Машинное обучение на основе анализа выпуклых оболочек классов

Анатолий Немирко

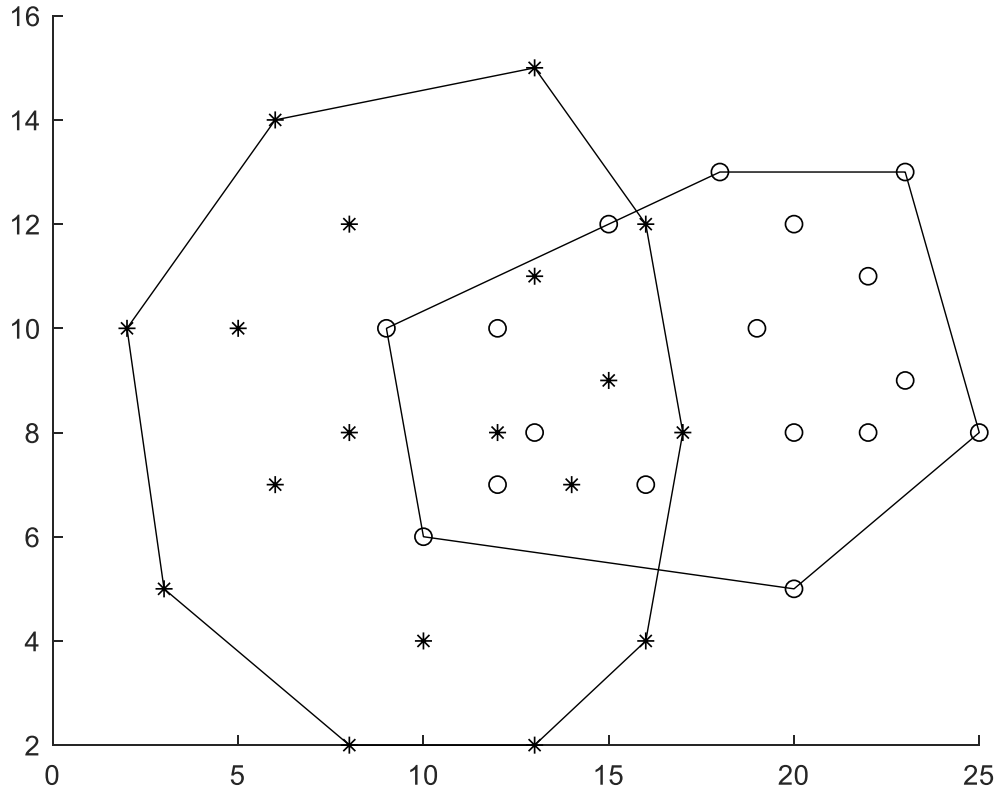
Department of Biotechnical Systems
Saint Petersburg Electrotechnical University "LETI",
Saint Petersburg, Russian Federation

apn-bs@yandex.ru

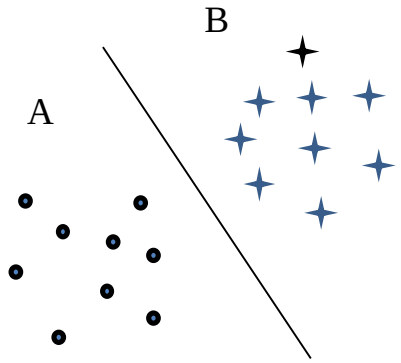
План доклада

- Введение
- Новый метод оценки близости тестовой точки к выпуклой оболочке класса
  - Глубина проникновения
  - DISTANCE алгоритм
  - Выбор направления проекции
- Облегченный классификатор
- Эксперименты
- Заключение

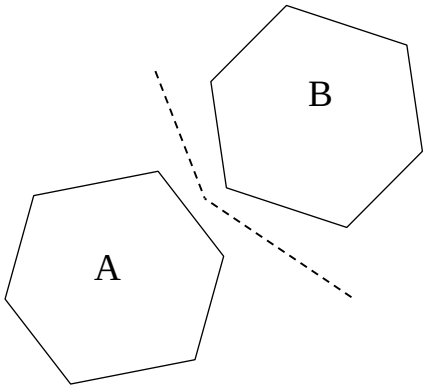
Представление классов в виде выпуклых оболочек



Introduction



Statistical approach



Convex hull approach

1. Support vector machine liner classifier (SVM)
2. Nearest convex hull classifier (NCH)

Для метрических задач отображения, классификации, кластеризации и др. важно измерять расстояния от точки до множества, между двумя множествами, между несколькими множествами.

