Wykład 5 Zagadnienie klasyfikacji c.d.

dr hab. Konrad Furmańczyk, prof. SGGW

Instytut Informatyki Technicznej/KZM, SGGW, bud. 34, pok 3/87
email: konrad_furmanczyk@sggw.edu.pl

November 6, 2021

Metoda najblizszego sąsiada 1NN (kNN)

1NN -metoda najbliższego sąsiada, nowy obiekt przyporządkowujemy do klasy obiektu, który jest najbliżej (w sensie pewnej odległości).

kNN -metoda k najbliższych sąsiadów, nowy obiekt otrzymuje klasę która występuje najczęściej pośród jego k sąsiadów.

Metoda najbliższego sąsiada jest odporna na występowanie obserwacji odstających oraz zakłóceń.
Główna wada: długi czas obliczeń, który rośnie bardzo szybko wraz ze wzrostem liczby obserwacji.
Zmienne (atrybuty) tylko ilościowe.

Potrzebna normalizacja, która transformuje zmienne o różnych jednostkach w wielkości niemianowane i porównywalne.

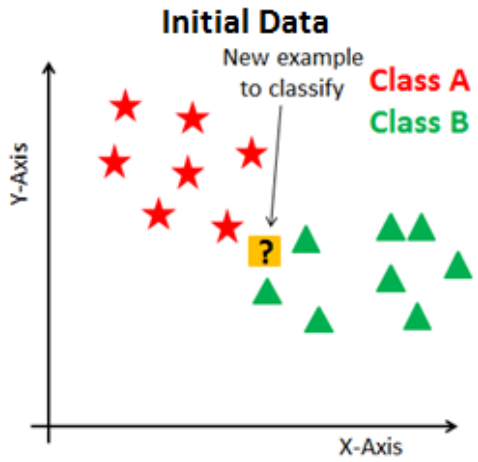
normalizacja min-max

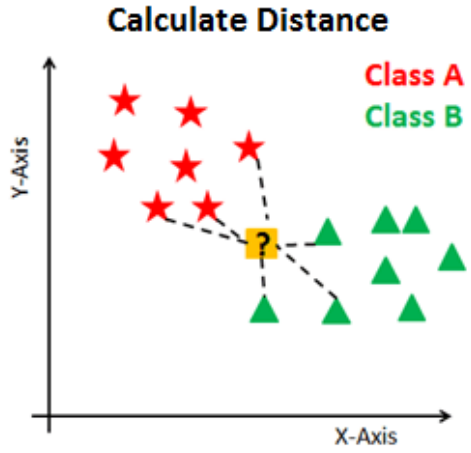
$$x = \frac{x - min(x)}{max(x) - min(x)}$$

