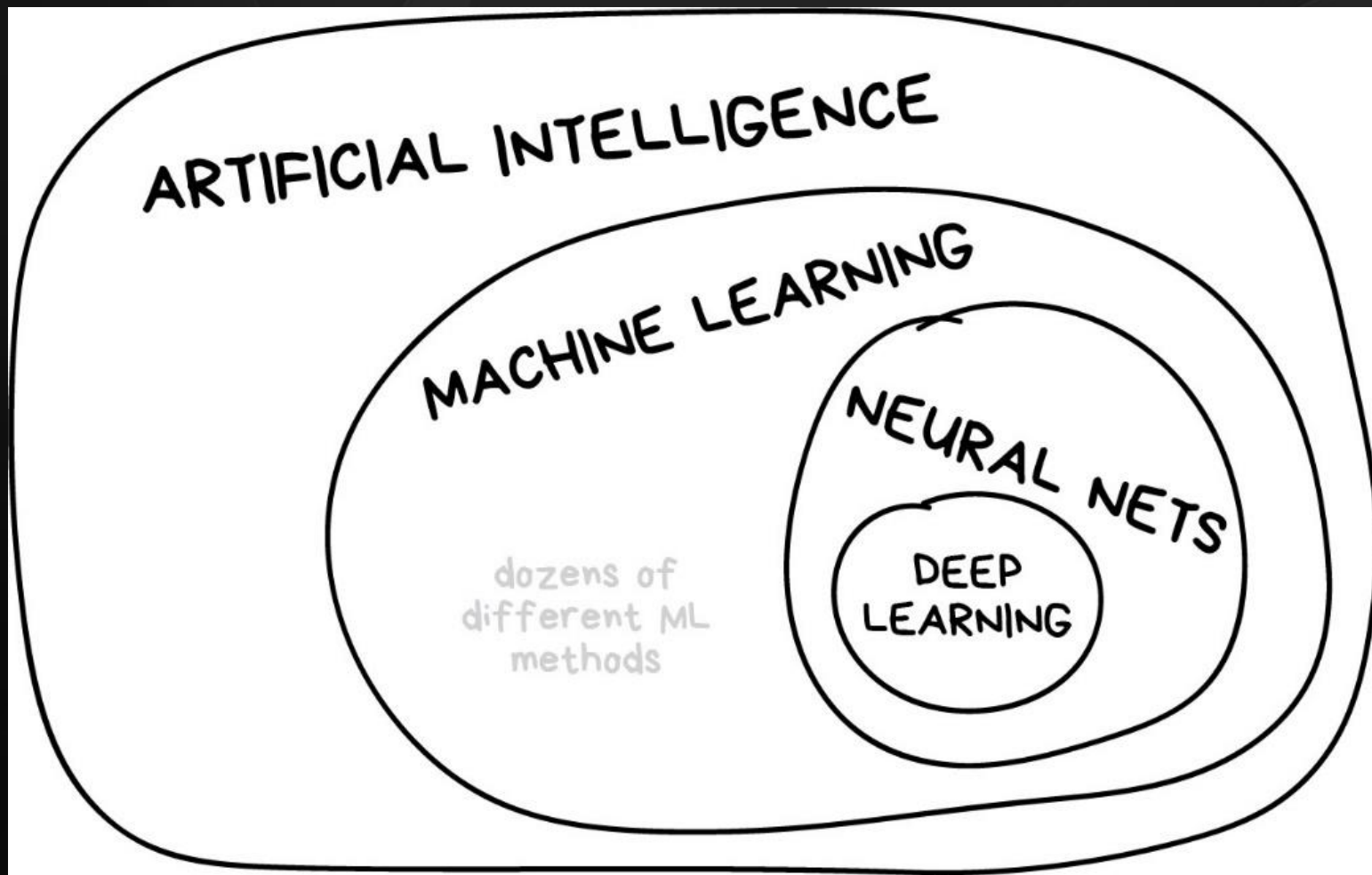


COGSYS: KOGNITIVNÍ SYSTÉMY A NEUROVĚDY

Český institut informatiky robotiky a kybernetiky (CIIRC)

Ing. Václav Gerla, Ph.D.



Co je klasifikace?

- Zařazení objektu nebo události do jedné z několika předem daných kategorií (tříd)
- Typicky na objektu pozorujeme/měříme nějaké vlastnosti, které popíšeme vektorem příznaků
- Pomocí příznaků klasifikujeme (predikujeme) třídy

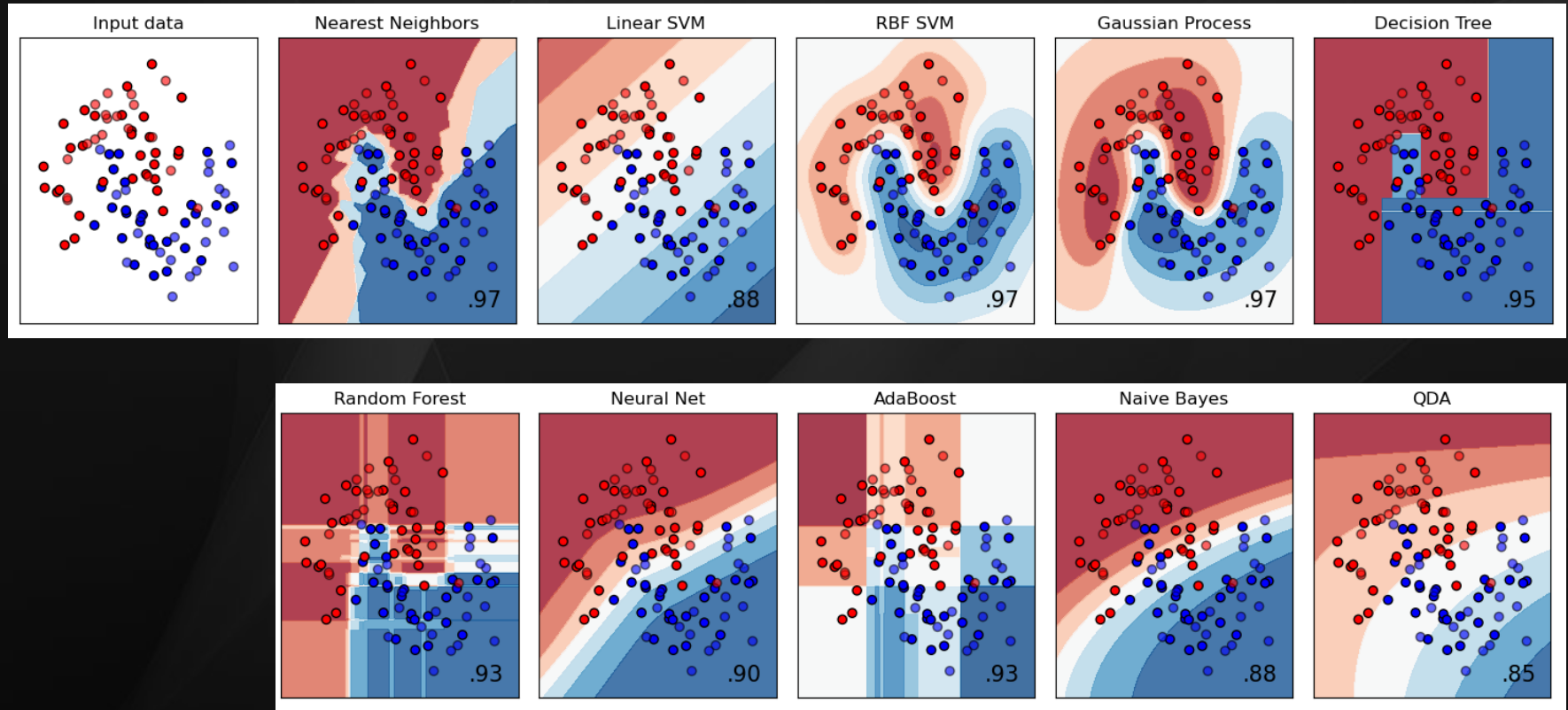
česky	anglicky
třídy (klasifikační třídy)	classes
příznaky (atributy)	features
klasifikace	pattern classification, recognition

- Máme např. 100 záznamů o studentech, kteří buď zkoušku udělali nebo neudělali
- Na základě dat dostupných již na začátku semestru chceme predikovat, zda daný student zkoušku a předmět dokončí
- Z dostupných dat můžeme extrahovat příznaky,
 - Některé příznaky jsou lepší
 - předchozí zkušenost s programováním, známky z jiných předmětů
 - Některé jsou špatné
 - pohlaví, barva vlasů, měsíc narození

- Máme sadu možných symptomů, výsledků testů a vyšetření provedených na pacientovi
 - Např. měření krevního tlaku, saturace O₂, ohodnocený záznam celonočního EEG záznamu, čas usnutí a probuzení, apod.
- Chceme predikovat, zda pacient trpí např. spánkovou poruchou.

