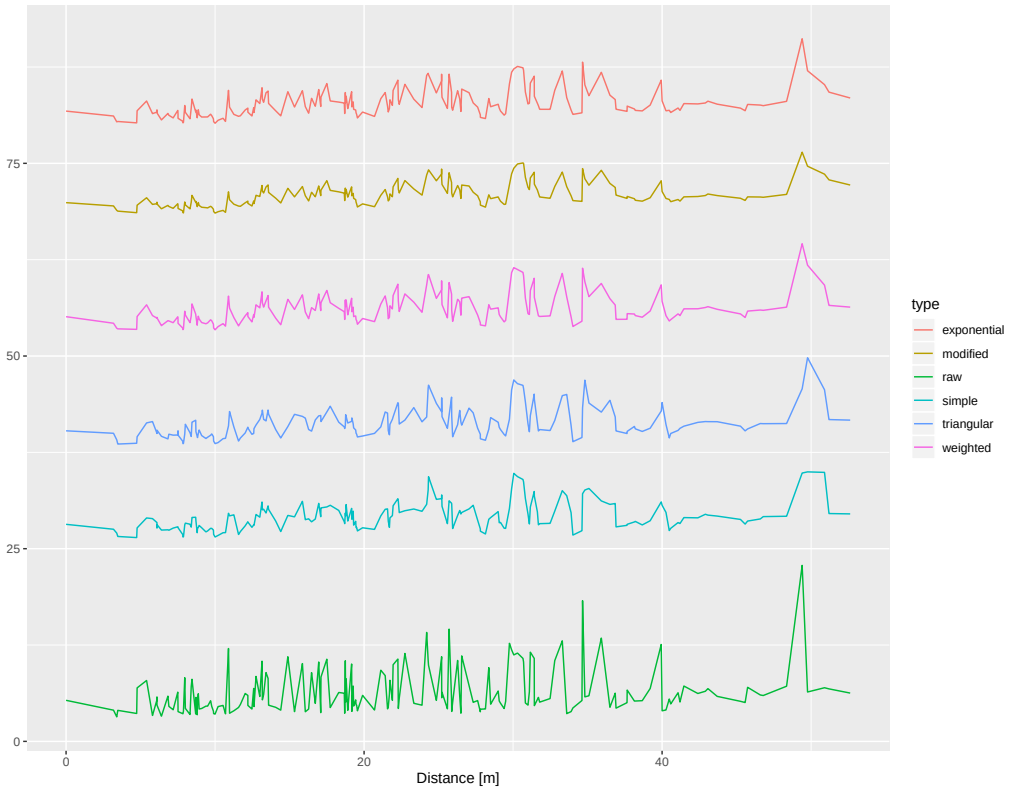
$$WMA = \frac{\sum_{i=1}^{n} i * s_i}{\sum_{i=1}^{n} i}.$$

- Wykładnicza średnia ruchoma:

$$WMA = \frac{\sum_{i=1}^{n} (1-\alpha)^{n-i} * s_i}{\sum_{i=1}^{n} (1-\alpha)^{i}},$$

gdzie $\alpha = \frac{2}{n+1}$.

Zastosowanie różnych rodzajów średnich



Rysunek 21: Wpływ rodzaju średniej na wygładzenie (n=3)

- Zastosowanie innych rodzajów wyliczenia średniej jest mniej inwazyjne.

Przetwarzanie obrazów

- Zaprezentowane podejście, najczęściej w bardziej zaawansowanej formie, stosuje się także do obrazów.
- Traktując obraz jako przestrzeń dwuwymiarową, wygładzamy jego elementy bazując na odpowiednio zdefiniowanym sąsiedztwie.
- Podobnie można wyostrzyć obraz lub dokonać innych modyfikacji, zwiększających jego czytelność.
- Metody wygładzania stosowane dla obrazów nazywane są filtrami.

Filtr bilateralny

- Filtr bilateralny jest filtrem nieliniowym, który wygładza obraz z zachowaniem krawędzi.
- Jest zależny od odległości między pikselami i różnicy w ich jasności.

Działanie filtra bilateralnego

- Filtr bilateralny wylicza nową jasność piksela jako:

$$I_{bilateral}(x) = \frac{\sum_{x_i \in \Omega} I(x_i) f_r(\|I(x_i) - I(x)\|) g_s(\|x_i - x\|)}{\sum_{x_i \in \Omega} f_r(\|I(x_i) - I(x)\|) g_s(\|x_i - x\|)},$$

gdzie $I_{bilateral}$ to obraz po filtracji, I klatka początkowa, x to współrzędne piksela,
Ω to okno filtru, którego środkowym elementem jest x.
g_s - współczynnik wygładzenia zależnej od odległości,
f_r - współczynnik wygładzenia zależnej od intensywności.