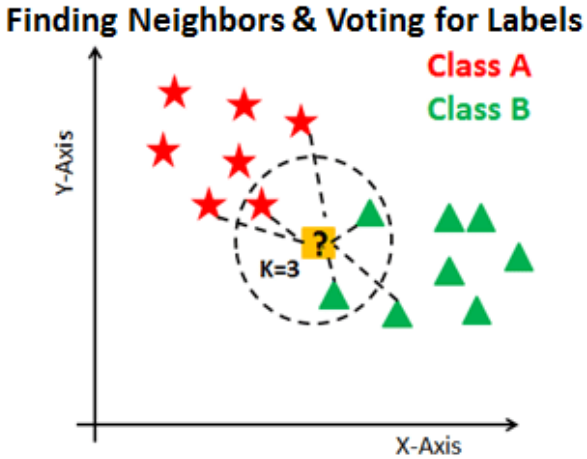
standaryzacja

$$x = \frac{x - mean(x)}{sd(x)}.$$

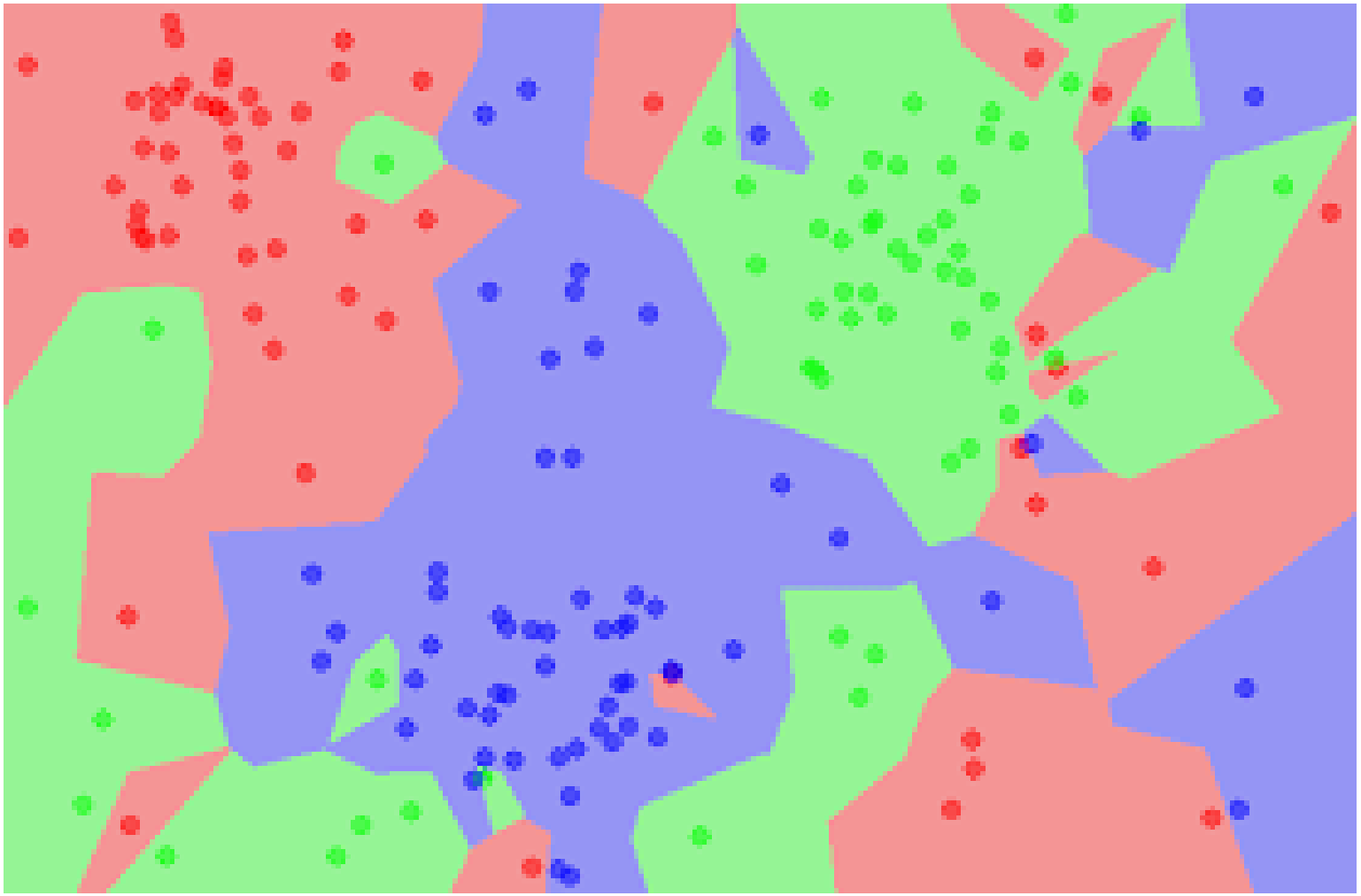
https://blakelobato1.medium.com/k-nearest-neighbor-classifier-implement-homemade-class-compare-with-sklearn-import-6896f49b89e