Existují různé typy klasifikátorů

CIIRC

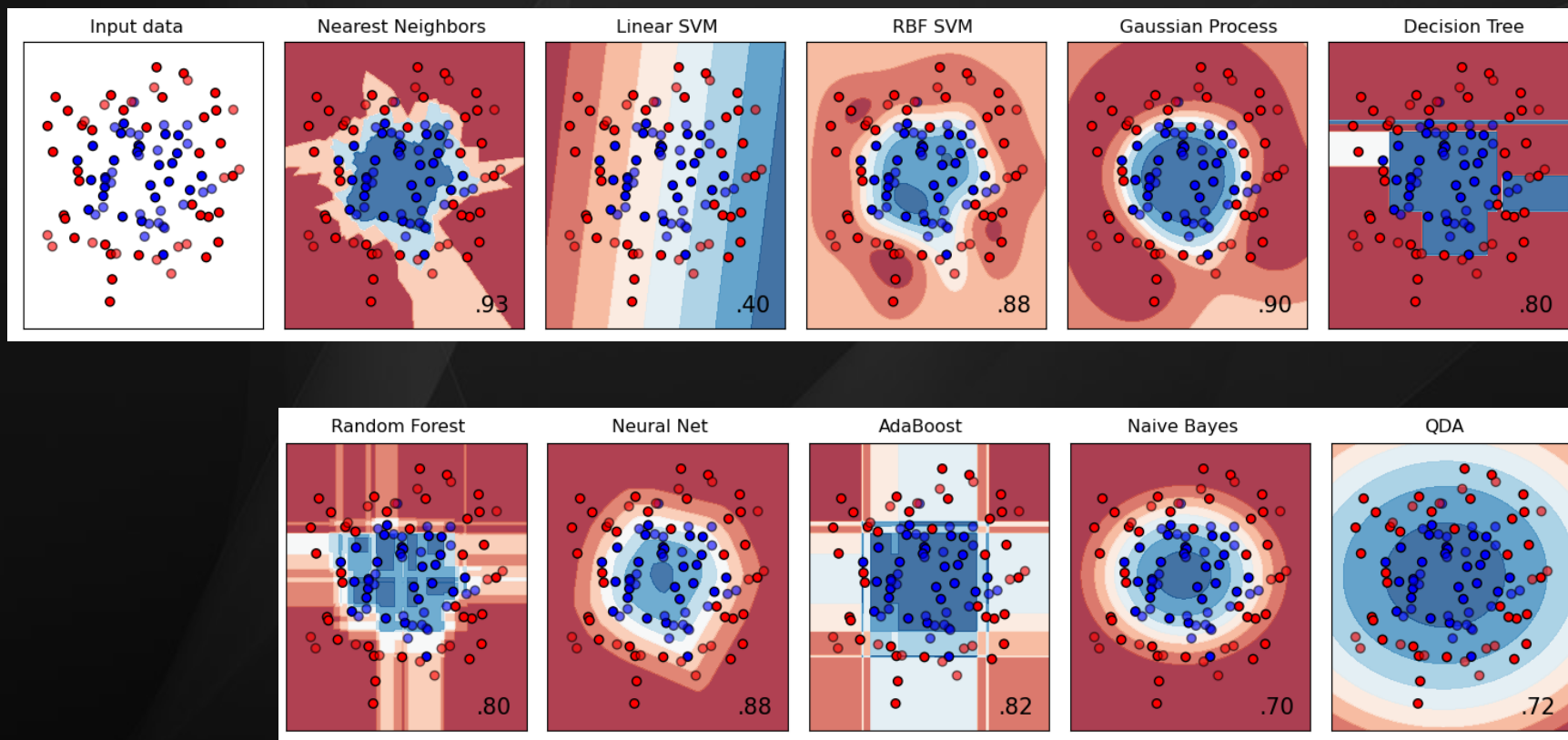


Rozhodovací hranice je různá pro různé klasifikátory

<https://scikit-learn.org/>

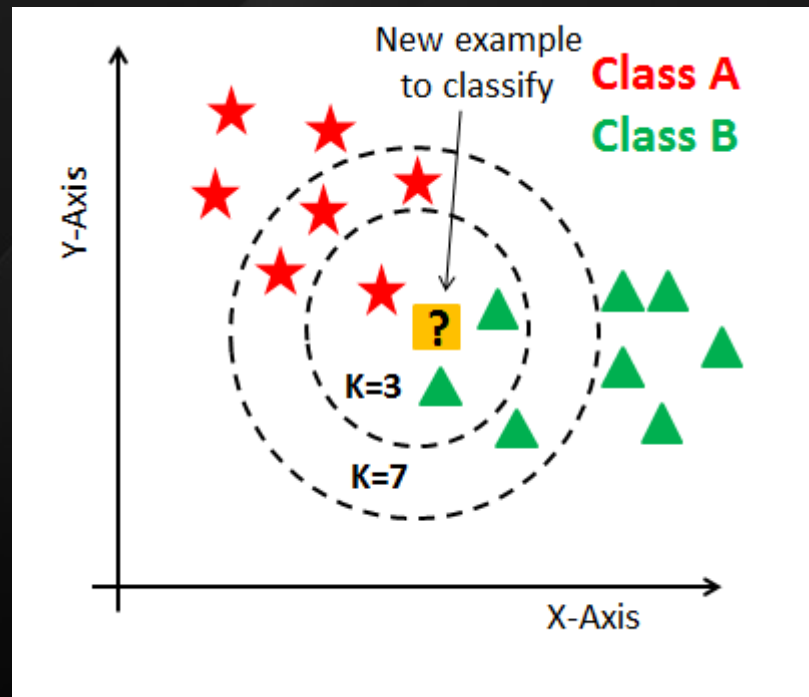
Existují různé typy klasifikátorů

CIIRC

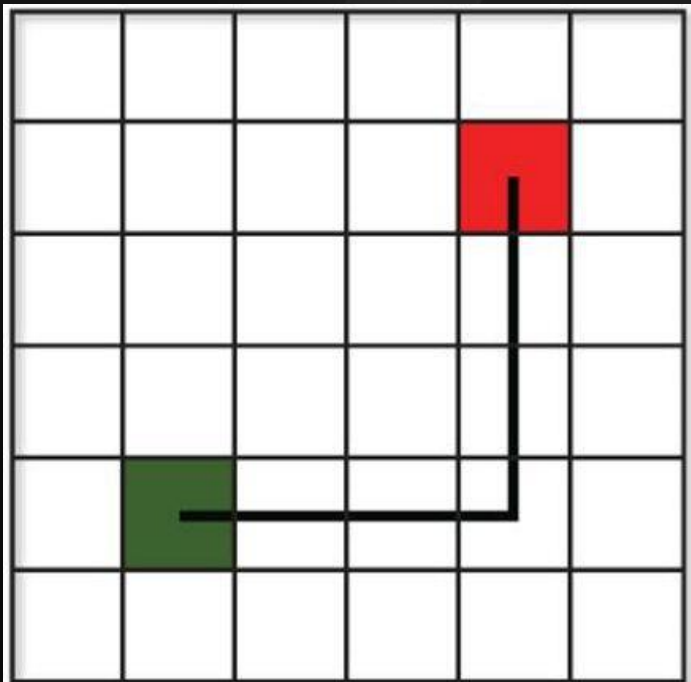


1-NN: Klasifikovaný bod se zařadí do stejné třídy jako jeho nejbližší soused z trénovací množiny

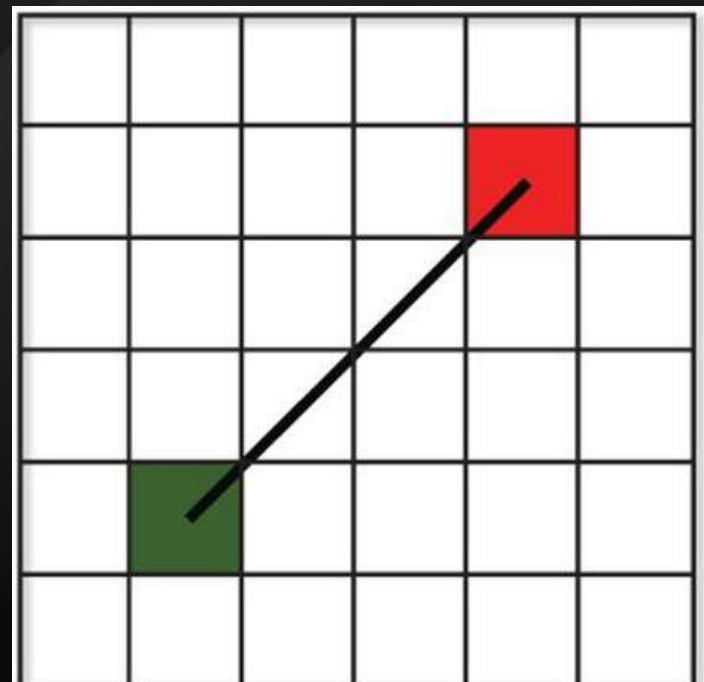
K-NN: Najde se k nejbližších sousedů a rozhodne se podle majoritní třídy



Jak měřit vzdálenost mezi body?



Manhattan Distance

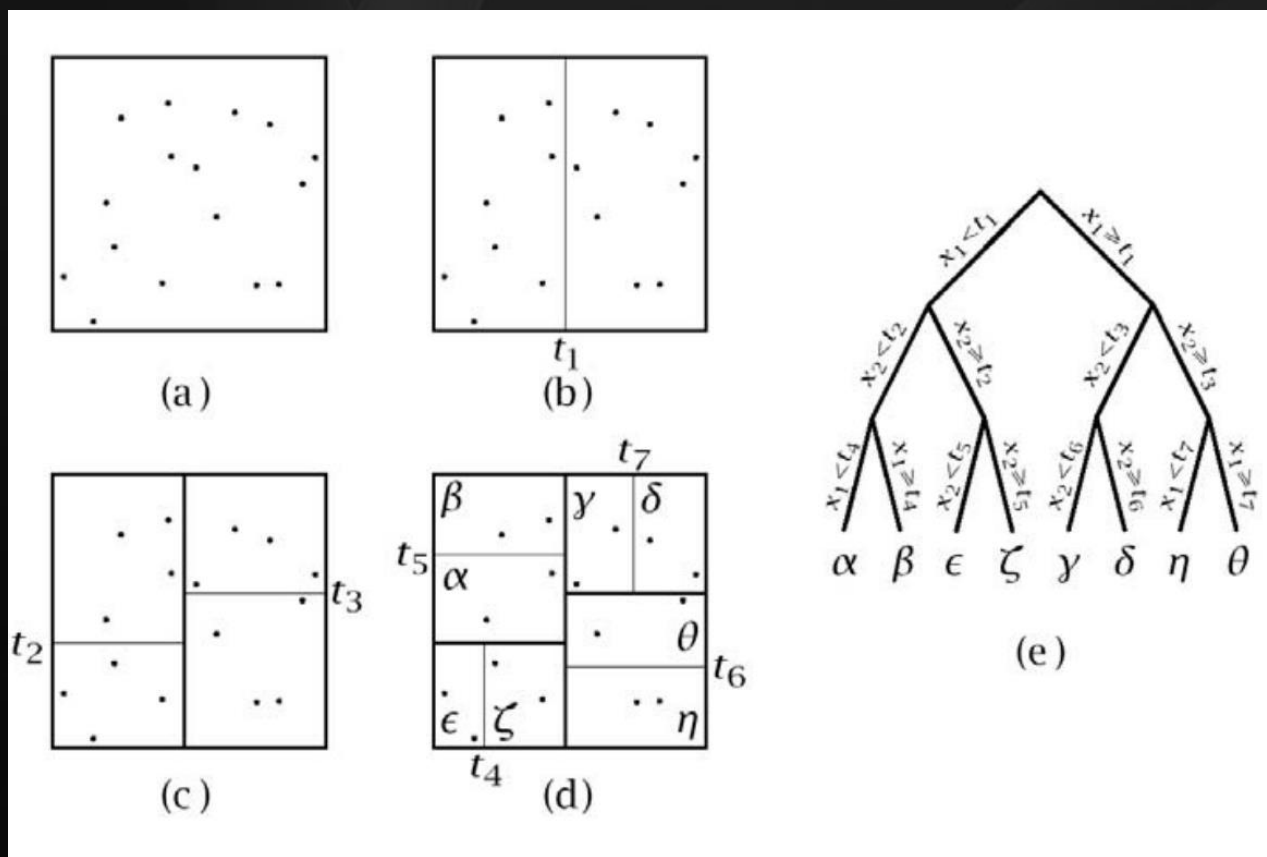


Euclidean Distance

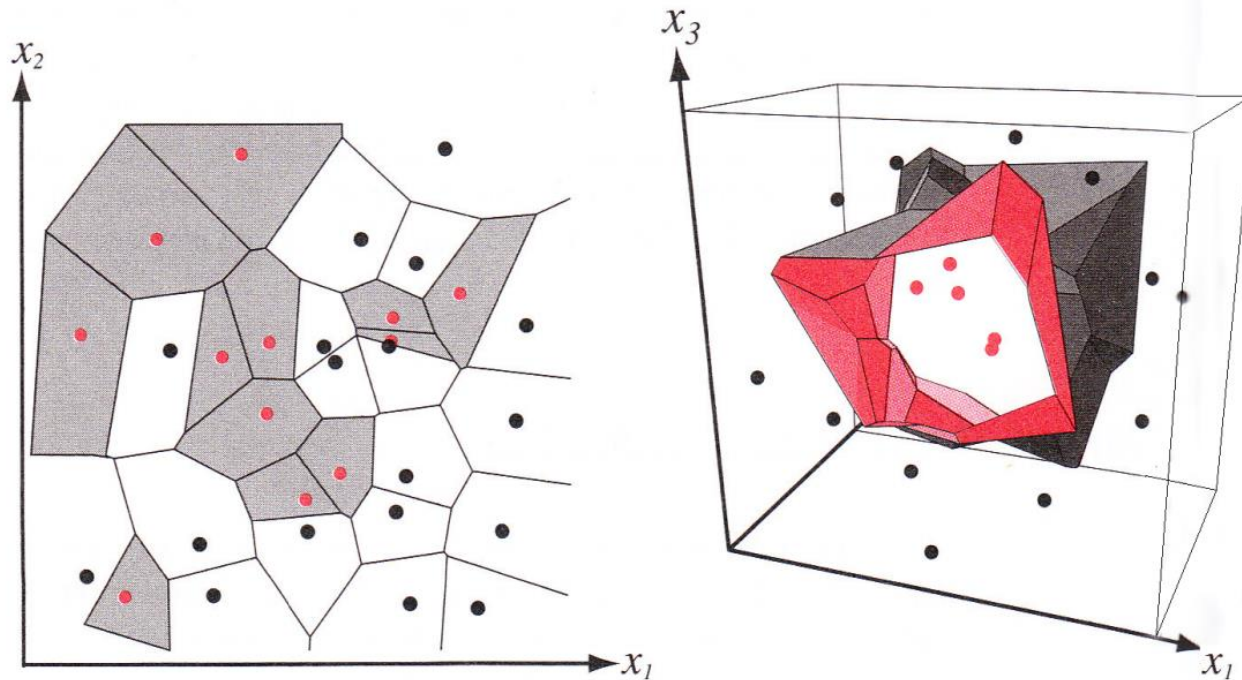
Výhody: Rychlé učení, jednoduchá implementace, snadné pochopení

Nevýhody: Pomalá predikce (mnoho výpočtů vzdáleností), při větších počtech příznaků nespolehlivý

K-D stromy jsou modifikací algoritmu K-NN, která využívá rozdělení trénovací množiny na oblasti.



Metoda je založena na odstranění přebytečných prvků, které neovlivňují klasifikaci.



- Každá třída je popsána nějakým pravděpodobnostním rozložením v prostoru příznaků
- $P(x|\omega_j)$ Podmíněná pravděpodobnost, že u objektu ze třídy ω_j pozorujeme x
- $P(\omega_j)$ Apriorní pravděpodobnost reflektuje naši apriorní znalost o objektu, aniž bychom objekt viděli
- $P(\omega_j|x)$ Aposteriorní pravděpodobnost, že objekt, na kterém pozorujeme x je ze třídy ω_j

Bayesovská rozhodovací teorie

If $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$, přiřad' $\mathbf{x}$ do ω_1 ,
else přiřad' $\mathbf{x}$ do ω_2 .