Zastosowanie filtra bilateralnego

Rysunek 22: Oryginalny obraz

Rysunek 23: Obraz po filtracji

- Zastosowanie filtra bilateralnego [Tomasi and Manduchi, 1998]

Filtr odszumiający

- Zadaniem filtra odszumiającego jest oczyszczenie obrazu z występujących na nim szumów i zakłóceń.
- Zakłócenie definiuje się formułą:

$$v(i) = u(i) + n(i),$$

gdzie $v(i)$ to aktualna wartość piksela, $u(i)$ właściwa wartość piksela, a $n(i)$ to zakłócenie.

- Istnieje wiele implementacji filtrów odszumiajacych.
  - Gauss filtering,
  - Anisotropic filtering,
  - Total variation,
  - Neighborhood filtering,
  - Non-Local Means Denoising.

Przykład działania algorytmu redukującego szumu

Rysunek 24: Zaszumiony obraz

Rysunek 25: Zastosowanie filtra

- Zastosowanie filtra redukcji szumów (filtr adaptacyjnego wygładzania) [Polzehl and Tabelow, 2007]

Non-Local Means Denoising

- Filtr Non-Local Means Denoising bazuje na średniej wartości pikseli na całym obrazie, a nie tylko w otoczeniu poprawianego piksela.
- Dla pikseli p i q wartość filtra wyliczamy jako:

$$u(q) = \frac{\sum_{x_i \in \Omega} v(q) f(p,q)}{\sum_{x_i \in \Omega} f(p,q)},$$

- gdzie u(q) to wartość po filtracji, $v(q)$ wartość początkowa, a Ω to obszar obrazu.
- f - funkcja wagowa.

Funkcja wagowa

- Rolą funkcji wagowej jest określenie bliskości relacji między obrazem w punkcie p i q
- Może przyjmować wartość:

$$f(p,q) = e^{-\frac{\|B(q)-B(p)\|^2}{h^2}},$$

gdzie

$$B(p) - \frac{1}{\|R(p)\|} \sum_{q \in R(p)} v(i).$$

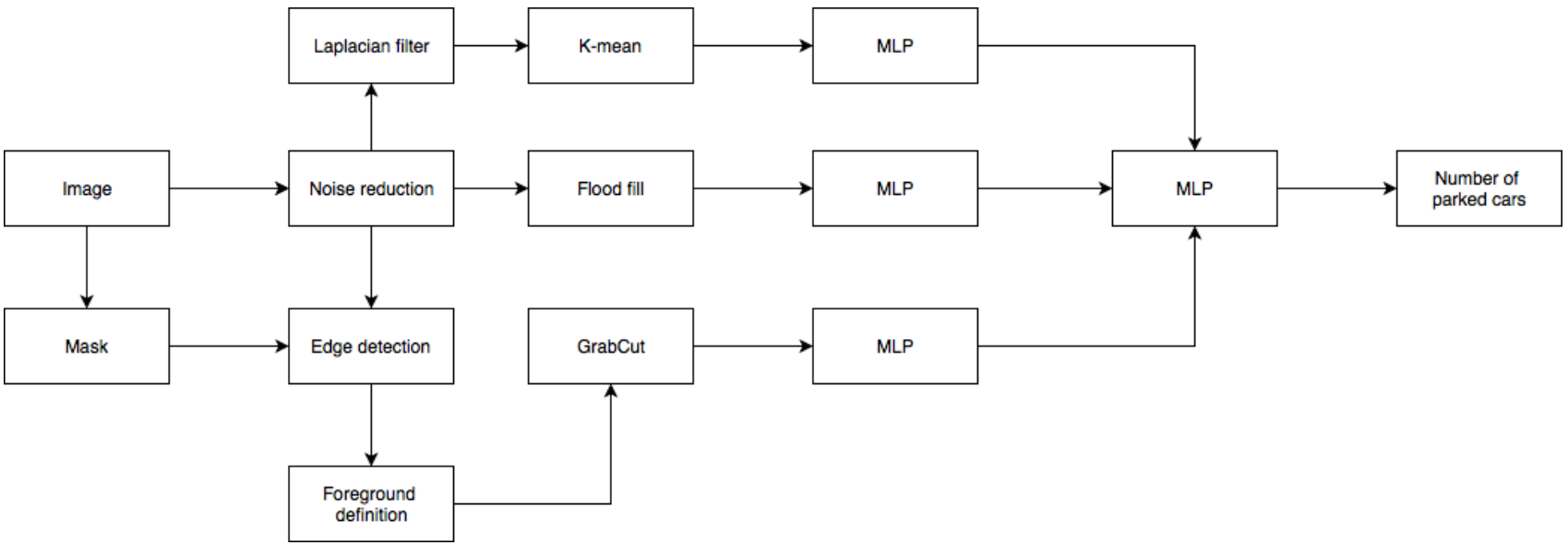
- $R(p) \in \Omega$ to kwadratowy obszar otaczający p, a $\|R(p)\|$ to pole tego obszaru
- $h > 0$ jest współczynnikiem filtrowania.