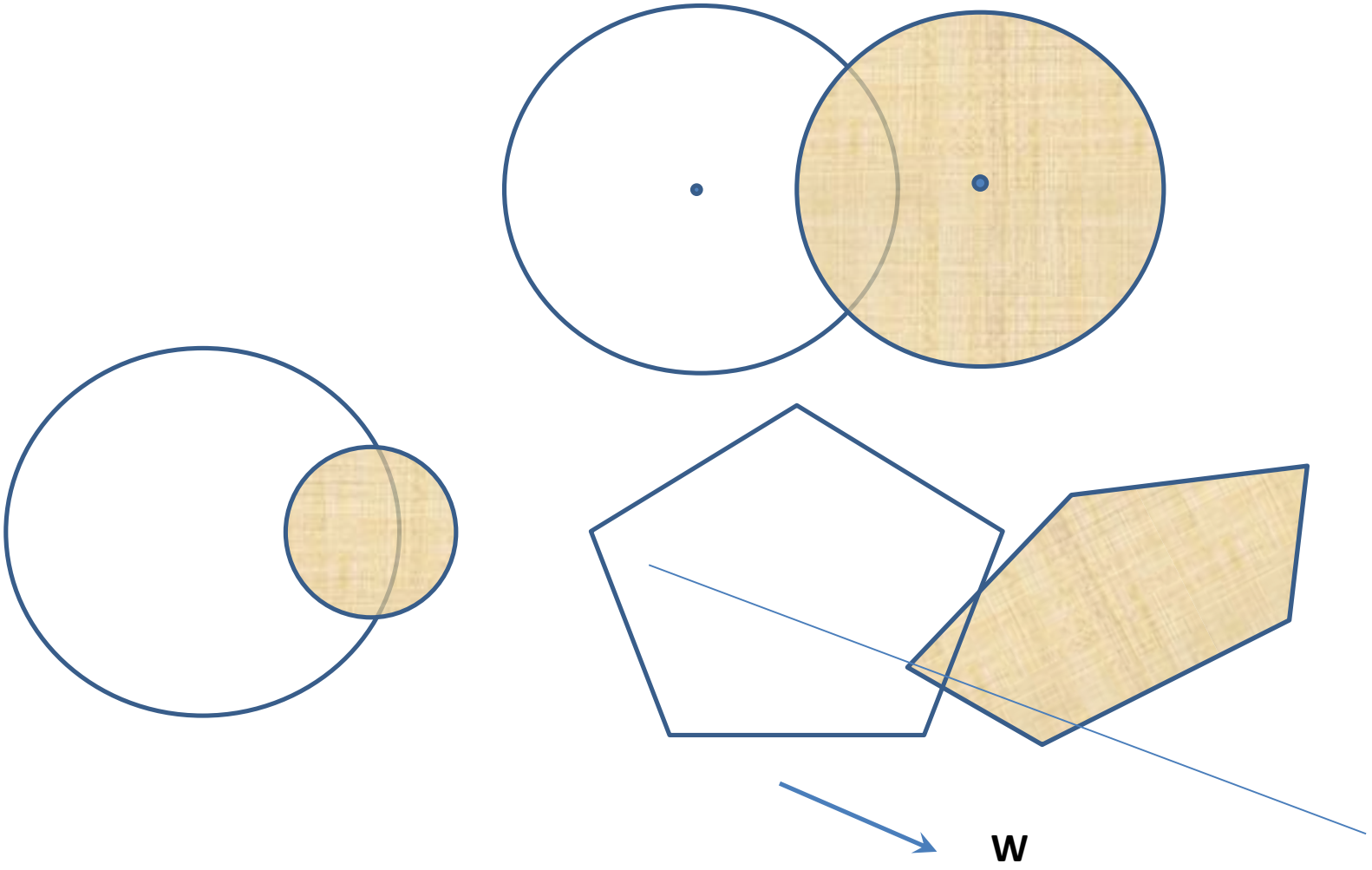
Для отображения точек многомерного пространства на плоскости для определения плоскости нужно найти 2 перпендикулярных вектора. Назовем их весовыми. В случае перспективы в дальнейшем решения задачи классификации критерии для определения этих весовых векторов могут быть следующими:

1. На основе максимизации расстояния между центрами классов;
2. На основе максимизации критерия Фишера (использование ковариационных данных);
3. На основе максимизации расстояния между выпуклыми оболочками классов;
4. На основе применения других критериев.