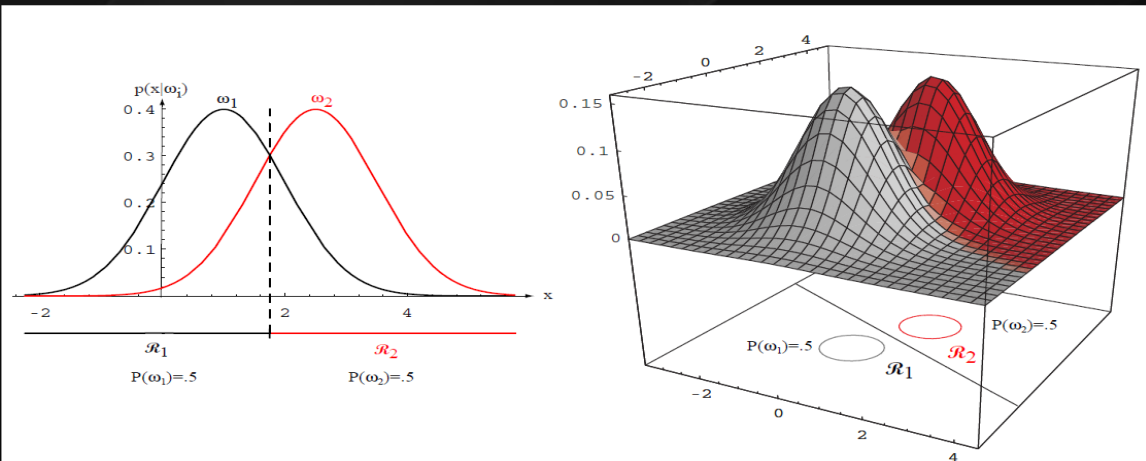
$P(\omega_1|\mathbf{x})$ a $P(\omega_2|\mathbf{x})$ jsou aposteriorní
pravděpodobnosti.

Minimalizujeme střední pravděpodobnost chyby.

Pro výpočet aposteriorních pravděpodobností lze využít Bayesův vztah

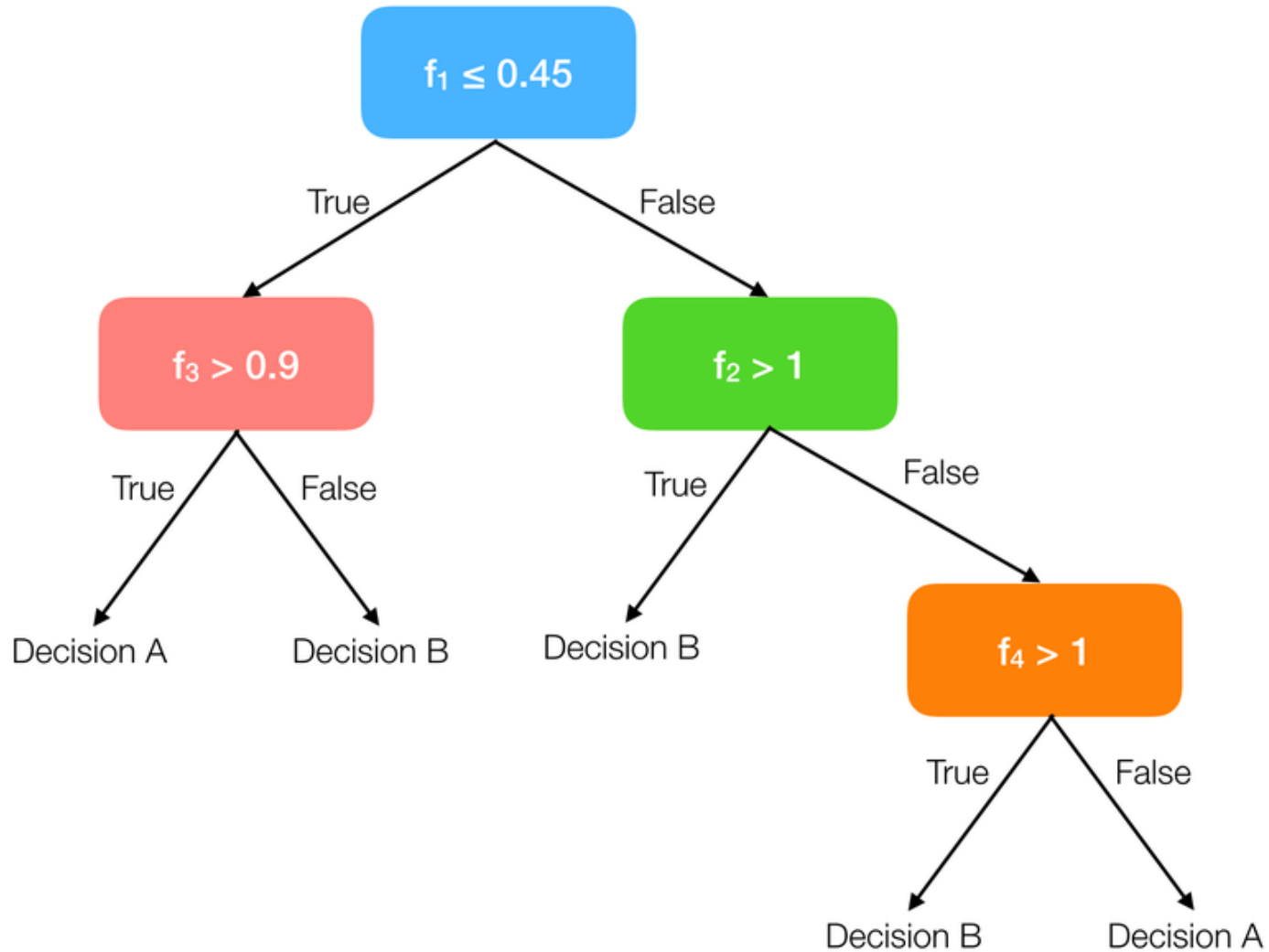
$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})}$$

Učení je odhad parametrů pravděpodobnostních rozložení $p(\mathbf{x}|\omega_j)$



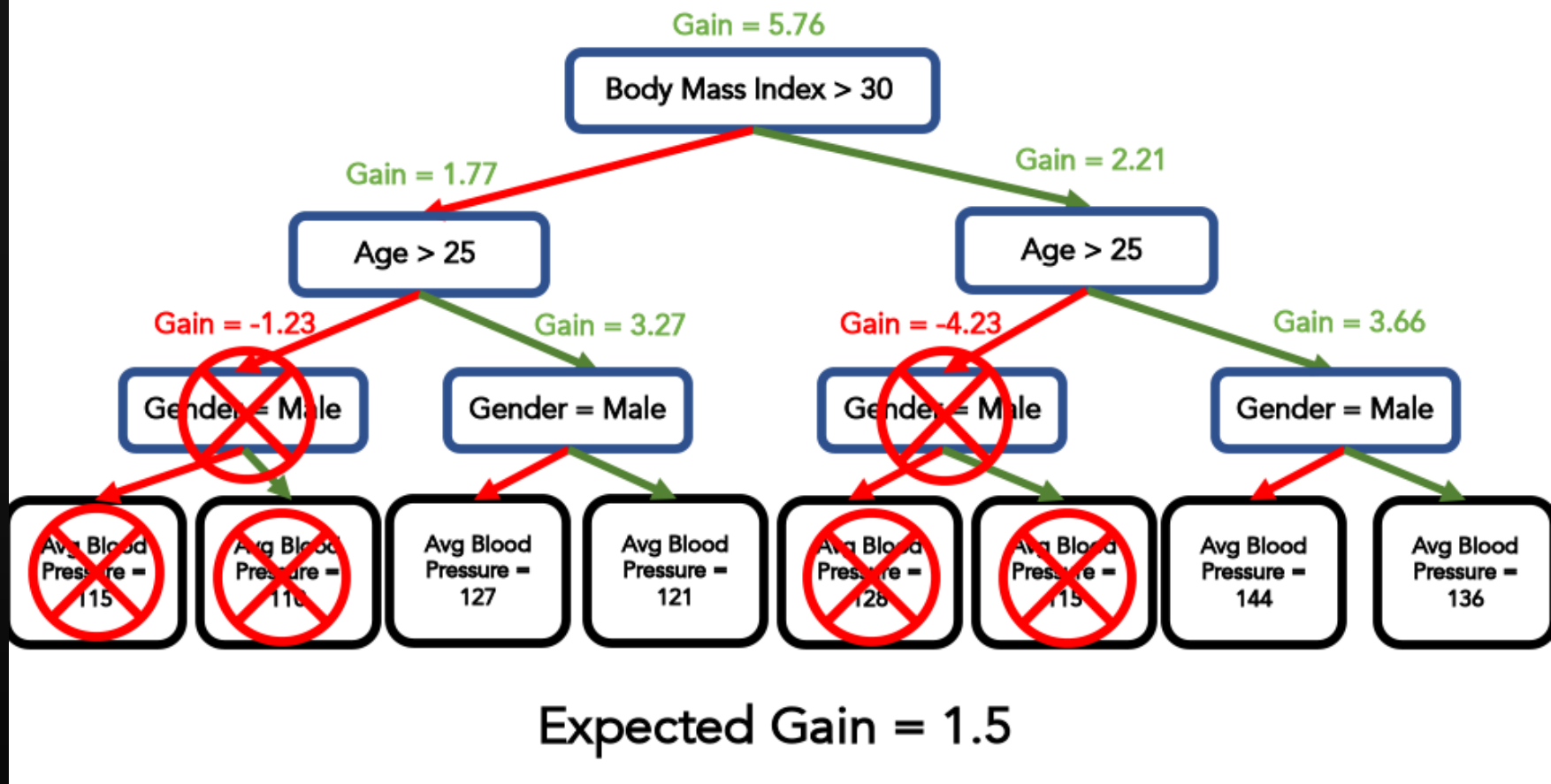
Řešení klasifikačních úloh pomocí rozhodovacích stromů vyžaduje dva kroky:

1. Indukce rozhodovacího stromu (podle trénovacích dat)
2. Použij rozhodovací strom a urči třídu



Prořezávání rozhodovacích stromů

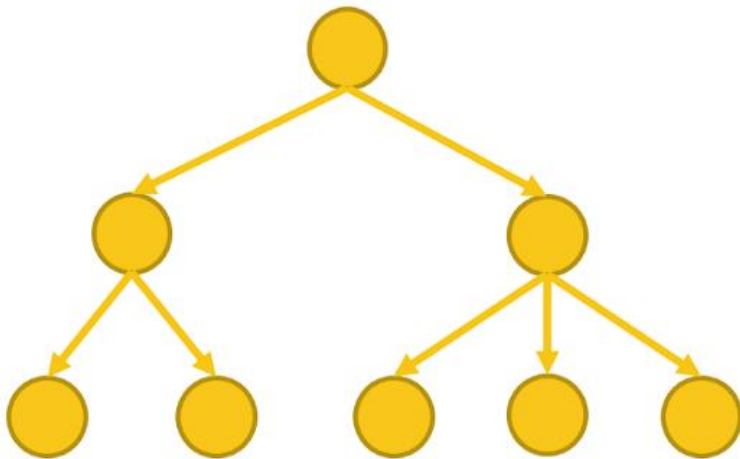
CIIRC



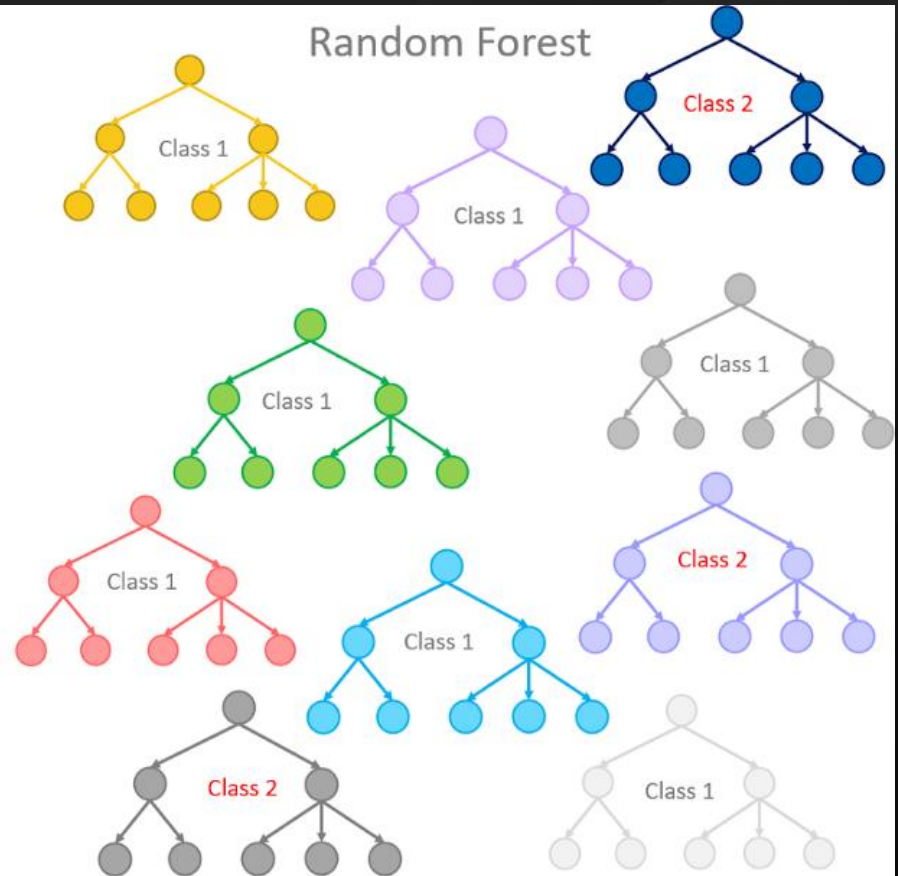
Random decision forest

CIIRC

Single Decision Tree



Random Forest



Výhody:

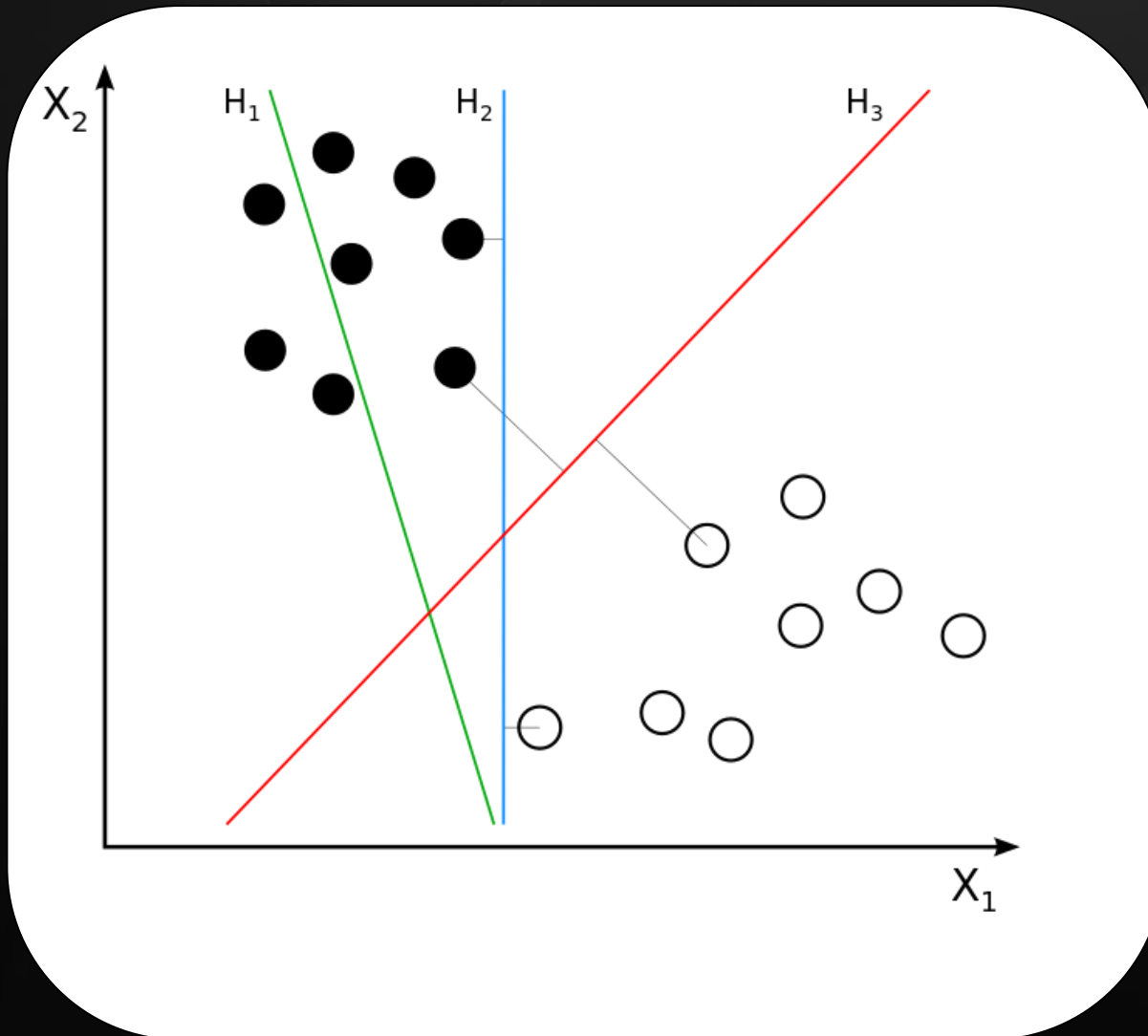
- Jednoduché
- Efektivní
- Extrakce jednoduchých pravidel
- Použitelné i pro velké datasety

Nevýhody:

- Obtížnější zpracování spojitých dat (kategorizace atributů)
- Obtížné zpracování při chybějících údajích

Support-vector machine, SVM

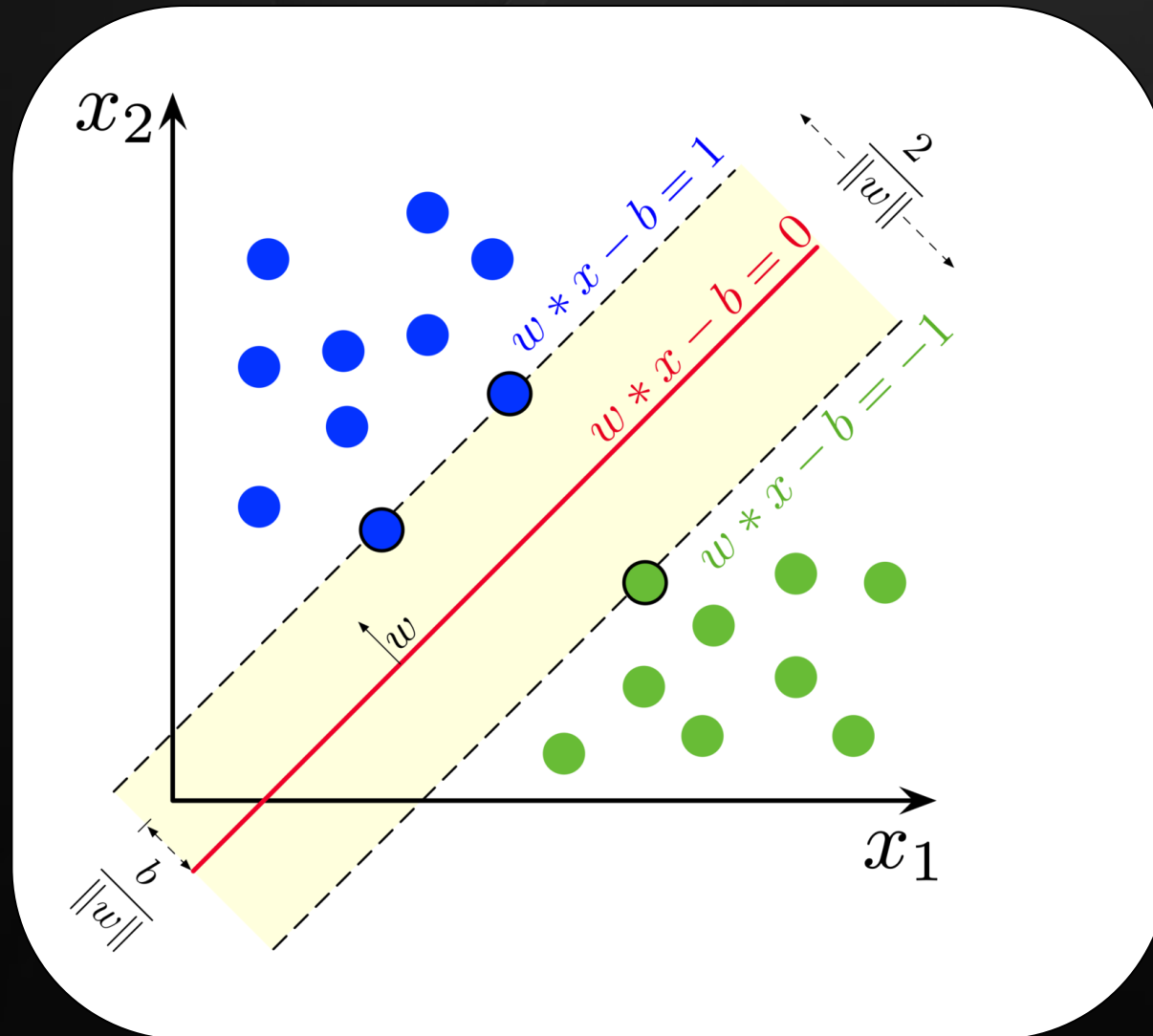
CIIRC



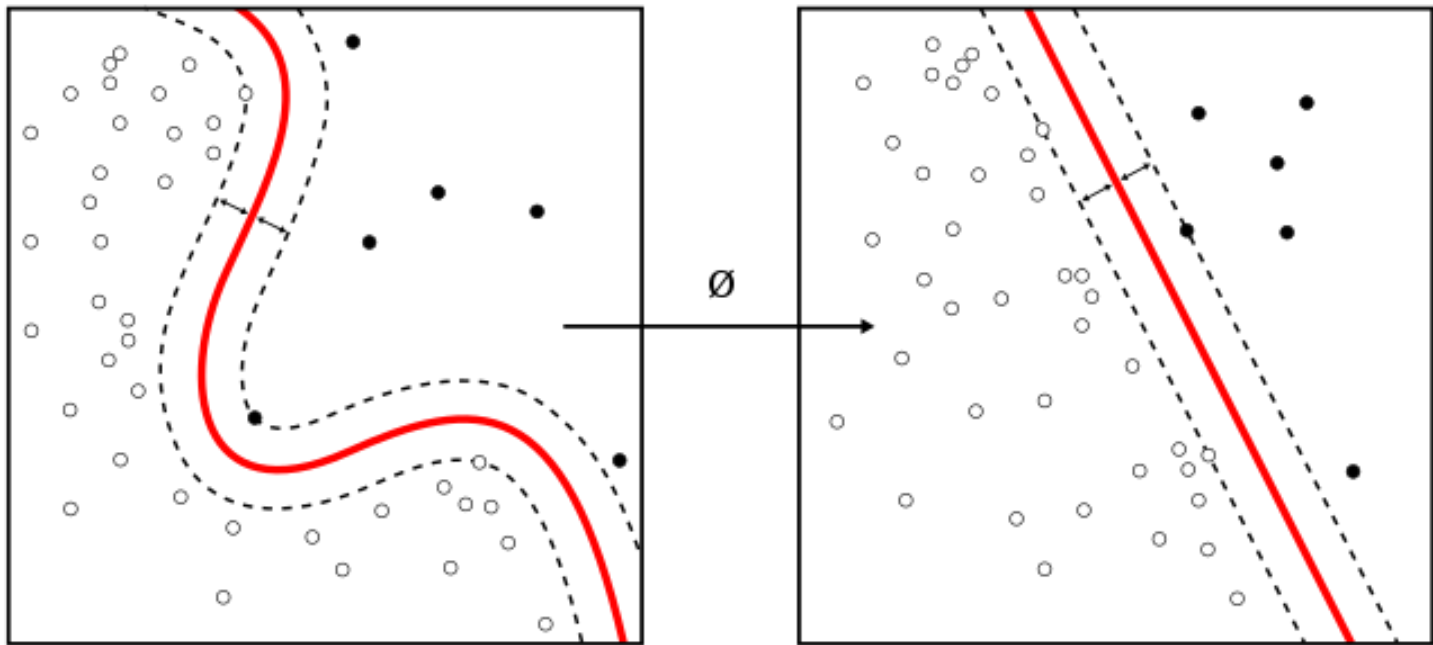
H2? H3?

Support-vector machine, SVM

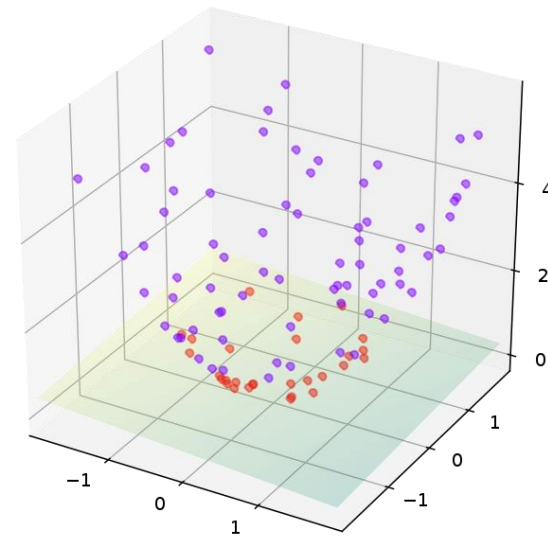
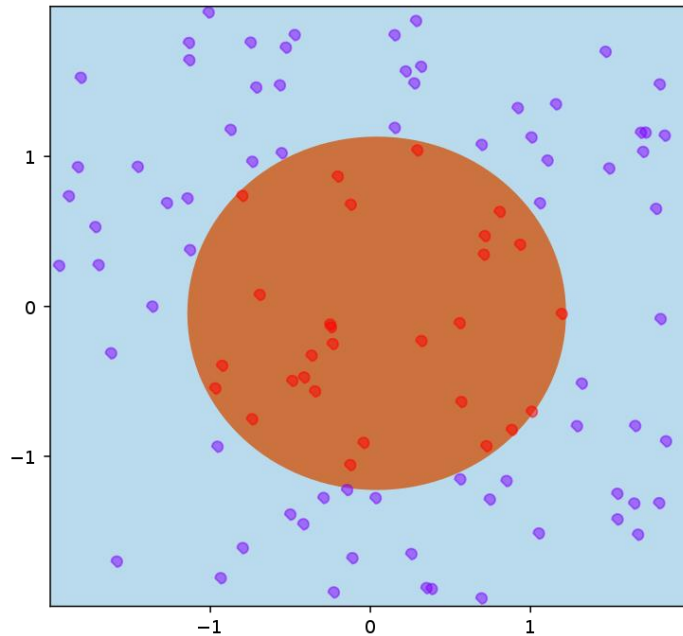
CIIRC