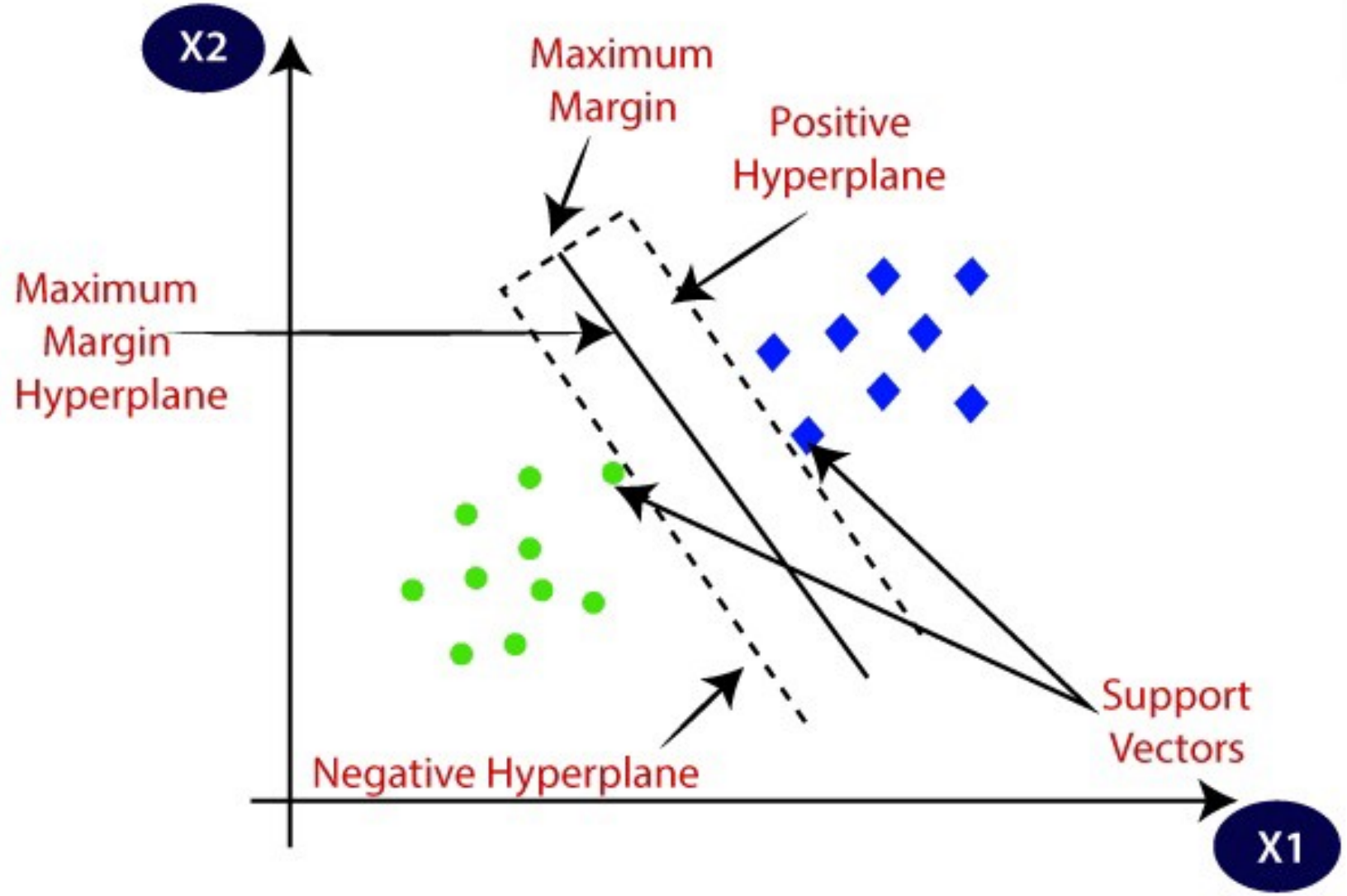




kNN



SVM (support vector machine-metoda wektorów nośnych)



Populacje liniowo separowalne

Niech $\mathcal{L} = \{(\mathbf{x_1}, y_1), ..., (\mathbf{x_n}, y_n)\}$ -próba uczaca, mamy dwie populacje liniowo separowalne $G_1(y_j = +1), G_2(y_j = -1)$

$$g(\mathbf{x}) = \mathbf{w^T x} + w_0$$

Rozważamy klasyfikator liniowy postaci

$$d(\mathbf{x}) = \begin{cases} +1 & g(\mathbf{x}) > 0 \\ -1 & g(\mathbf{x}) \leq 0 \end{cases}.$$

Szukamy klasyfikatora d który separuje dane z próby uczącej, tak że hiperpłaszczyzna $g(\mathbf{x}) = 0$ jest maksymalnie odległa (maksymalny margines) od najbliższej obserwacji z próby uczącej L.