В случае применения 3-го критерия и линейно непересекающихся классов задача может быть легко решена с помощью технологии SVM или имеющихся других средств.

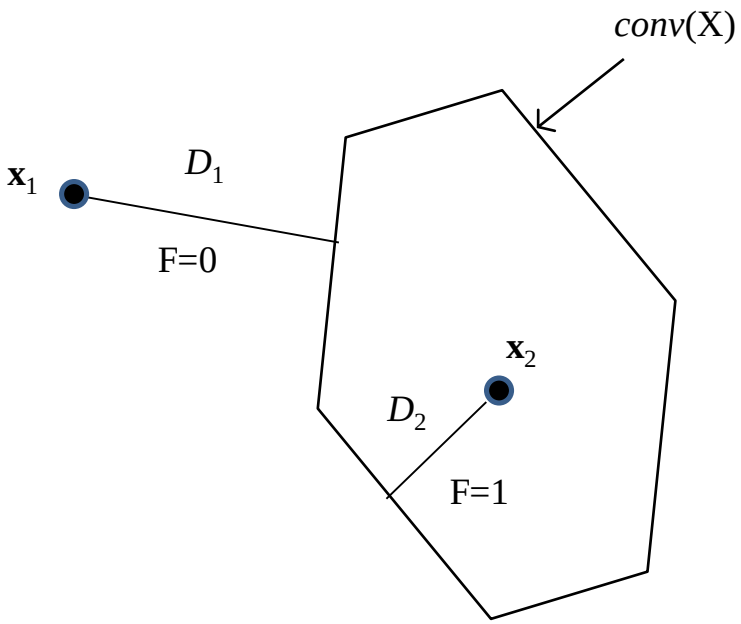
При непересекающихся классах задача **намного сложнее**. SVM за счет влияния ошибок классификации не всегда правильно (и всегда неточно) определяет расстояние между выпуклыми оболочками классов или расстояние от точки до выпуклой оболочки.