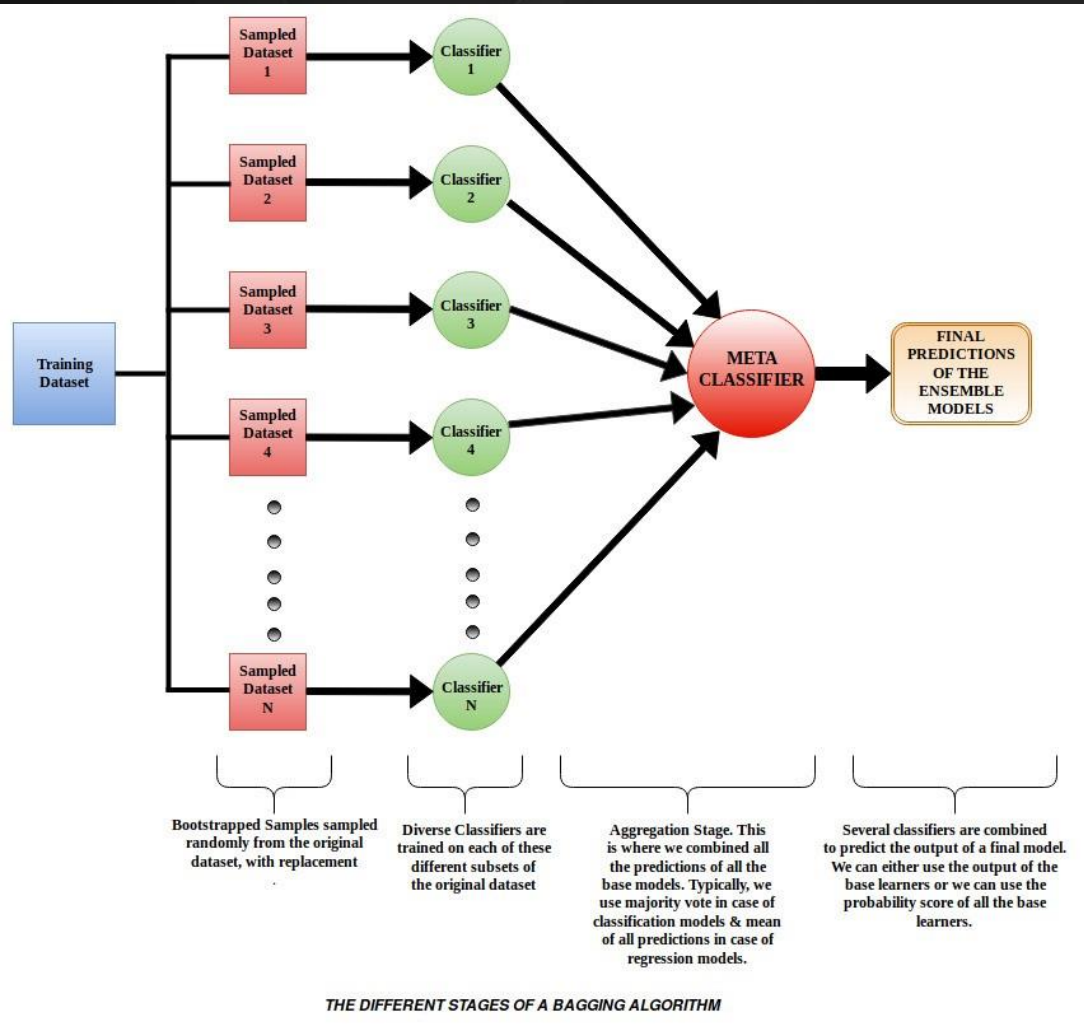
Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.



A training example of SVM with kernel given by $\phi((a, b)) = (a, b, a^2 + b^2)$



- Bagging
- Boosting
- Stacking

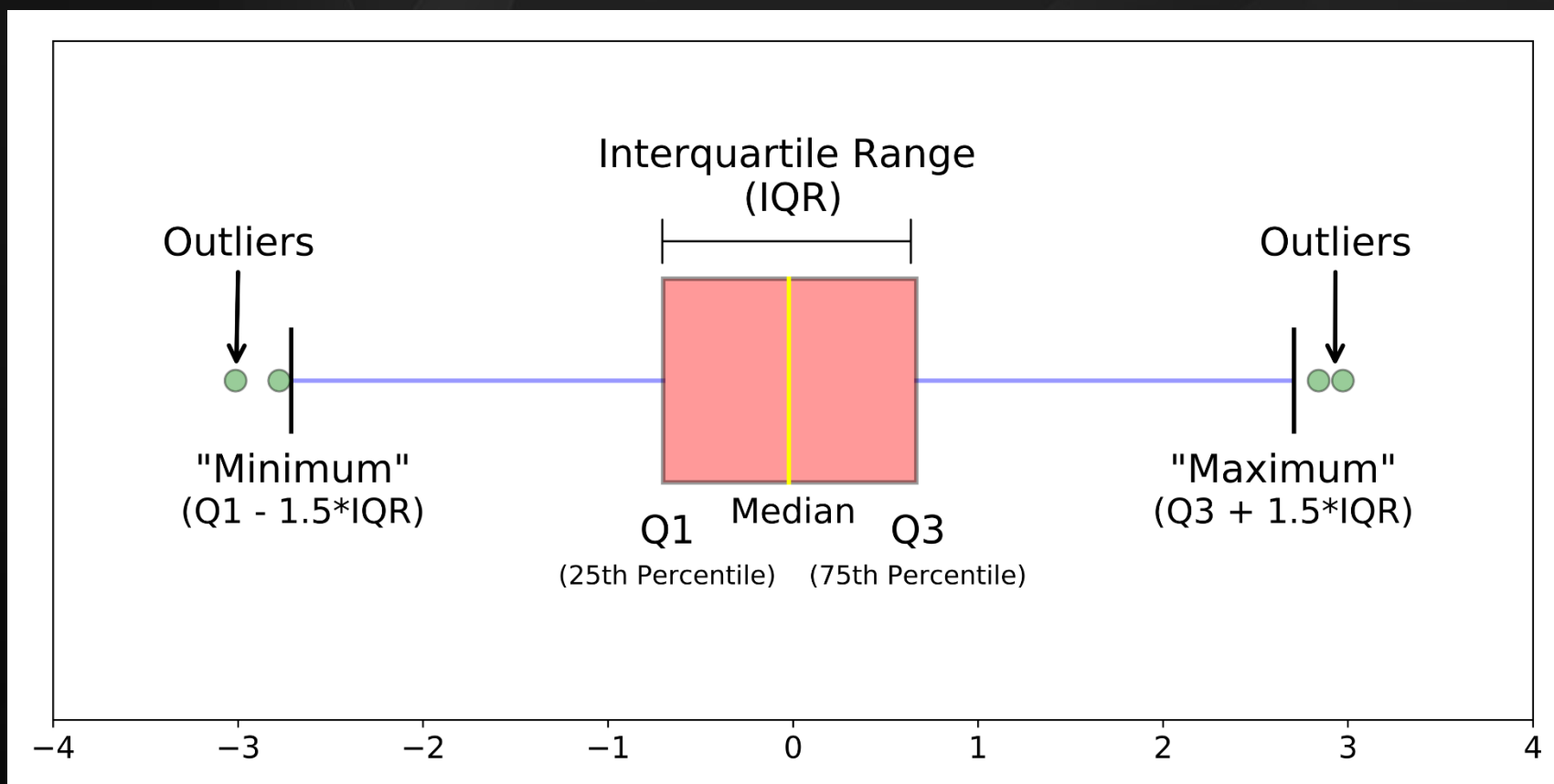


Outlier je neobvyklá hodnota, která se výrazně odlišuje od normálních objektů jako kdyby byla generována jiným mechanismem.

Může mít velký vliv na výsledný klasifikátor, proto je většinou lépe ho odstranit

Zobrazení typu Boxplot

CIIRC



Velké hodnoty příznaku mohou pro některé klasifikátory způsobovat dominanci onoho příznaku při rozhodování.

Normalizace, základní dvě metody:

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

$$x' = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

Některé typy klasifikátorů vyžadují normalizaci. Např. KNN klasifikátor.

U jiných to není potřebné, např. rozhodovací stromy.

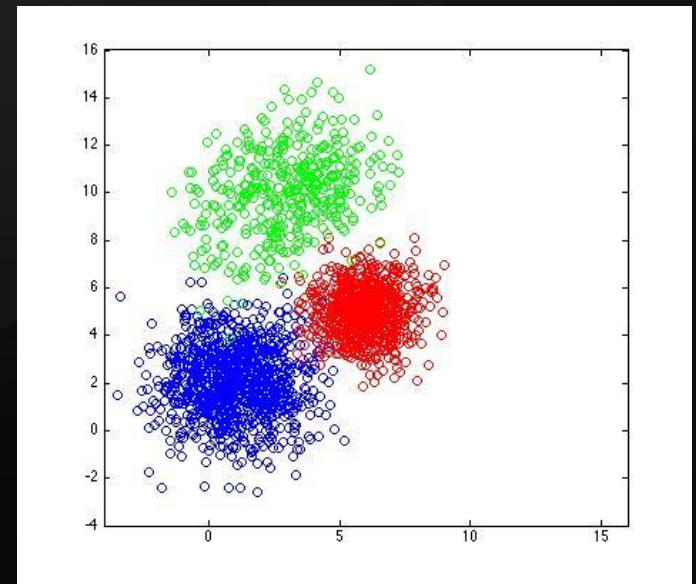
Hledání podsouboru příznaků, který maximalizuje nějaké kritérium.

Kritérium

Přesnost klasifikátoru, inter-intra class distance, vzájemná informace, entropie.....

Prohledávání

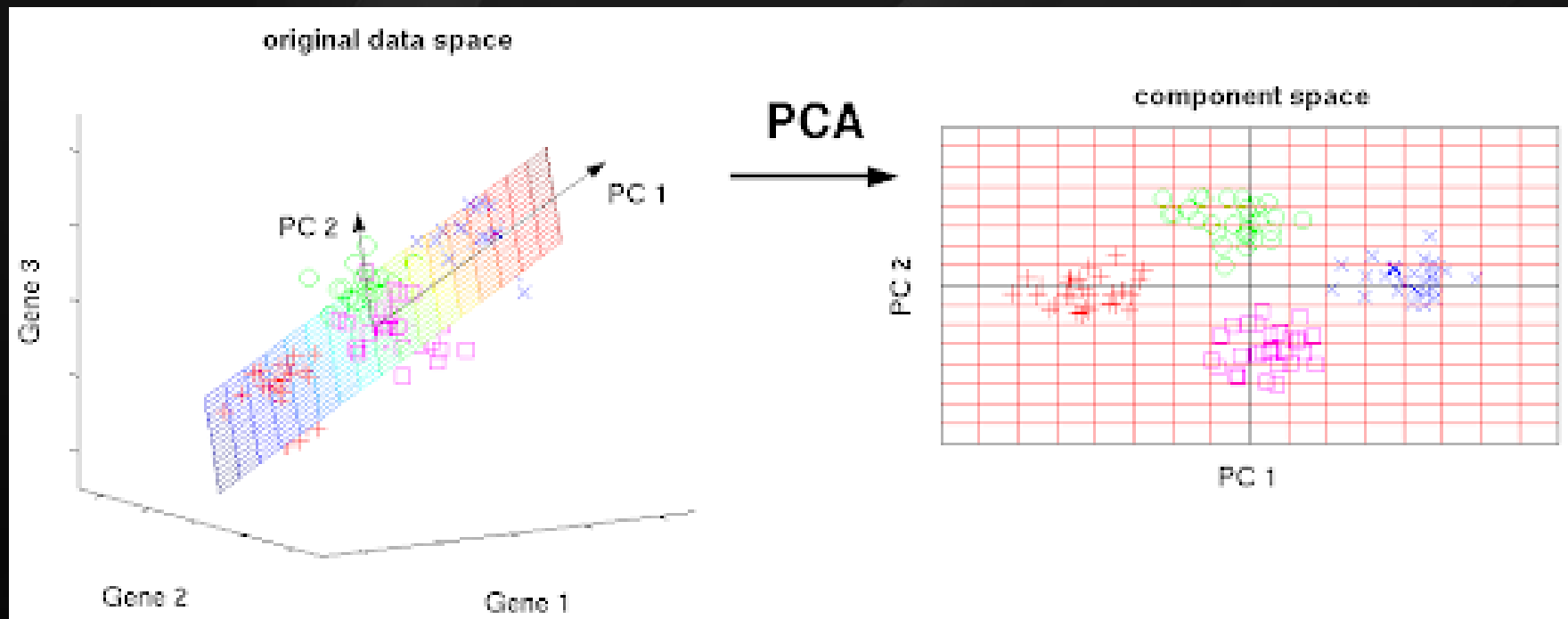
Individual ranking
Sequential Forward Selection
Sequential Backward Elimination
Evolutionary algorithms



Redukce příznakového prostoru

Např. Analýza hlavních komponent (Principle Component Analysis - PCA)

Ukázka snížení dimenze z 3D do 2D



Chceme predikovat, zda pacient trpí např. Wilsonovou chorobou.