Nieseparowalne liniowo populacje

Dopuszczamy dla pewnych elementów próby możliwość błędnej klasyfikacji $\xi_j > 1$ tak, że

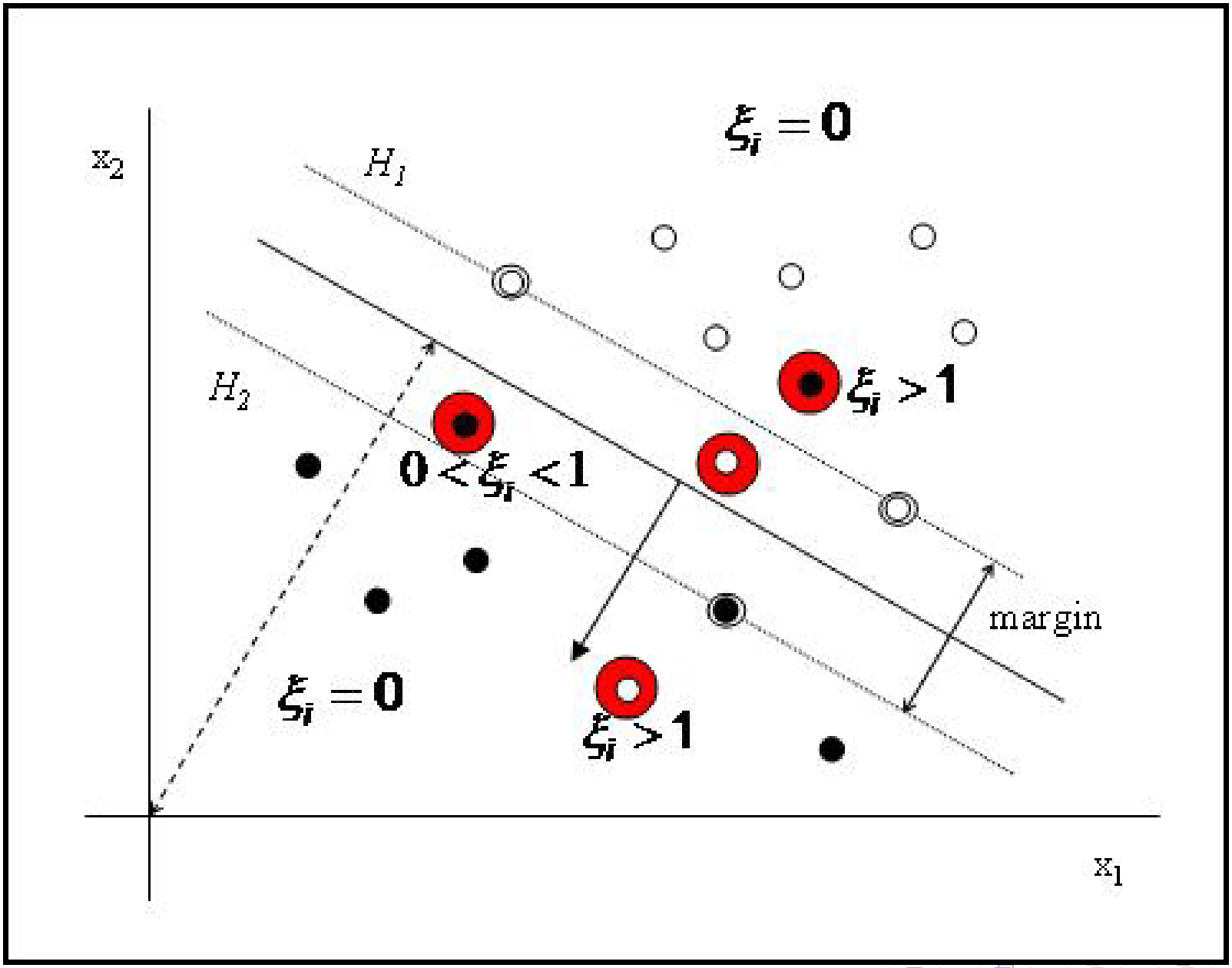
$$y_j g(\mathbf{x}) \geq 1 - \xi_j.$$

Szukamy minimum wyrażenia (przy powyższym warunku dla $\xi_j \geq 0$)

$$\frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w} + C\sum_{j=1}^{n}\xi_j,$$

gdzie C-pewna stała.

