Klasifikace a shluková analýza

CIIRC ČESKÝ INSTITUT INFORMATIKY, ROBOTIKY A KYBERNETIKY

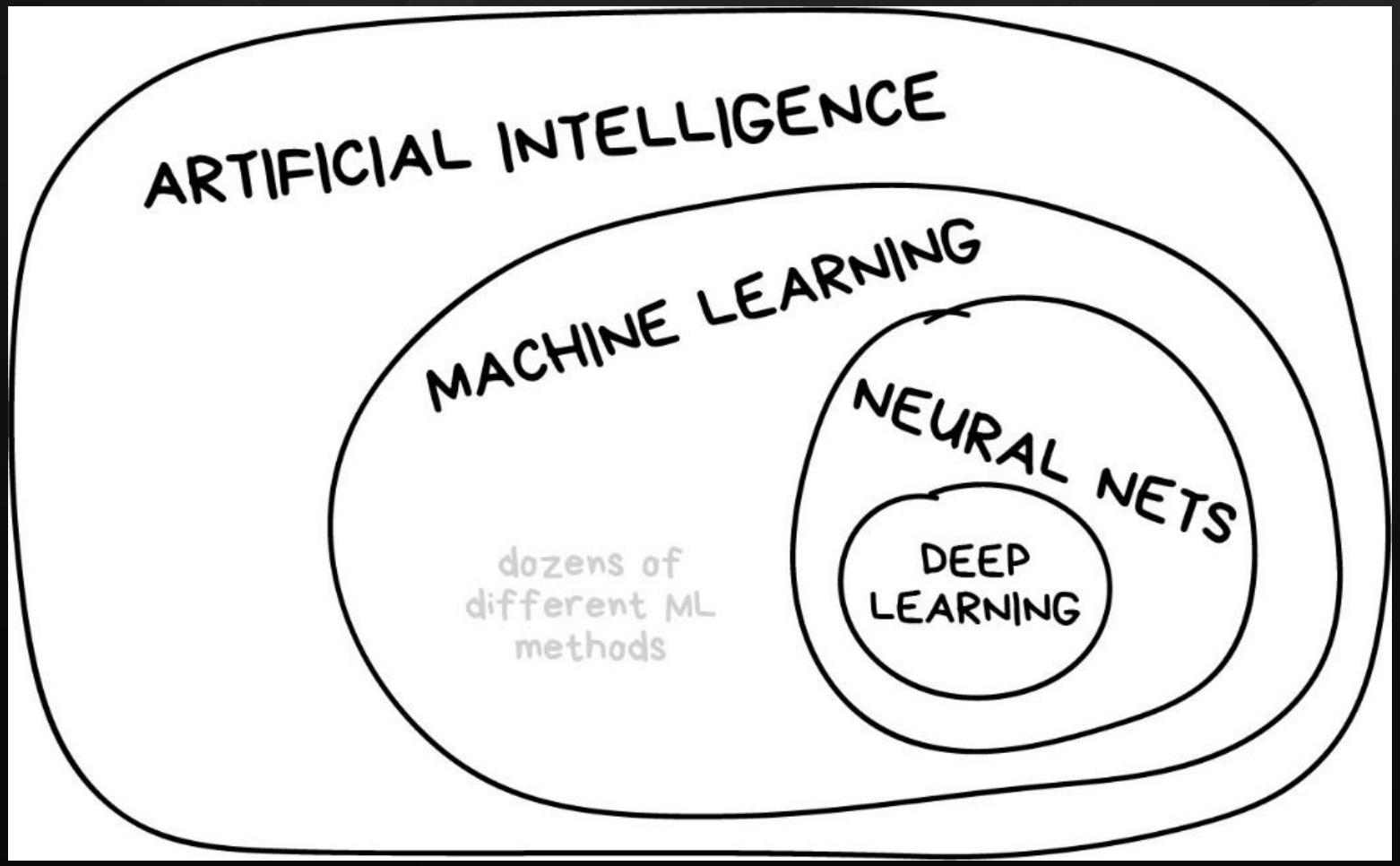
Václav Gerla

KOGNITIVNÍ SYSTÉMY A NEUROVĚDY

Český institut informatiky robotiky a kybernetiky

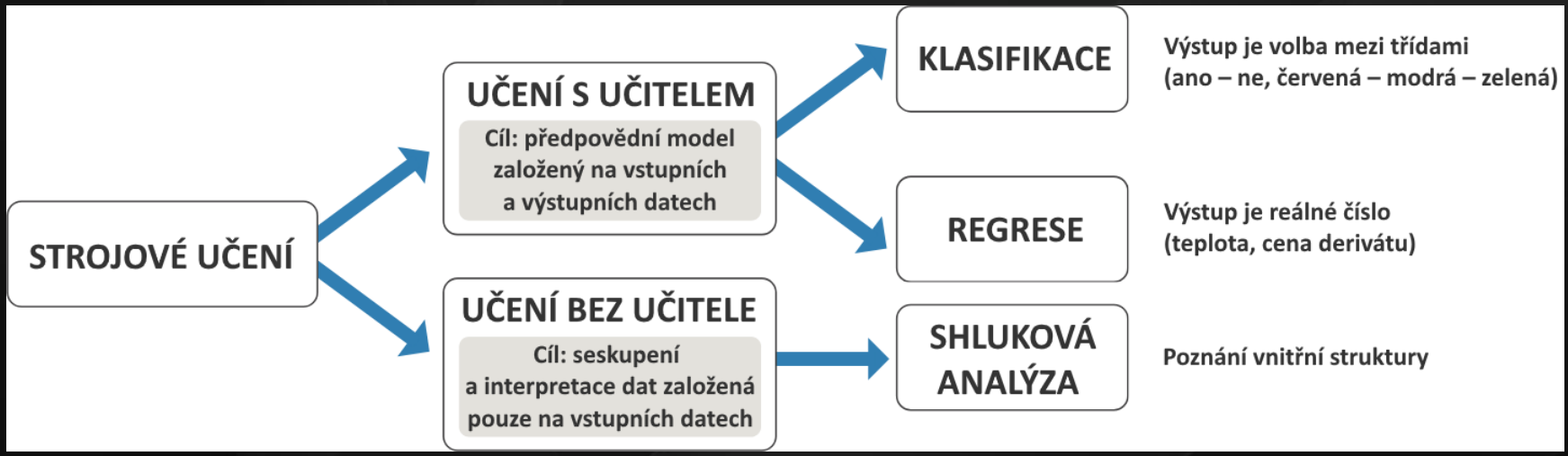
Strojové učení – Machine Learning (ML)