Проблема неразделимых классов

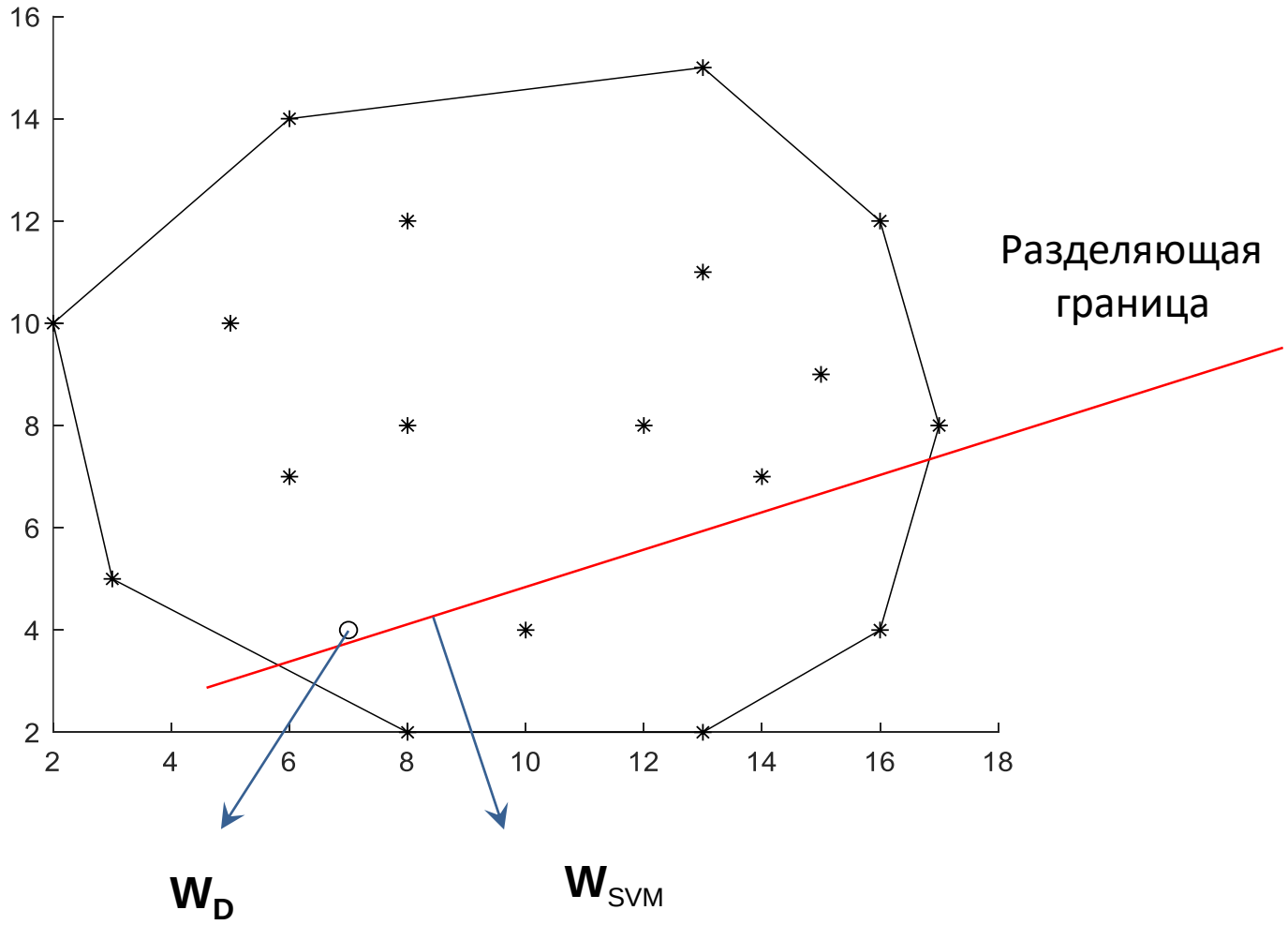


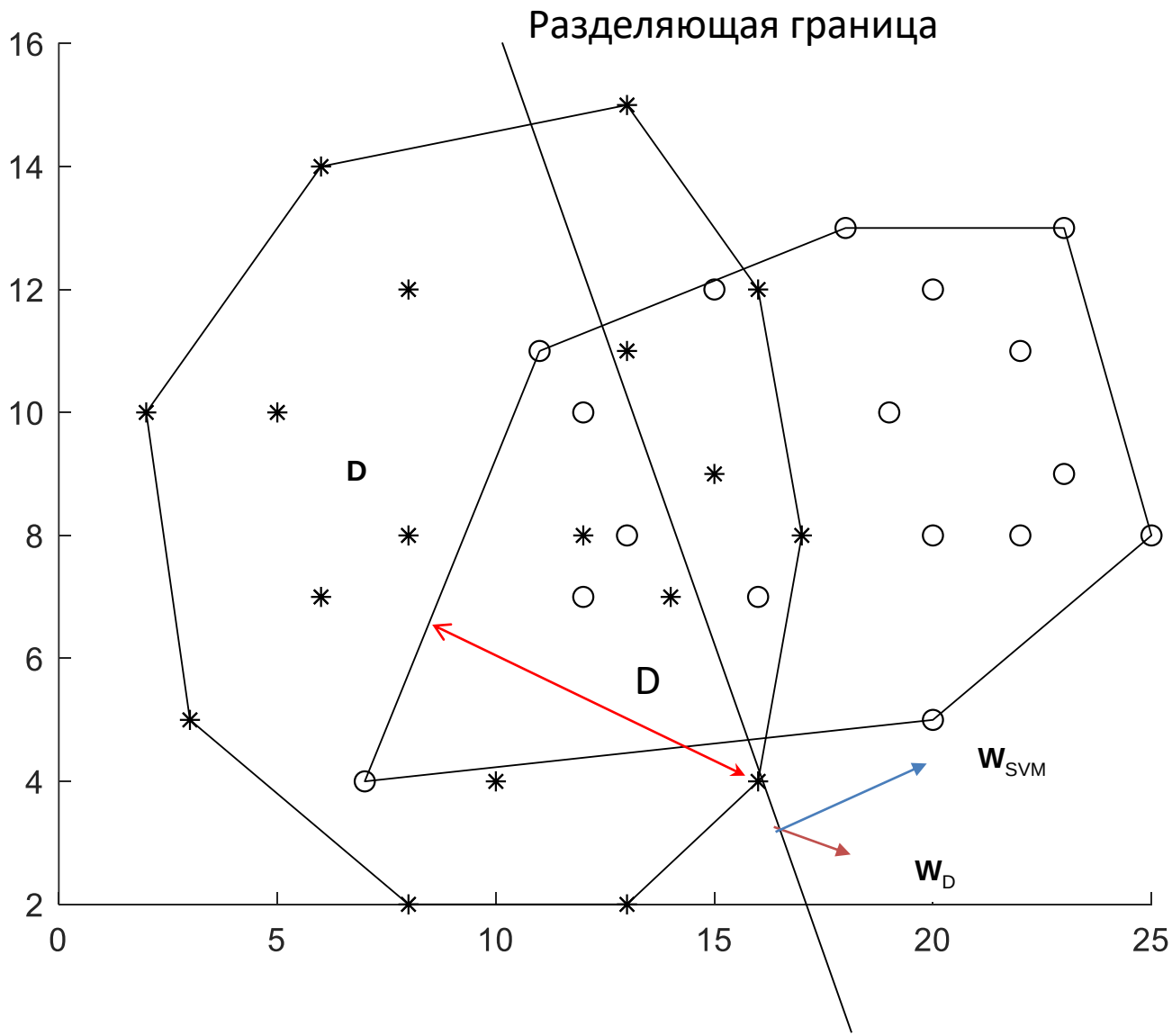
Determination of the distance from point to convex hull