SVM-populacje niesparowalne liniowo



Nieliniowe SVM

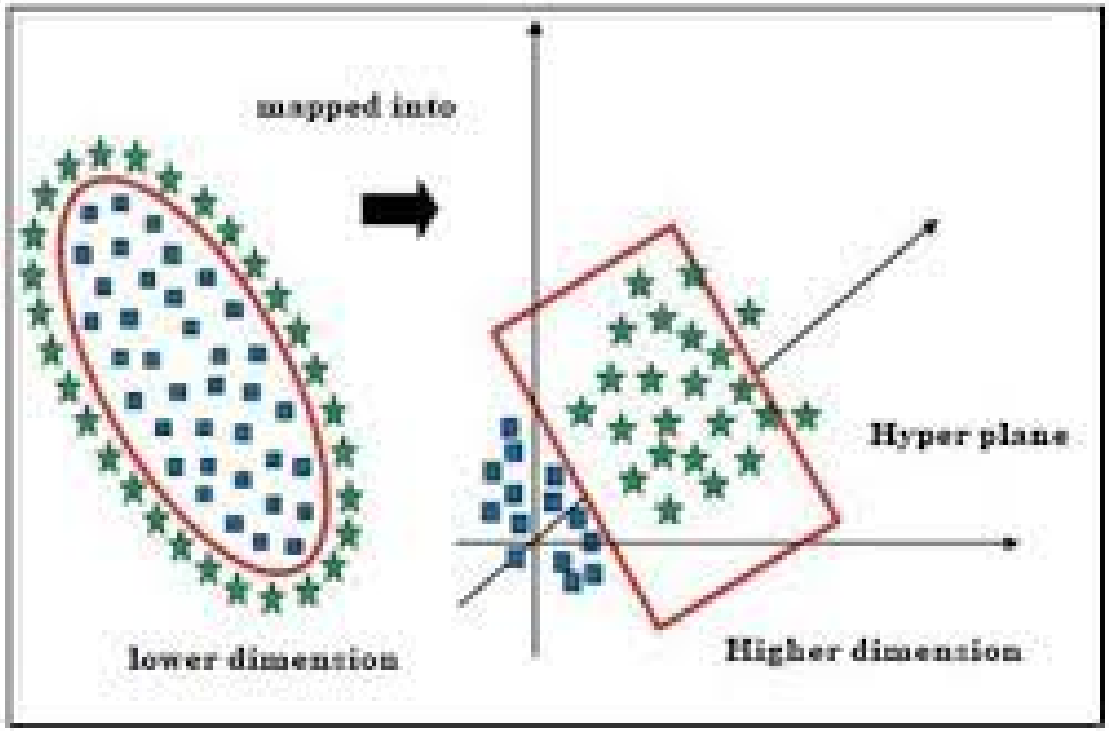


Figure: Overview of SVM non- linear problem

Nieliniowe SVM

Używamy pewnego przekształcenia nieliniowego przestrzeni próby $\mathcal{X}$ na przestrzeń $\phi(\mathcal{X})$. W tej nowej przestrzeni stosujemy model liniowy

$$g(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\phi(\mathbf{x}) + w_0.$$

Jądro przestrzeni $\phi(\mathcal{X})$

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T\phi(\mathbf{z}).$$

Other Types of Kernels

type of SVM	K(x,y)	Comments
Polynomial learning machine	$(\mathbf{x}^T\mathbf{y} + 1)^p$	p: selected a priori
Radial basis function	$\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|^2\right)$	σ^2: selected a priori
Two-layer perceptron	$\tanh(\beta_o\mathbf{x}^T\mathbf{y} + \beta_1)$	only some β_o and β_1 values are feasible.

What kernel is feasible? It must satisfy the "Mercer's theorem"!