CIIRC ČESKÝ INSTITUT INFORMATIKY, ROBOTIKY A KYBERNETIKY

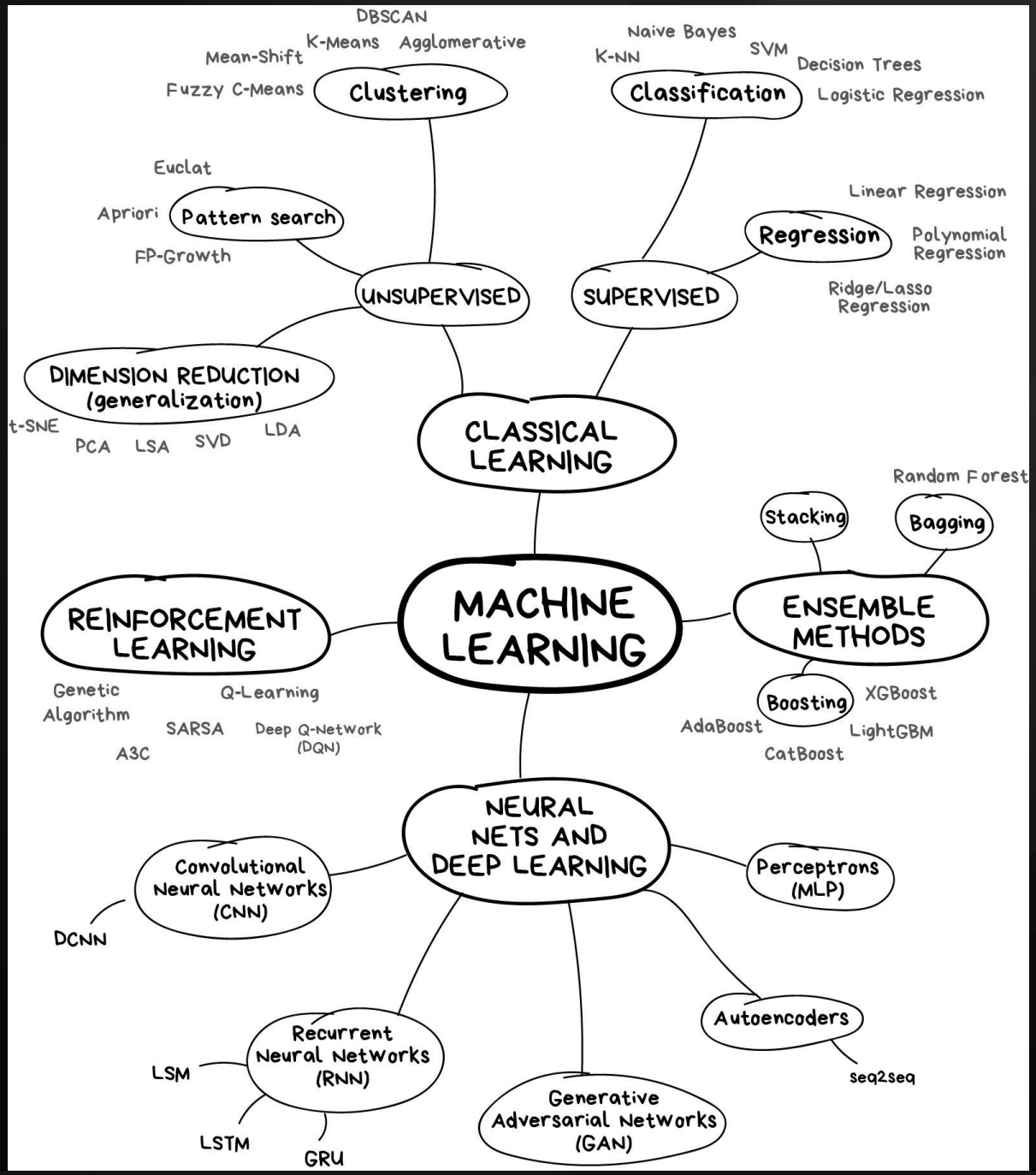


https://vas3k.com/blog/machine_learning/index.html

Rozdělení strojového učení podle typu úlohy



https://humusoft.cz/blog/?view=20180817-strojove-uceni



CIIRC ČESKÝ INSTITUT INFORMATIKY, ROBOTIKY A KYBERNETIKY

https://vas3k.com/blog/machine_learning/index.html

Co je klasifikace?

CIIRC