Prevalence choroby je 1:25.000
Budu-li stále predikovat negativní výsledek,
klasifikuju 99.996% případů správně!!!!!!!!!!!!

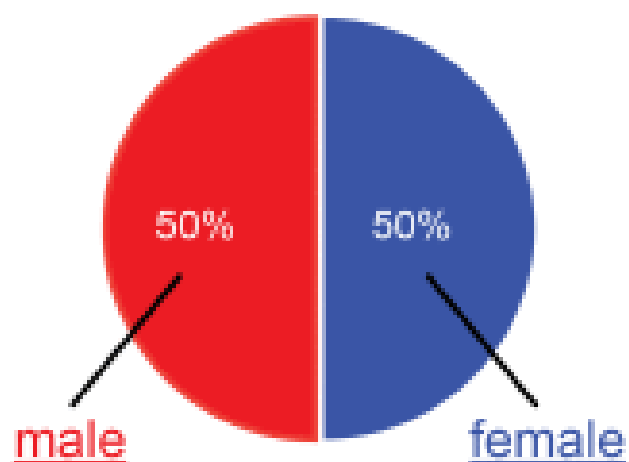
**ACCURACY (PŘESNOST) KLASIFIKÁTORU NEMUSÍ BÝT
VHODNÁ MÍRA**

Nerovnoměrné zastoupení tříd

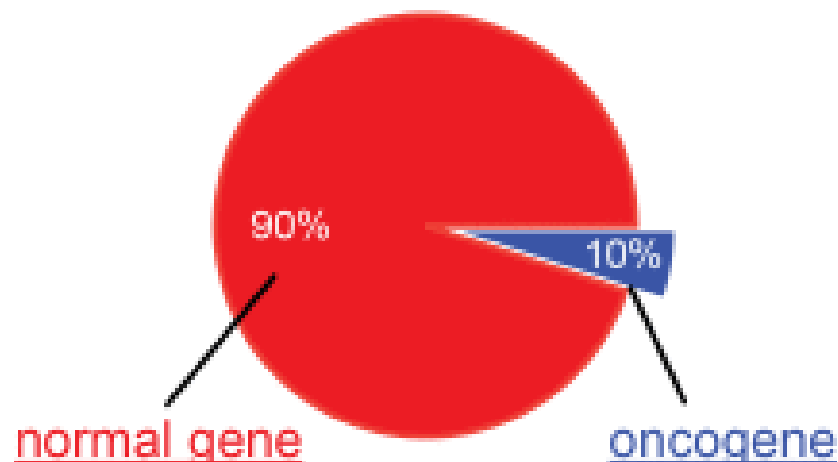
CIIRC



Example of balanced and imbalanced data



Negatives $\approx$ Positives
Balanced

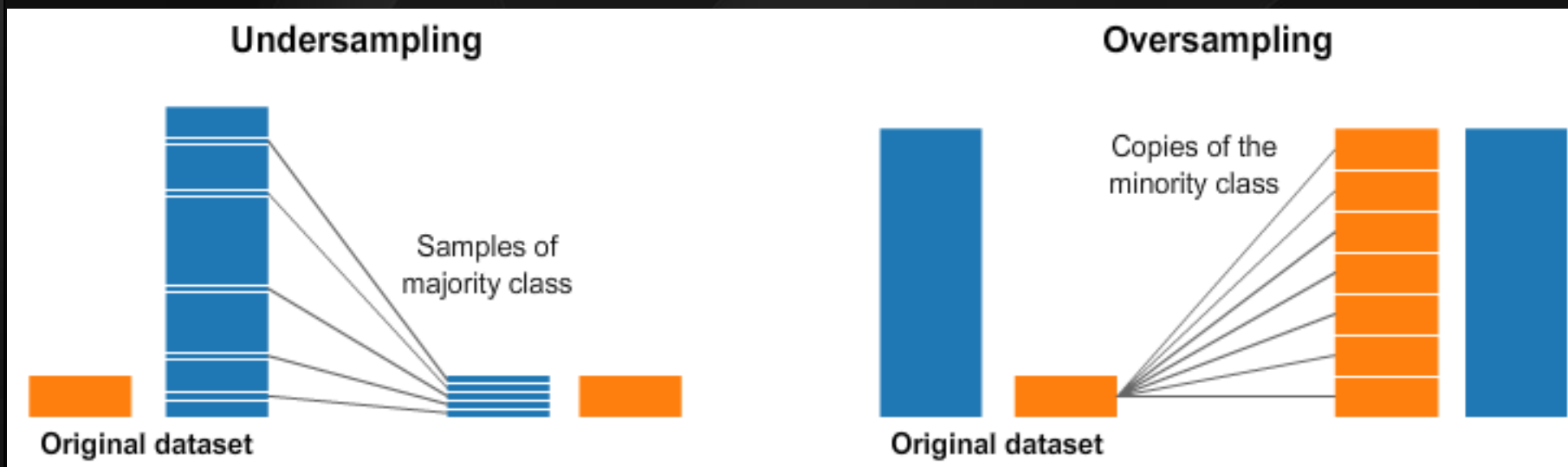


Negatives $>$ Positives
Imbalanced

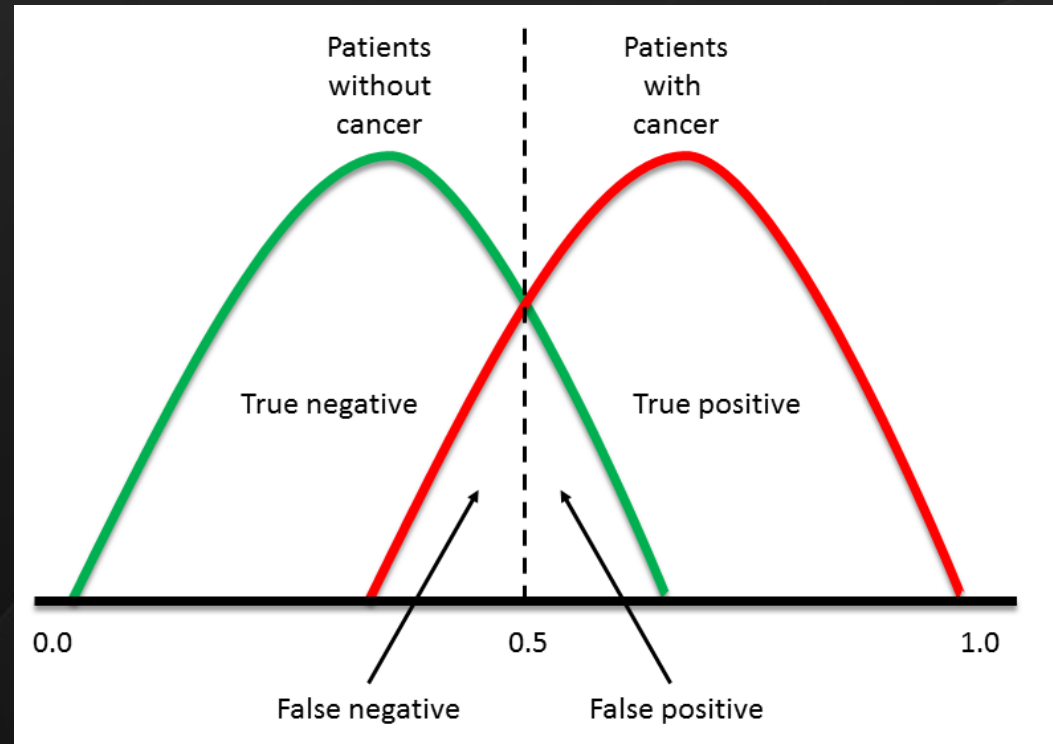
Nerovnoměrné zastoupení tříd

Pokud jsou třídy nevyvážené, může docházet k chybnému trénování klasifikátoru.

Řešení – např. **Undersampling**, nebo **Oversampling**.



positive samples (P)
negative samples (N)
true positive (TP)
true negative (TN)
false positive (FP)
false negative (FN)



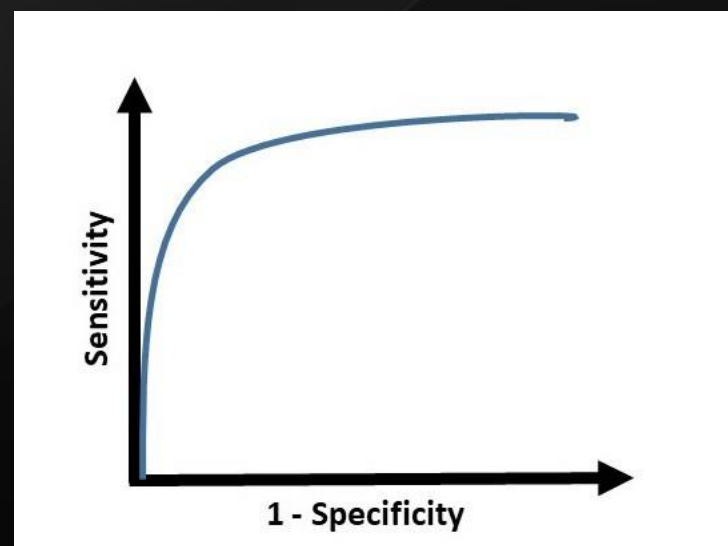
$$\text{ACCURACY} = (TP+TN)/(TP+FP+FN+TN)$$

$$\text{SENSITIVITY} = TP/P = TP/(TP+FN)$$

$$\text{SPECIFICITY} = TN/N = TN/(TN+FP)$$

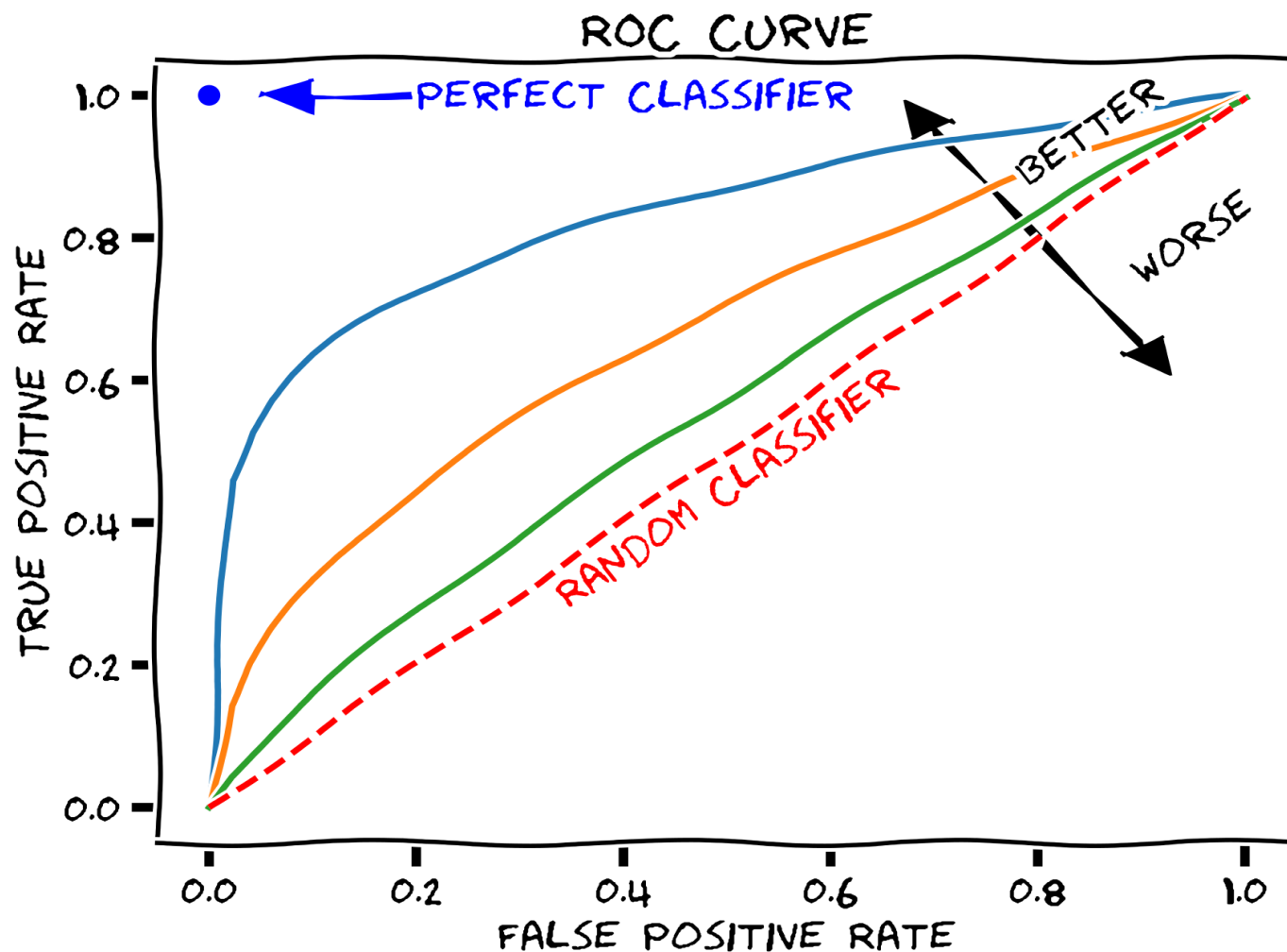
ROC (Receiver Operating Characteristic) křivka je nástroj pro hodnocení a optimalizaci binárního klasifikačního systému (testu), který ukazuje vztah mezi **specificitou a senzitivitou** pro všechny přípustné hodnoty prahu.