$$D(A,B) = \min_{\substack{a \in A \\ b \in B}} \|a - b\|$$

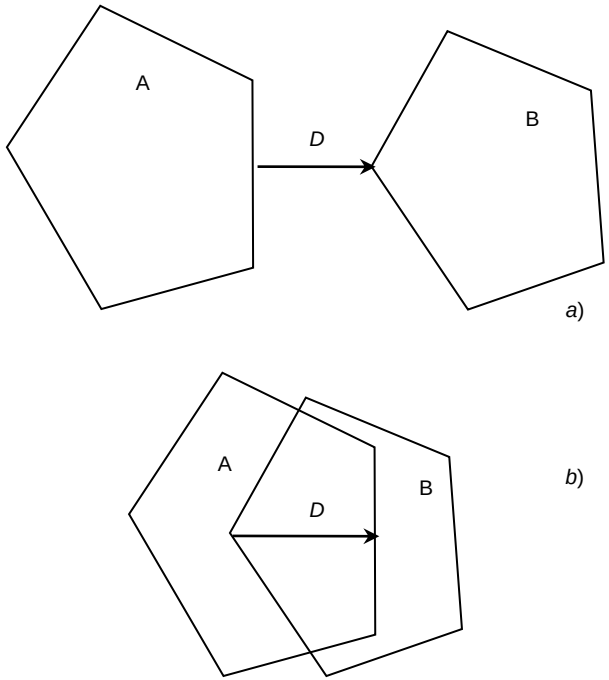


Расстояние от точки до выпуклой оболочки




Разделяющая граница
16
14
12
10
8
6
4
2
0
5
10
15
20
25
D
D
W_SVM
W_D

Directional depth of penetration



For any $\{A, B\} \subseteq \mathbf{R}^n$

$$D(A,B) = \min_{\substack{a \in A \\ b \in B}} \|a - b\|$$

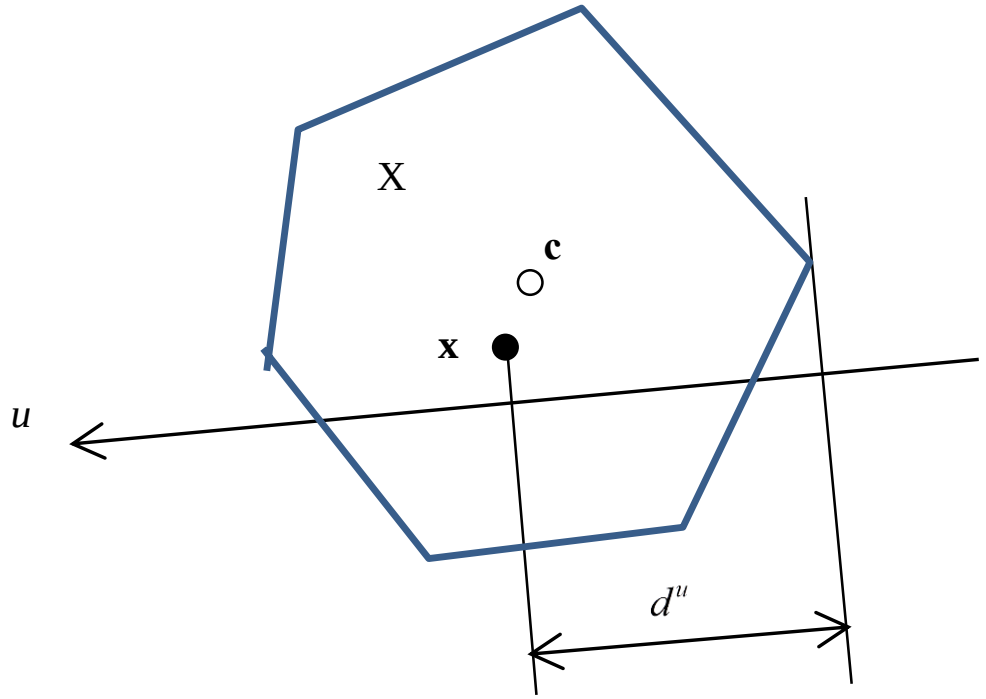
a) *D* – the minimum distance between A and B and the corresponding weight vector;