Zastosowanie filtrów odszumiających

Rysunek 26: Porównanie filtrów odszumiających. Zaszumiony obraz, Gauss filtering, Anisotropic filtering, Total variation, Neighborhood filtering, Non-Local Means Denoising [Buades et al., 2005].

Zastosowanie filtrów w eksperymencie

- Wstępne przetwarzanie obrazu, poprzez zastosowanie szeregu filtrowanie jest konieczne przed zastosowaniem uczenia maszynowego.
- Dobór rodzajów filtrów i ich implementacji zależy od stosowanych algorytmów i rozwiązywanego zadania.



Rysunek 27: Schemat przetwarzania danych w zadaniu predykcji liczby wolnych miejsc parkingowych [Bukowski et al., 2019].

Bibliografia I

[Bishop, 2006] Bishop, C. M. (2006).

Bishop, Pattern recognition and machine learning.

Springer.

[Buades et al., 2005] Buades, A., Coll, B., and Morel, J. . (2005).

A non-local algorithm for image denoising.

In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65 vol. 2.

[Bukowski et al., 2019] Bukowski, M., Luckner, M., and Kunicki, R. (2019).

Estimation of free space on car park using computer vision algorithms.

In Szewczyk, R., Zielinski, C., and Kaliczynska, M., editors, *Automation 2019 - Progress in Automation, Robotics and Measurement Techniques, outcomes of the international conference AUTOMATION 2019, 27-29 March, 2019, Warsaw, Poland*, volume 920 of *Advances in Intelligent Systems and Computing*, pages 316–325. Springer.

Bibliografia II

[Homenda et al., 2014] Homenda, W., Luckner, M., and Pedrycz, W. (2014).

Classification with rejection based on various svm techniques.

In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 3480–3487.

[Jankowski, 2003] Jankowski, N. (2003).

Ontogeniczne sieci neuronowe: o sieciach zmieniających swoją strukturę.

Akademicka Oficyna Wydawnicza EXIT.

[Koronacki and Ćwik, 2006] Koronacki, J. and Ćwik, J. (2006).

Statystyczne systemy uczące się.

Akademicka Oficyna Wydawnicza EXIT.

[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.

Bibliografia III

[Liu et al., 2012] Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2012).
Isolation-based anomaly detection.
ACM Trans. Knowl. Discov. Data, 6(1).

[Mandot, 2017] Mandot, P. (2017).
What is the significance of c value in support vector machine?

[Manevitz and Yousef, 2002] Manevitz, L. M. and Yousef, M. (2002).
One-class svms for document classification.
J. Mach. Learn. Res., 2:139–154.

[Osowski, 2006] Osowski, S. (2006).
Sieci neuronowe do przetwarzania informacji.
Oficyna Wydawnicza Politechniki Warszawskiej.

[Polzehl and Tabelow, 2007] Polzehl, J. and Tabelow, K. (2007).
Adaptive smoothing of digital images: The r package adimpro.
Journal of Statistical Software, Articles, 19(1):1–17.

Bibliografia IV

[Schölkopf et al., 2001] Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., and Williamson, R. C. (2001).

Estimating the support of a high-dimensional distribution.

Neural Comput., 13(7):1443–1471.

[Sharma, 2019] Sharma, S. (2019).

Kernel trick in svm.

[Taylor, 1995] Taylor, J. R. (1995).

Wstęp do analizy błędu pomiarowego.

[Tomasi and Manduchi, 1998] Tomasi, C. and Manduchi, R. (1998).

Bilateral filtering for gray and color images.

In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 839–846.

[UC Business Analytics R Programming Guide, 2000] UC Business Analytics R Programming Guide (2000).

Support vector machine.

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.