Z ROC je možné spočítat AUC (Area Under Curve)



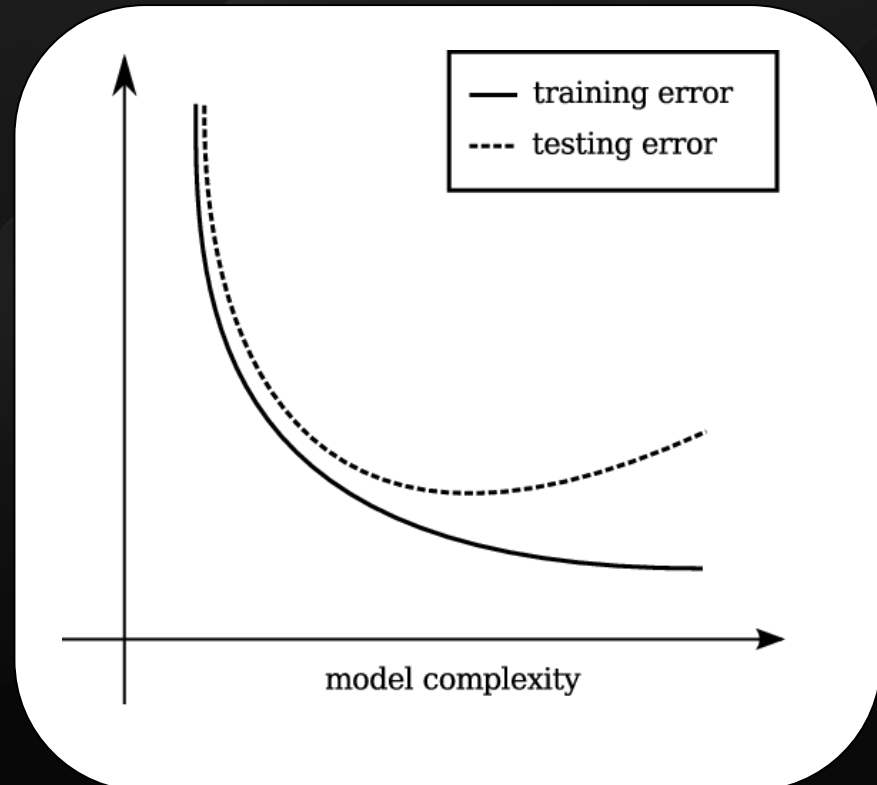
ROC křivka

CIIRC

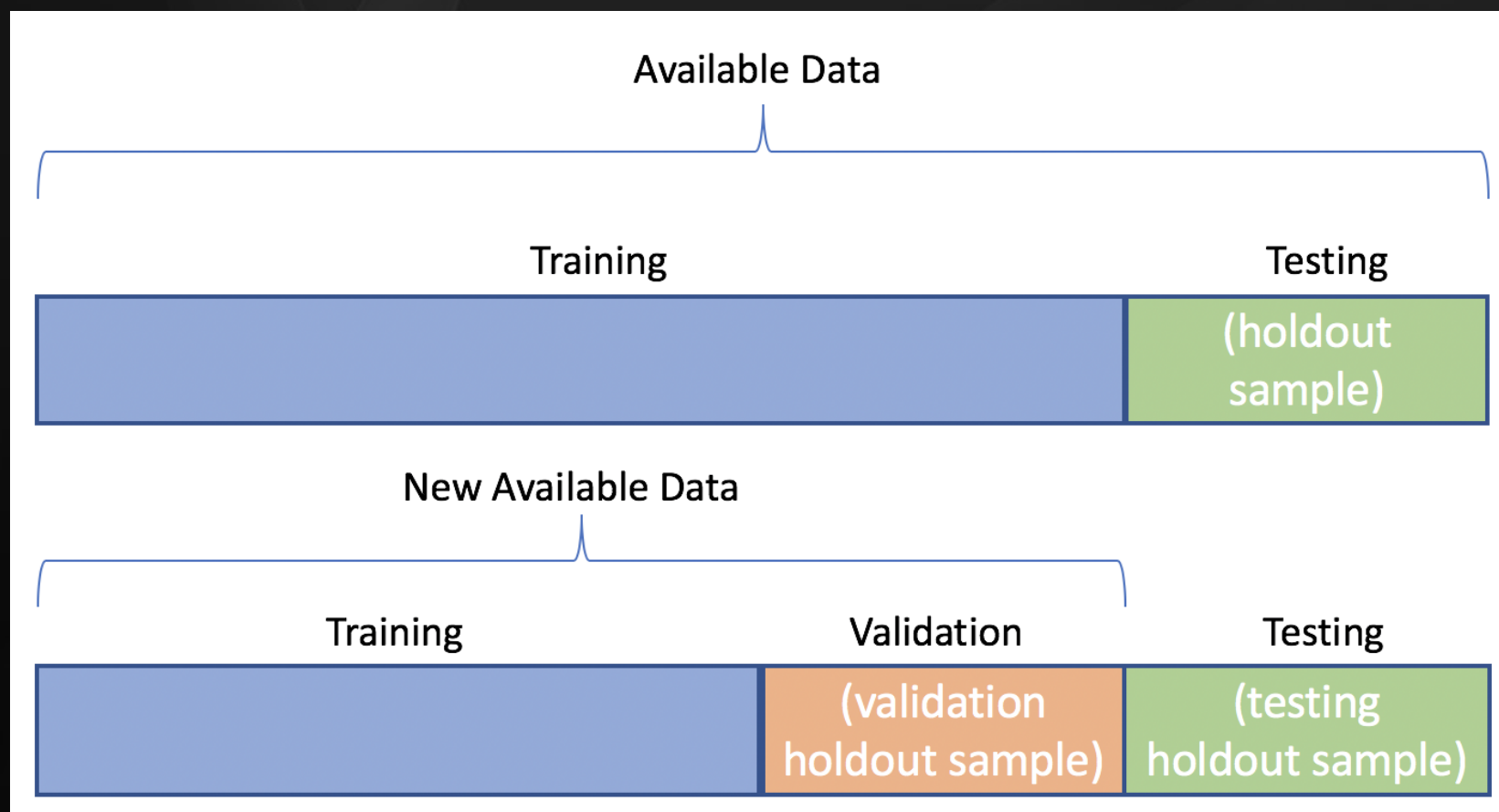


Zjištění generalizační schopnosti klasifikátoru

Tj. zda klasifikátor funguje i na datech, která nikdy neviděl

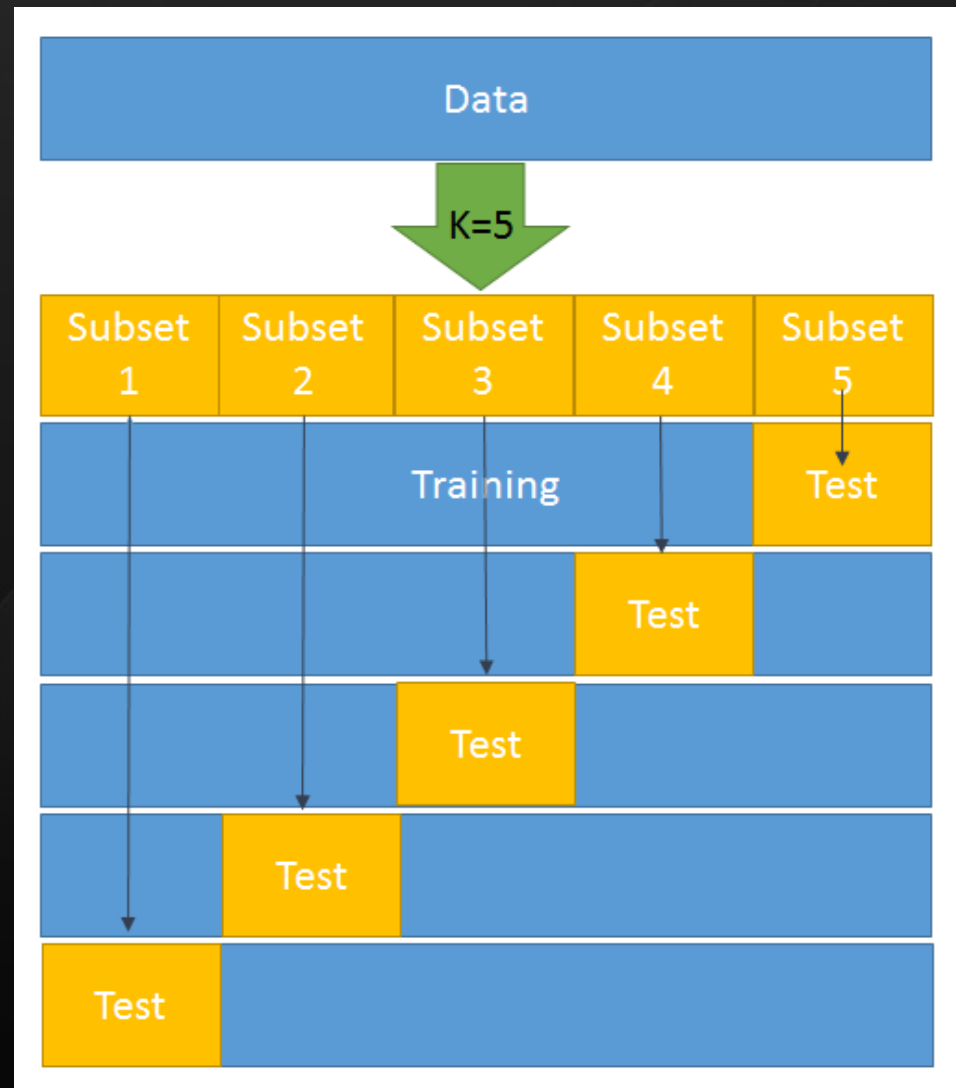


Hold-out metoda – rozseknutí dat na trénovací a testovací data



Křížová validace –
rozdělení na části,
iterativní použití vždy
jen jedné části jako
testovací a zbytku
jako trénovací

**Např. 5 skupin
(foldů).**



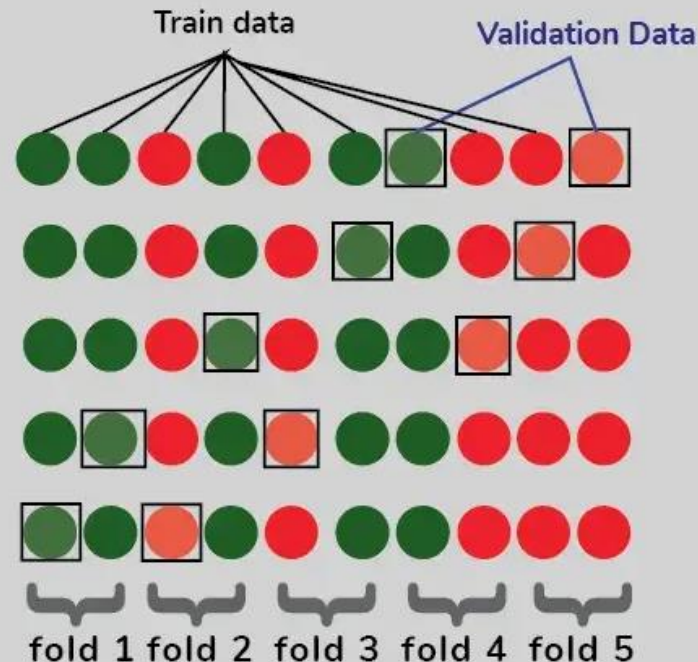
Stratify K-Fold CV

Lets we have total 10 datapoints and we have set $k=5$

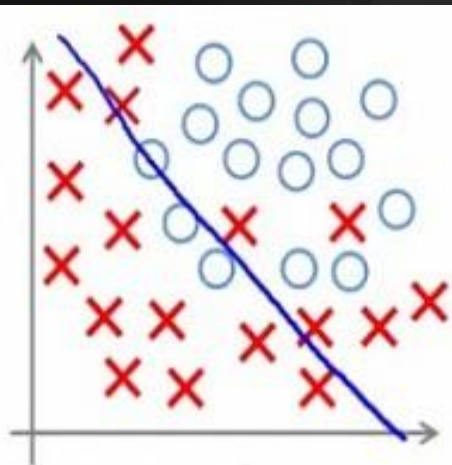
● Class 1

● Class 0

The process will iterate or you can say the dataset n-one fold will pas through model $k=5$ times, everytime model will be tested on one fold. Both sets contains equal ratio of both classes.