b) *D* – the minimum penetration depth between A and B and the corresponding weight vector.

In the case *b)*, the *penetration depth* is used. It reflects the intersection degree. For a given direction vector **u**, the ***directional penetration depth*** between A and B is defined as the minimum shift to be applied to B towards **u**, so that inner region A could not intersect with B. The ***total penetration depth*** between A and B is defined as the minimum shift to be applied to B *in any direction*, so that inner region A could not intersect with B .

d^u

Directional depth of penetration



On the determination of the directional depth of penetration d^u of a vector **x** into the convex hull *conv*(X) of a set X in the direction *u*. In the figure, **x** – the test vector, **c** – the centroid of the set X.

Calculating $d_i^u\left(\mathbf{x}, conv(\mathrm{X}_i)\right)$

DISTANCE algorithm

Input: $\mathbf{u}, \mathbf{x}, \mathrm{X}_i$ {direction vector, test point, set of elements of the i-th class}

Output: F, $d_i^u\left(\mathbf{x}, conv(\mathrm{X}_i)\right)$ {intersection mark, distance to a convex hull of the *i*-th class}

1. For set X_i , define the data matrix $\mathbf{X}_i$
2. If $\min\left(\mathbf{u}^T\mathbf{X}_i\right) \le \mathbf{u}^T\mathbf{x} \le \max\left(\mathbf{u}^T\mathbf{X}_i\right)$ than
3. F = 1; $d_i^u\left(\mathbf{x}, conv(\mathrm{X}_i)\right) = \left|\mathbf{u}^T\mathbf{x} - \min\left(\mathbf{u}^T\mathbf{X}_i\right)\right|$
4. else if $\mathbf{u}^T\mathbf{x} < \min\left(\mathbf{u}^T\mathbf{X}_i\right)$ than
5. F = 0; $d_i^u\left(\mathbf{x}, conv(\mathrm{X}_i)\right) = \left|\mathbf{u}^T\mathbf{x} - \min\left(\mathbf{u}^T\mathbf{X}_i\right)\right|$
6. else F = 0; $d_i^u\left(\mathbf{x}, conv(\mathrm{X}_i)\right) = \left|\mathbf{u}^T\mathbf{x} - \max\left(\mathbf{u}^T\mathbf{X}_i\right)\right|$
7. end

Here $\mathbf{X}_i$ is the data matrix formed by the set X_i, F is the flag of intersection between $\mathbf{u}^T\mathbf{x}$ and $\mathbf{u}^T\mathbf{X}_i$; F = 1 stands for "intersect", F = 0 stands for "do not intersect".