Drzewa klasyfikacyjne

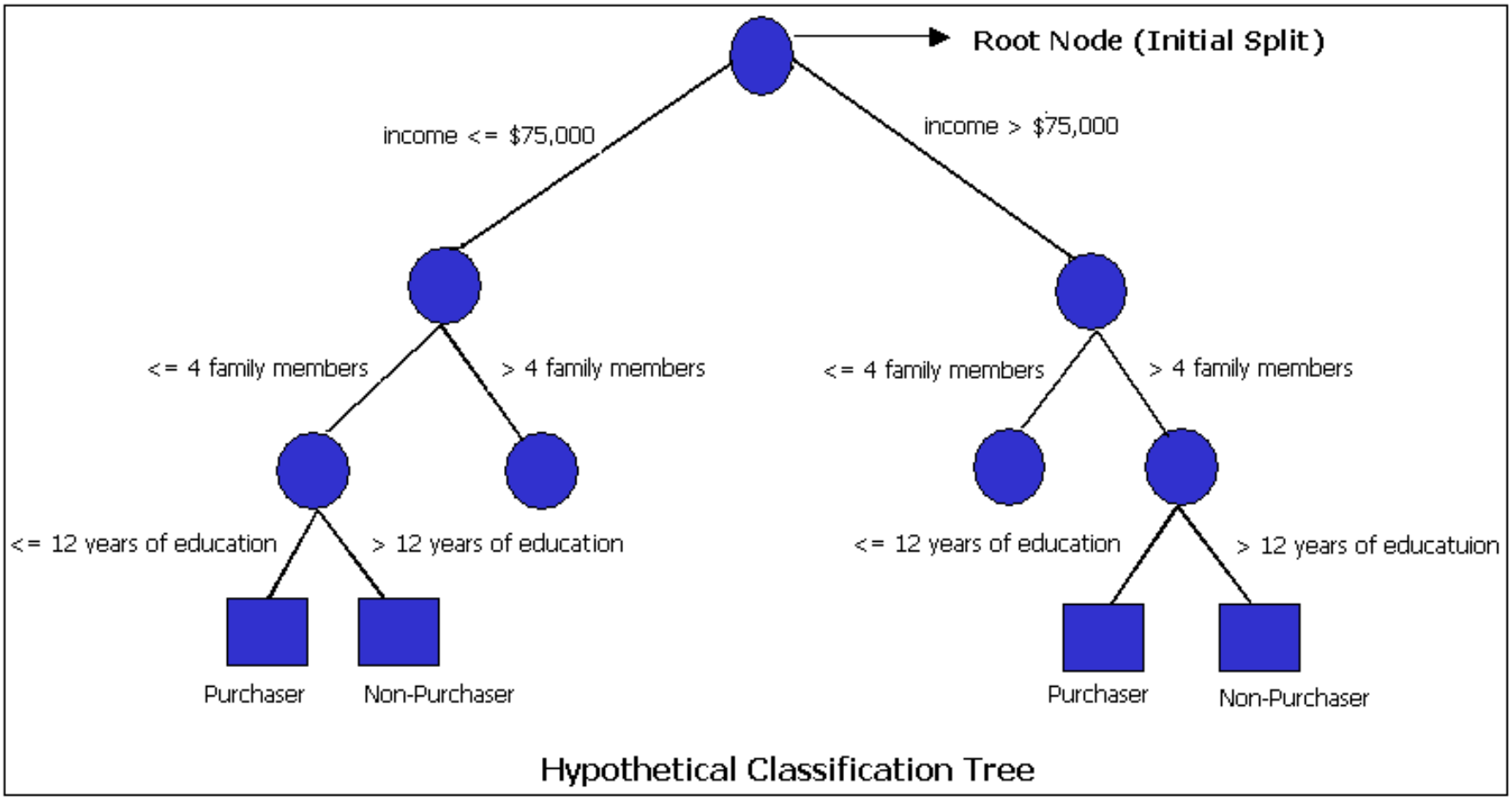
Drzewo klasyfikacyjne składa się z korzenia oraz gałęzi odchodzących z korzenia do kolejnych węzłów.

Dla klasyfikowanej obserwacji w każdym węźle sprawdzany jest pewien warunek, i na jego podstawie wybierana jest jedna z gałęzi prowadząca do kolejnego węzła poniżej.

Na samym dole znajdują się liście, w których odczytujemy etykietę klasy do której należy przypisać badaną obserwację.

Klasyfikacja obserwacji polega na przejściu od korzenia, przez węzły do liścia i przypisaniu tej obserwacji klasy zapisanej w danym liściu.

Drzewa klasyfikacyjne



Konstrukcja drzewa

Sekewncyjne dzielenie podzbiorów przestrzeni próby na dwa rozłączne i dopełniające sie podzbiory (węzły), startując od całej przestrzeni próby.

Podziały są uwarunkowane przez obserwacje ze zbioru uczącego należące do danego węzła.

