Nadzorowane algorytmy minimalno-odległościowe: NM, kNN

Dr inż. Urszula Libal

Nadzorowane algorytmy minimalno-odległościowe

SPIS TREŚCI:

- ❑ Nearest Mean (NM)
- ❑ Nearest Neighbor (NN)
- ❑ k-Nearest Neighbors (kNN)

1. Nadzorowane algorytmy minimalno-odległościowe

— *nearest mean* (NM) - najbliższa średnia,

— *nearest neighbor* (NN) - najbliższy sąsiad,

— k *nearest neighbors* (kNN) - k najbliższych sąsiadów.

Algorytmy oparte o ciągi uczące:

1. podejście globalne (NM),
2. lokalne (NN),
3. pośrednie (kNN).

2. Klasyfikator najbliższa średnia (NM)

W nadzorowanej wersji algorytmu, wyliczamy **centra klas** na podstawie ciągów uczących: w klasie 1 $\left\{\mathbf{x}_j^{(1)}\right\}_{j=1}^{N_1}$ oraz w klasie 2 $\left\{\mathbf{x}_j^{(2)}\right\}_{j=1}^{N_2}$:

$$\mu_1 = (\mu_{11}, \mu_{21}, \ldots, \mu_{D1}) = \frac{1}{N_1} \sum_{j=1}^{N_1} \mathbf{x}_j^{(1)}, \tag{1}$$

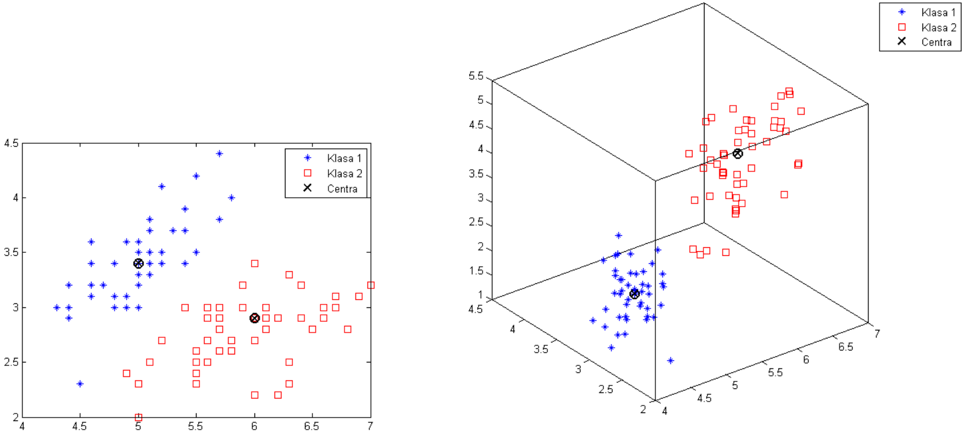
$$\mu_2 = (\mu_{12}, \mu_{22}, \ldots, \mu_{D2}) = \frac{1}{N_2} \sum_{j=1}^{N_2} \mathbf{x}_j^{(2)}. \tag{2}$$

Obraz $\mathbf{x} = (x_1 x_2, \ldots, x_D)$ jest klasyfikowany do tej klasy, z której średnią dzieli go mniejsza odległość w ustalonej metryce

$$\Psi_{NM}(\mathbf{x}) = \begin{cases} 1, & \text{gdy } ||\mathbf{x} - \mu_1|| < ||\mathbf{x} - \mu_2||, \\ 2, & \text{w przeciwnym wypadku.} \end{cases} \tag{3}$$

Dla metryki euklidesowej warunek upraszcza się do

$$\Psi_{NM}(\mathbf{x}) = \begin{cases} 1, & \text{gdy } \sum_{i=1}^{D} (x_i - \mu_{i1})^2 < \sum_{i=1}^{D} (x_i - \mu_{i2})^2, \\ 2, & \text{w przeciwnym wypadku.} \end{cases} \tag{4}$$



a) b)

Rysunek 1. Klasyfikator najbliższa średnia: (a) widok 2D, (b) widok 3D.

Nazwa metryki	Wzór $d(\mathbf{x},\mathbf{y}) = \|\|\mathbf{x}-\mathbf{y}\|\|$
euklidesowa	$d(\mathbf{x},\mathbf{y}) = \left\{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})\right\}^{\frac{1}{2}} = \sqrt{\sum_{i=1}^{D}(x_i-y_i)^2}$
taksówkowa, Manhattan	$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{D}\lvert x_i-y_i\rvert$
Czebyszewa	$d(\mathbf{x},\mathbf{y}) = \max_{i=1,\ldots,D}\lvert x_i-y_i\rvert$
Canberry [1]	$d(\mathbf{x},\mathbf{y}) = \frac{1}{D}\sum_{i=1}^{D}\frac{\lvert x_i-y_i\rvert}{x_i+y_i}$
Lance'a-Williamsa [1]	$d(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{D}\lvert x_i-y_i\rvert}{\sum_{i=1}^{D}(x_i+y_i)}$

3. Klasyfikator najbliższy sąsiad (NN)

Obliczamy $N = N_1 + N_2$ odległości $||\mathbf{x} - \mathbf{x}_j^{(k)}||$ między klasyfikowanym obrazem $\mathbf{x}$ a wektorami cech $\mathbf{x}_j^{(k)}$ $(j = 1, 2, \ldots, N_k)$ z ciągów uczących dla obu klas, $k = 1, 2$.