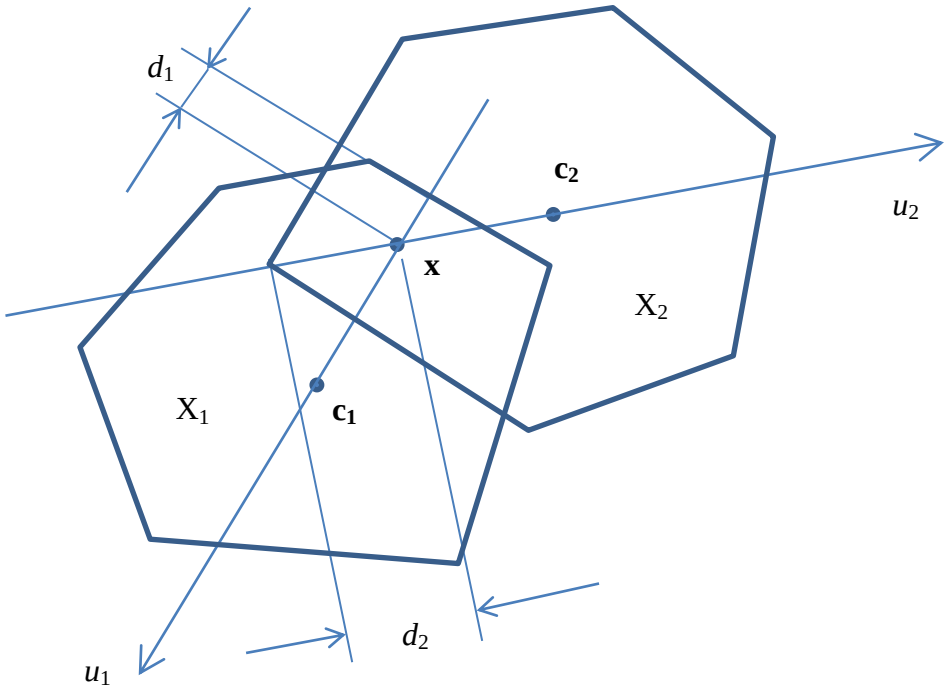
$\mathbf{u} = \left(\mathbf{c} - \mathbf{x}\right) / norm\left(\mathbf{c} - \mathbf{x}\right)$, where $\mathbf{c}$ is the class centroid

Lite Nearest Convex Hull Classifier

For m classes X_i, $i = 1,2,...,m$, as a result of the DISTANCE procedure we will obtain m pairs of $(\mathrm{F}_i, d_i), i = 1,2,...,m$. Then, based on the obtained data, we will carry out classification by the following decision rule.

1. If no pair contains F = 1, then the number of the recognized class is calculated using $class(\mathbf{x}) = \arg\min_{i=1,2,...m} d_i\left(\mathbf{x}, conv(\mathrm{X}_i)\right)$.
2. If only one pair contains F = 1, then the number of the recognized class will equal to the index of this pair.
3. If several pairs (or possibly all) contain F = 1 and the indices of these pairs form set G, then the number of the recognized class will be selected from these classes, so that $class(\mathbf{x}) = \arg\max_{i \in G} d_i\left(\mathbf{x}, conv(\mathrm{X}_i)\right)$.

Lite Nearest Convex Hull Classifier



X_1 and X_2 are two classes, $\mathbf{c}_1$ and $\mathbf{c}_2$ are their centroids, $\mathbf{x}$ – test point, u_1 and u_2 – the directions from the test point to the centroids of the classes, d_1 and d_2 are the directed penetration depths of $\mathbf{x}$ into the convex hulls of the classes. Since $\boldsymbol{d_2 > d_1}$ (the penetration of the point in X_2 is greater than in X_1), then according to the above decision rule, **the point x belongs to the class X_2.**

Experimental studies

1. The synthesized digital data. 2 classes in the 5-dimensional space with 100 copies each have been used. The signs were assigned uniformly distributed random values in the intervals (0.0, 1.0) for the first class and (0.3, 1.3) for the second.

2. Breast cancer diagnosis problem. Real data from the UCI Machine Learning Repository. The data includes 683 cases: 444 cases of benign tumors B (1st class) and 239 cases of malignant cancer M (2nd class). 9 cytological features are integers ranging from 1 to 10.