W każdym kroku podział jest tak dokonywany aby uzyskane części (węzły) były możliwie jednorodne.

Każdy końcowy podzbiór (liść) ma przypisaną jedną etykietę klasy.

Konstrukcja drzewa

Dla każdego węzła t określmy pewną miarę $I(t)$ niejednorodności elementów w tym węźle.

Dla każdego podziału węzła t określamy niejednorodność elementów w tym węźle oraz w węzłach jego potomkówt_L (lewy "prawda") i t_R (prawy "fałsz").

Zmiana jednorodności podziału s wezła t określamy jako

$$\Delta I(s,t) = I(t) - p_L I(t_L) - p_R I(t_R),$$

gdzie $p_L = n(t_L)/n(t)$, $p_R = n(t_R)/n(t)$, $n(t)$-ilość elementów próby uczącej w węźle t, itd.

Konstrukcja drzewa

Optymalny podział $s*$ węzła t to taki, że

$$\Delta I(s*, t) = max_s \Delta I(s, t).$$

Miara niejdnorodności węzła to $I(t) = \phi(p(1|t), ..., p(K|t))$, gdzie $p(i|t) = P(\mathbf{X} \in t|Y = i)$,$K$-liczba klas. Inaczej, wybieramy podział, który daje maksymalną redukcję niejednorodności indeksu przynależności do klasy w węźle.

Funkcja ϕ spełnia warunki:

1. osiąga maksimum tylko dla punktu $\left(\frac{1}{K}, ..., \frac{1}{K}\right)$.
2. osiąga minimum tylko w punktach:
$(1, 0, ...0), (0, 1, 0, ..., 0), ..., (0, ..., 0, 1)$.
3. jest symetryczną funkcją swoich argumentów.

Wybóry funkcji ϕ

1. Błąd klasyfikacji

$$\phi(p_1, ..., p_K) = 1 - max\{p_1, ..., p_K\}$$

2. Entropia

$$\phi(p_1, ..., p_K) = -\sum_{i=1}^{K} p_i log p_i$$

3. Indeks Giniego

$$\phi(p_1, ..., p_K) = 1 - \sum_{i=1}^{K} p_i^2.$$

Przycinanie drzewa

Przy tworzeniu drzewa unikamy zbytniego rozbudowania struktury drzewa (złożoności modelu).

Złożoność modelu prowadzi do trudności w jego interpretacji oraz utraty właściwości generalizacji.