- 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

Klasifikace – příklad 1

CIIRC

- 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

Klasifikace – příklad 2

CIIRC

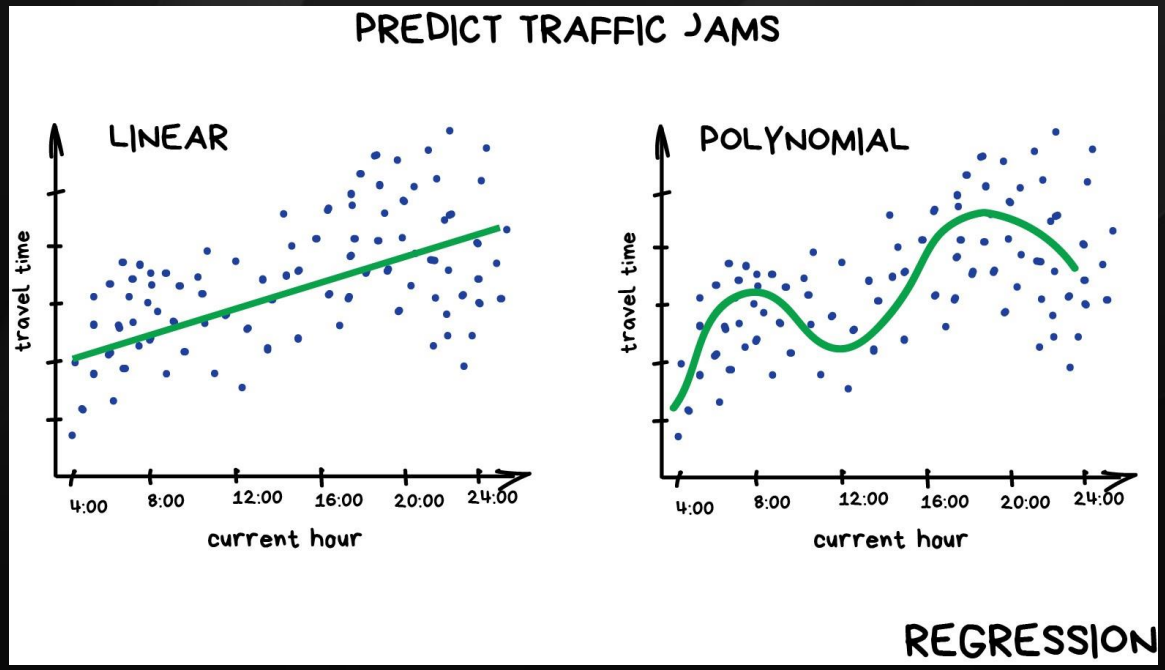
- Máme sadu možných symptomů, výsledků testů a vyšetření provedených na pacientovi
  - Např. měření krevního tlaku, saturace O2, ohodnocený záznam celonočního EEG záznamu, čas usnutí a probuzení, apod.