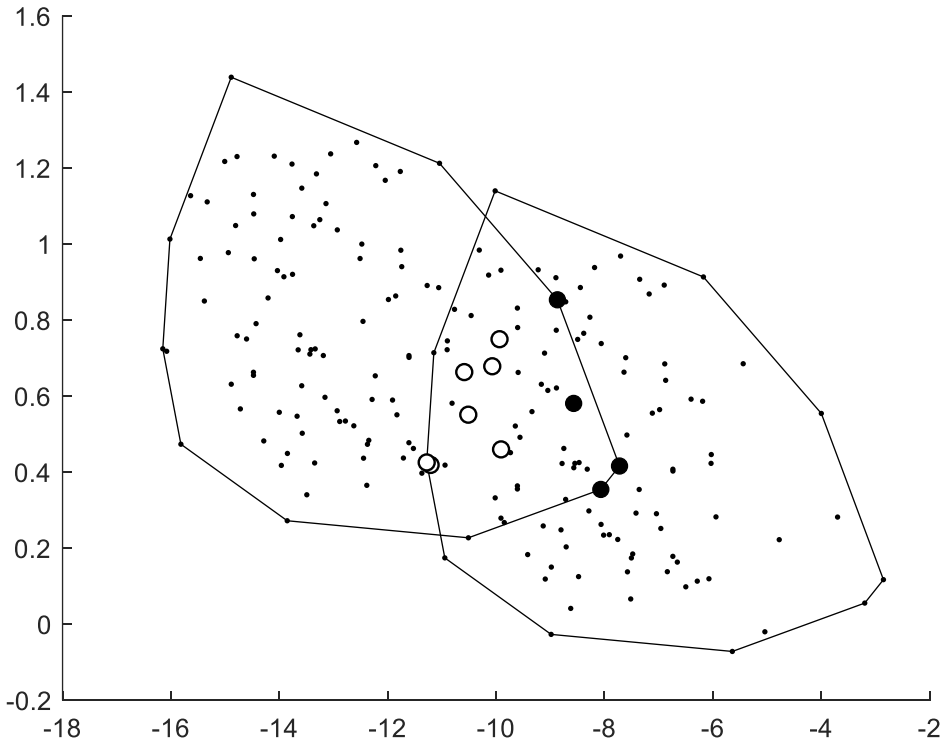
Assessment of class intersections

For 2 intersecting classes X_1 amd X_2 the intersection of the 1st class is $g_1 = r_1/k_1$, where r_1 - the number of elements of the class X_1, caught in the convex hull of the class X_2, k_1 - the total number of elements of the class X_1. The intersection of the second class, respectively, is equal $g_2 = r_2/k_2$. The average overlap is $g = (g_1 + g_2)/2$. The classification errors for the two classes will be p_1 and p_2, and the average classification error will be $p = (p_1 + p_2)/2$.

For synthesized digital data 4 and 7 points for different classes fell into the intersection zone. So g_1 = 4 %, g_2 = 7 %, g = 5.5 %.

For breast cancer diagnosis problem the study of the intersectability of classes B and M using the computational geometry procedures in the original 9-dimensional space has yielded the following results: no cases from class B fall into class M; 2 cases from class M fell into class B. So for this problem g_1 = 0 %, g_2 = 0.56 %, g = 0.28 %.

Visualization of the intersection area of two classes



Study on visualization of the area of intersection of two classes of digital data. Points that fall into the intersection zone of convex hulls of classes in 5-dimensional space: ○ - for the class located on the right, ● - for the class located on the left; abscissa axis - $\mathbf{w}_1$, ordinate axis - $\mathbf{w}_2$.

Recognition errors for various algorithms

Data	p_1	p_2	p
SVM (liner)	4.10	2.00	3.05
SNCH *)	----	----	2.70
LNCH (our method)	2.93	4.18	3.56
kNN (k = 5)	2.25	3.15	2.70

*) G. Nalbantov and E. Smirnov, “Soft nearest convex hull classifier,” Proc. 19th European Conference on Artificial Intelligence (ECAI-2010), H. Coelho et al. (Eds.), IOS Press, 2010, pp. 841-846. doi: 10.3233/978-1-60750-606-5-841.

The results show that the efficiency of LNCH is at the level of the efficiency of other similar algorithms in the absence of optimization procedures and the need to adjust the parameters of these procedures in other algorithms.

Advantages and Disadvantages of the LNCH method

Advantages of the LNCH method

1. Ease of implementation and use, especially for multi-class tasks
2. You should only remember the vertices of convex hulls without remembering all members of the training sample
3. Lack of optimization tasks and optimization parameters when implementing a decision rule
4. A simple way to calculate the distance to the convex hull for linearly inseparable classes

Disadvantages of the LNCH method

1. Low noise immunity
2. The decision surface between the classes is not explicitly computed.

Conclusion

In this paper, a new and simpler method for estimating the proximity of a test point to convex hulls of classes (DISTANCE algorithm) is proposed.

The method is based on the analysis of the projections of the vertices of convex hulls on the direction from a given point to the centroid of the class.

It underlies the proposed lightweight classification algorithm based on the nearest convex hull (LNCH).

The main advantage of the method is

1. **the ease of implementation and application.**
2. **LNCH can be easily applied to multi-class problems**.
3. **the lack of optimization procedures and the need to adjust the parameters of these procedures.**

Recognition accuracy and its generalizing ability are tested on two tasks, one of which is the task of medical diagnostics on real data. The experimental results showed the satisfactory quality of the LNCH algorithm.

Thank you