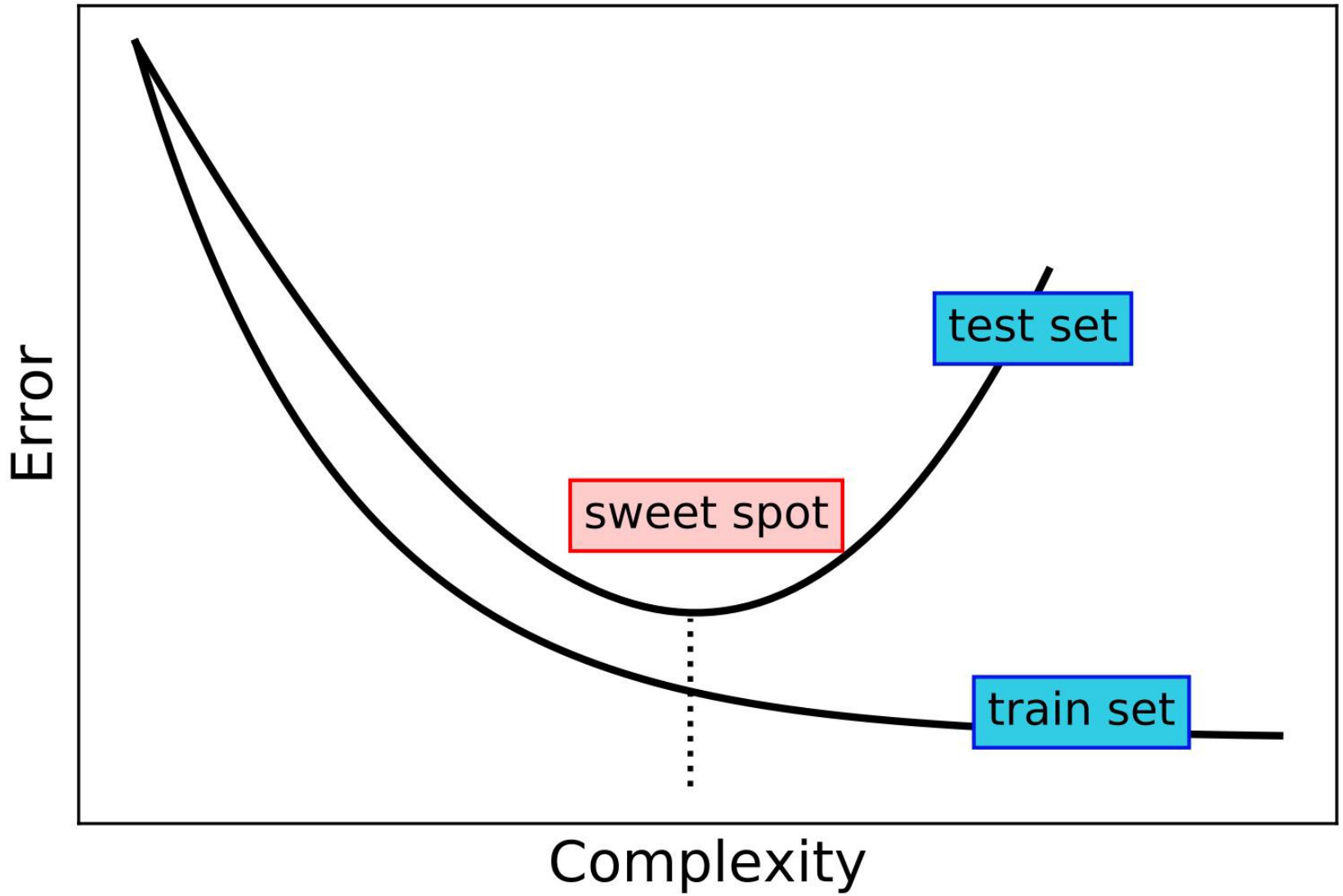
Zbyt duże drzewo to tzw. efekt przeuczenia (ang. overfitting). Polega on na tym, że drzewo doskonale klasyfikuje obiekty z próby uczącej lecz coraz słabiej (w miarę zwiększania liczby liści) nowe elementy.

W celu uniknięcią przeuczenia modelu konstruuje się drzewa maksymalnie złożone, a następnie stosuje się przycinanie drzewa (ang. pruning), która zmniejsza drzewo.

Drzewa decyzyjne-podsumowanie

1. Atrybuty mogą być ilościowe i jakościowe.
2. Atrakcyjna wizalizacja i prosta interpretacja.
3. Odporność na obserwacje odstające i braki danych.
4. Niestabilne. Niewielkie zmiany próby uczącej mogą dawać duże różnice w konstrukcji drzewa jak i predykcji. W celu poprawienia stabilności stosuje się techniki wzmacniania klasyfikatorów: bagging, boosting oraz lasy losowe.
5. Procedury konstrukcji drzew: CHAID, CART, C4.5, QUEST.

Overfitting



Lasy losowe

Łączymy wielu drzew klasyfikacyjnych. Losujemy B prób bootstrapowych (dużo), dla każdej z nich konstruujemy drzewo klasyfikacyjne w taki sposób, że w każdym węźle losujemy m cech, które będą uczestniczyły w wyborze najlepszego podziału.

Drzewa budowane są bez przycinania. Ostatecznie obserwacja klasyfikowana jest poprzez metodę większościowego głosowania (ang. majority voting).

Wybieramy $m = \sqrt{p}$, gdzie p -wymiar danych (ilość atrybutów).

Lasy losowe

Out of bag (OOB)-ocena jakości klasyfikatora. Zauważmy, że nie każda obserwacja jest używana do uczenia (przeciętnie 63% uczestniczy).
Jeśli zaklasyfikujemy obserwacje, które nie należą do zbioru uczącego dla konkretnego drzewa, to otrzymamy dla danej obserwacji przewidywane klasy przez te drzewa, które nie wykorzystywały jej w procesie uczenia. Ostateczną oceną jest klasa, która została wybrana przez większość drzew. W taki sposób dla każdej obserwacji otrzymamy ocenę klasy, z której pochodzi. Procent błędnych decyzji to OOB.