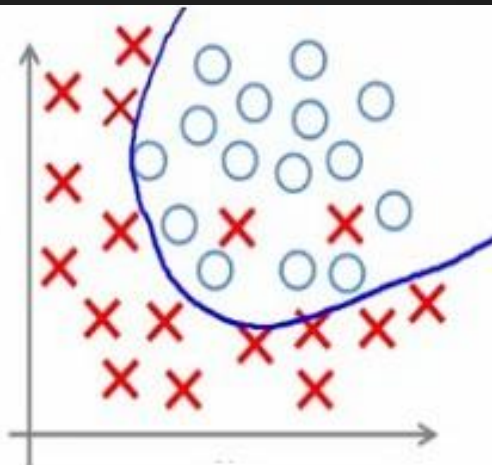


Klasifikátor učený na omezeném vzorku dat nesmí být příliš složitý, jinak dochází k tzv. **OVER-FITTINGu** (přeučení)

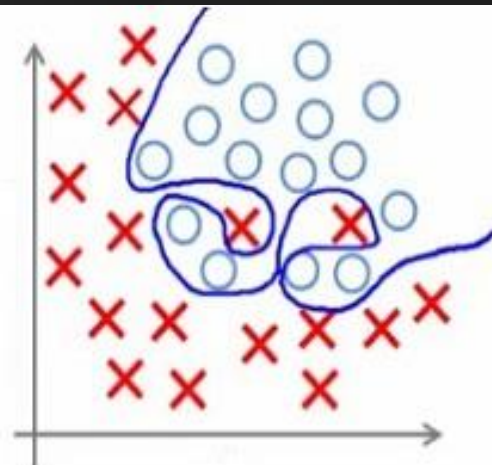


Under-fitting

(too simple to explain the variance)



Appropriate-fitting



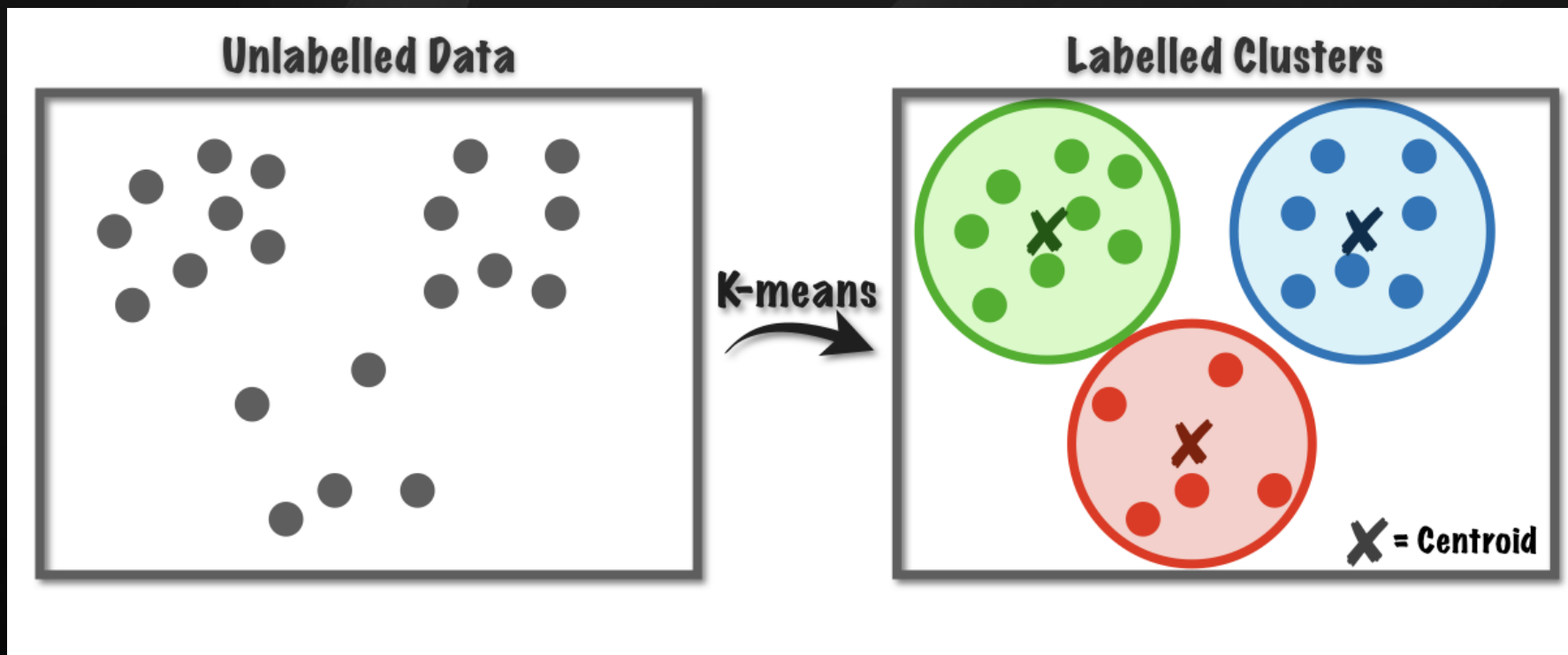
Over-fitting

(forcefitting -- too good to be true)

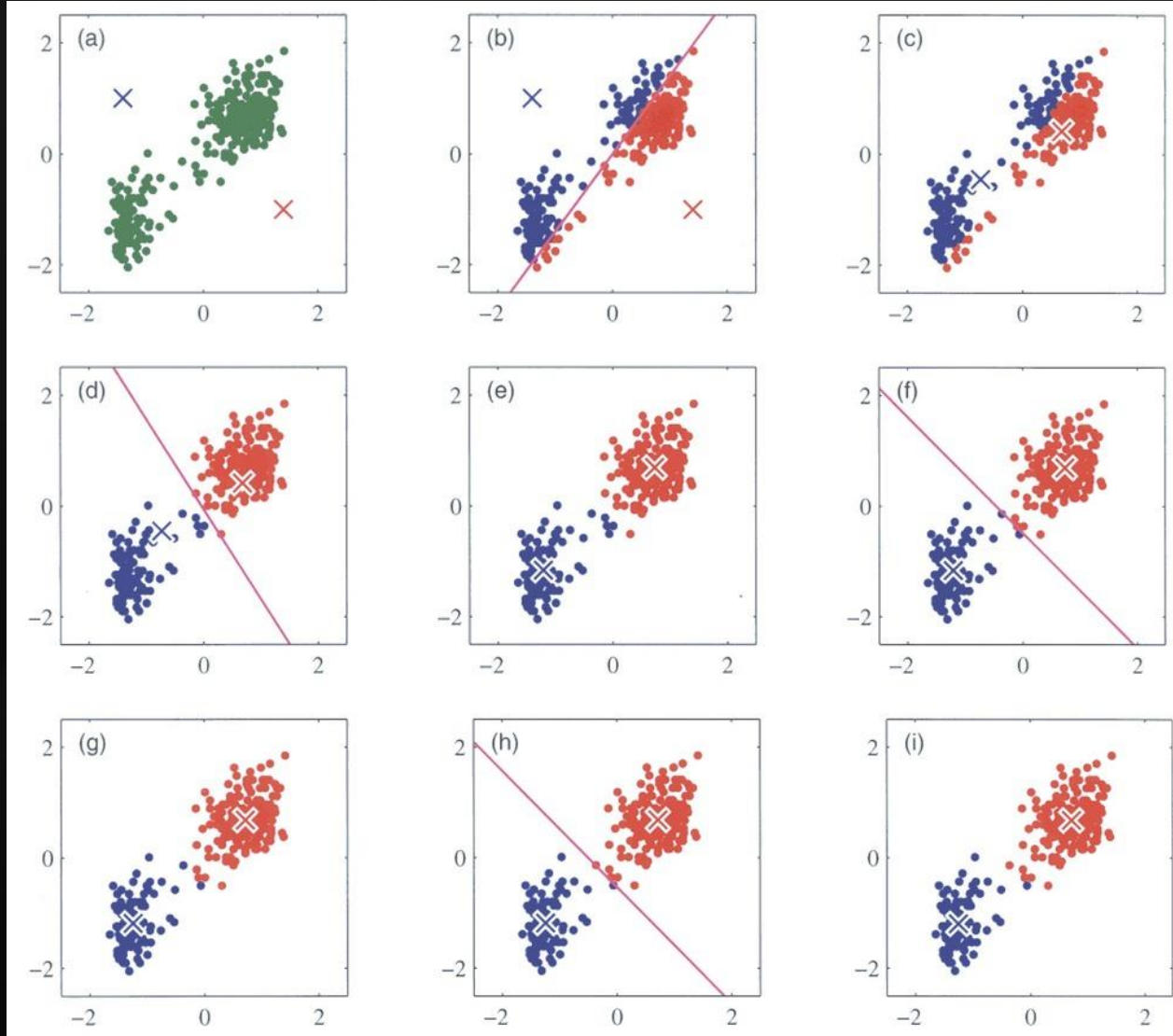
- Měření podobnosti objektů
- Existují různé přístupy, jak shlukovat objekty na základě jejich vzdálenosti či podobnosti.

K-means

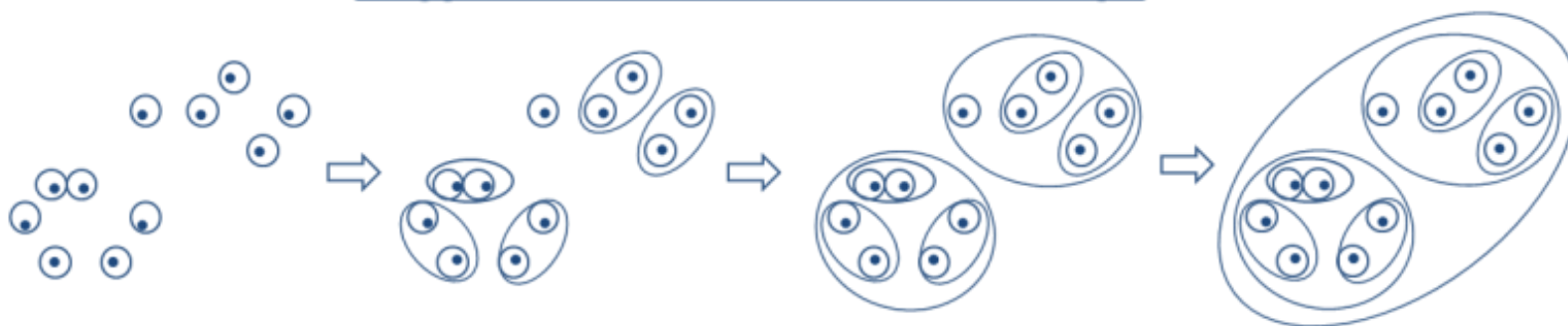
- Určíme hodnotu K
- V datech nalezneme K shluků
- Výhoda: jednoduchý iterační algoritmus



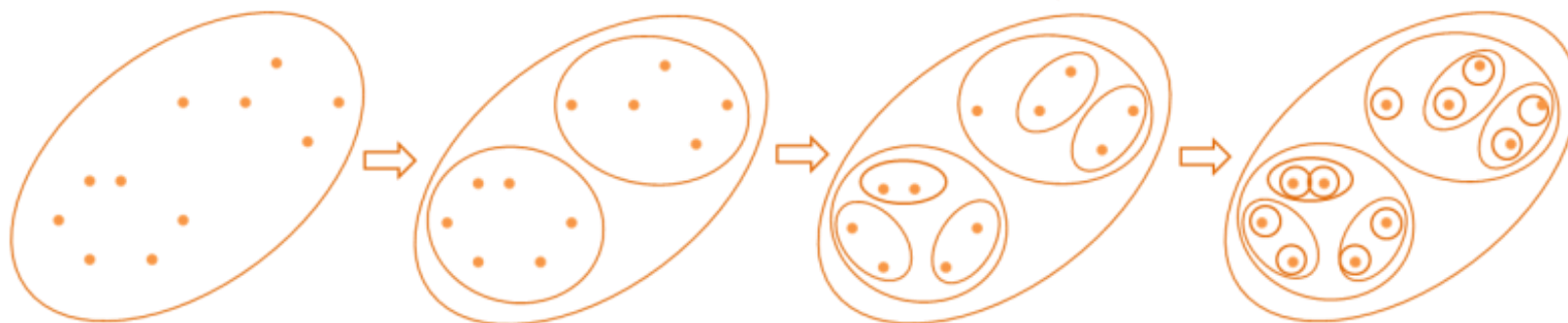
K-means algoritmus



Agglomerative Hierarchical Clustering



Divisive Hierarchical Clustering



- Kvalitní příznaky (jejich počet $\ll$ počet prvků datasetu)
- Vhodná volba trénovací a testovací sady, aby nedocházelo k preučování klasifikátoru
- Vhodný způsob hodnocení klasifikátoru (přesnost klasifikace, případně Senzitivita/Specificita/ROC)
- Volba klasifikátoru dle typu úlohy, lépe je začít jednoduchým modelem