Klasyfikujemy obraz $\mathbf{x}$ do klasy obrazu z ciągu uczącego, który jest położony najbliżej obrazu $\mathbf{x}$, czyli klasyfikujemy obraz do klasy pochodzenia jego najbliższego sąsiada.

Klasyfikację za pomocą algorytmu najbliższy sąsiad można formalnie zapisać następująco

$$\Psi_{NN}(\mathbf{x}) = \begin{cases} 1, & \text{gdy } \exists_i \forall_j \, ||\mathbf{x} - \mathbf{x}_i^{(1)}|| < ||\mathbf{x} - \mathbf{x}_j^{(2)}||, \\ 2, & \text{w przeciwnym wypadku.} \end{cases} \quad (5)$$

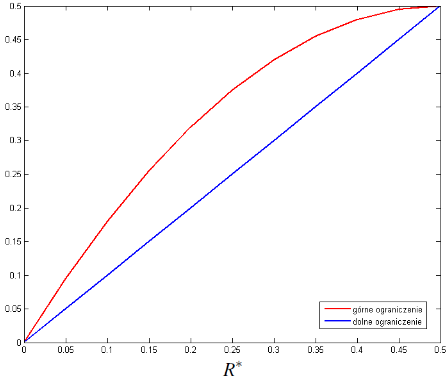
Cover i Hart [2] opublikowali w 1967 roku **oszacowanie ryzyka klasyfikatora najbliższy sąsiad** R_{NN} za pomocą ryzyka R^* optymalnego algorytmu Bayesa w asymptotycznym przypadku, gdy długość ciągu uczącego $N \to \infty$

$$R^* \leq R_{NN} \leq R^* \left(2 - \frac{M}{M-1} R^*\right), \tag{6}$$

M to liczba klas.

W przypadku problemu dwuklasowego ($M = 2$) otrzymujemy oszacowanie

$$R^* \leq R_{NN} \leq 2R^* \left(1 - R^*\right). \tag{7}$$



Rysunek 2. Górne i dolne ograniczenie ryzyka klasyfikatora najbliższy sąsiad dla dwóch klas - wzór (7).

4. Klasyfikator k-najbliższych sąsiadów (kNN)

Zamiast kierować się klasą tylko jednego (najbliższego) sąsiada, można decyzję oprzeć na informacji o klasach pochodzenia k-najbliższych sąsiadów.

Aby uniknąć sytuacji remisowych, najprościej jest przyjmować k **nieparzyste**. Wtedy “wygrywa” klasa, z której pochodzi większość sąsiadów z najbliższego otoczenia badanego obrazu.

Jak dobierać liczbę k sąsiadów?

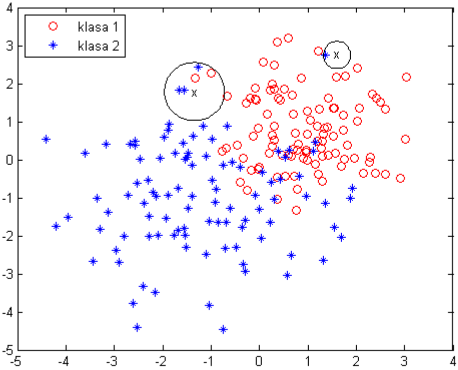
Liczba k musi być:

— na tyle duża, by redukować wrażliwość algorytmu na zakłócenia

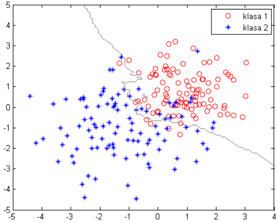
— na tyle mała, by nie wybierać sąsiadów mocno osadzonych w innych klasach

— trzeba także uwzględnić długości ciągów uczących

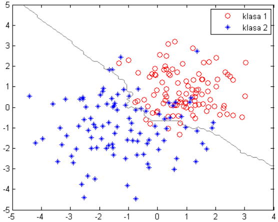
— można zastosować procedurę kroswalidacji



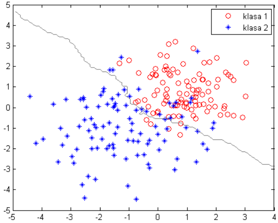
Rysunek 3. Sąsiedztwo.



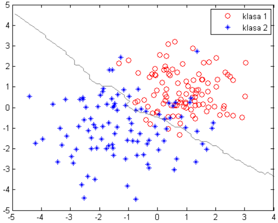
a)



b)



c)

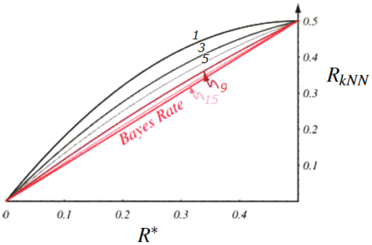


d)

Rysunek 4. Klasyfikator k-najbliższych sąsiadów: (a) $k = 9$, (b) $k = 19$, (c) $k = 29$, (d) $k = 59$.

Ryzyko klasyfikatora k-najbliższych sąsiadów R_{kNN} dla problemu dwuklasowego dąży do ryzyka Bayesa R^* przy liczbie k rosnącej do nieskończoności

$$\lim_{k \to \infty} R_{kNN} = R^*. \tag{8}$$



Rysunek 5. Ryzyko klasyfikatora kNN.