- Chceme predikovat, zda pacient trpí např. spánkovou poruchou.

Co je regrese?

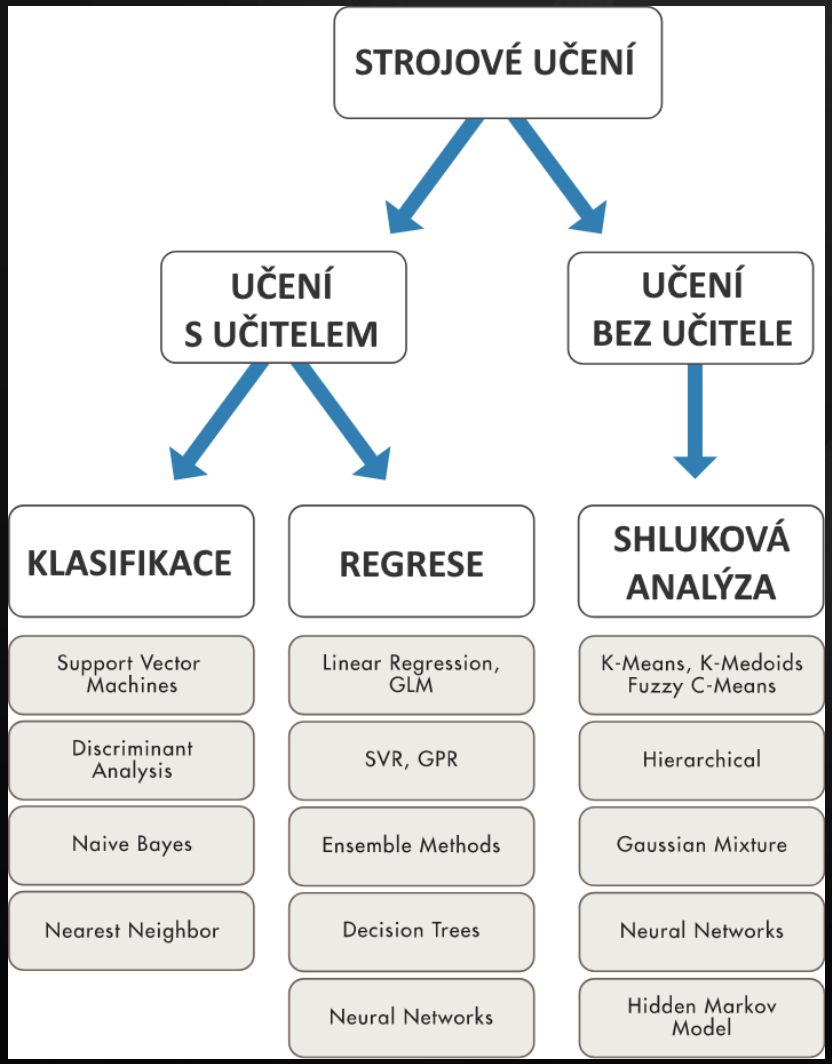
CIIRC

- Regrese je klasifikace, kde předpovídáme číslo místo kategorie
- Příklad: predikce dopravní situace v závislosti na čase, kdy vyjedeme



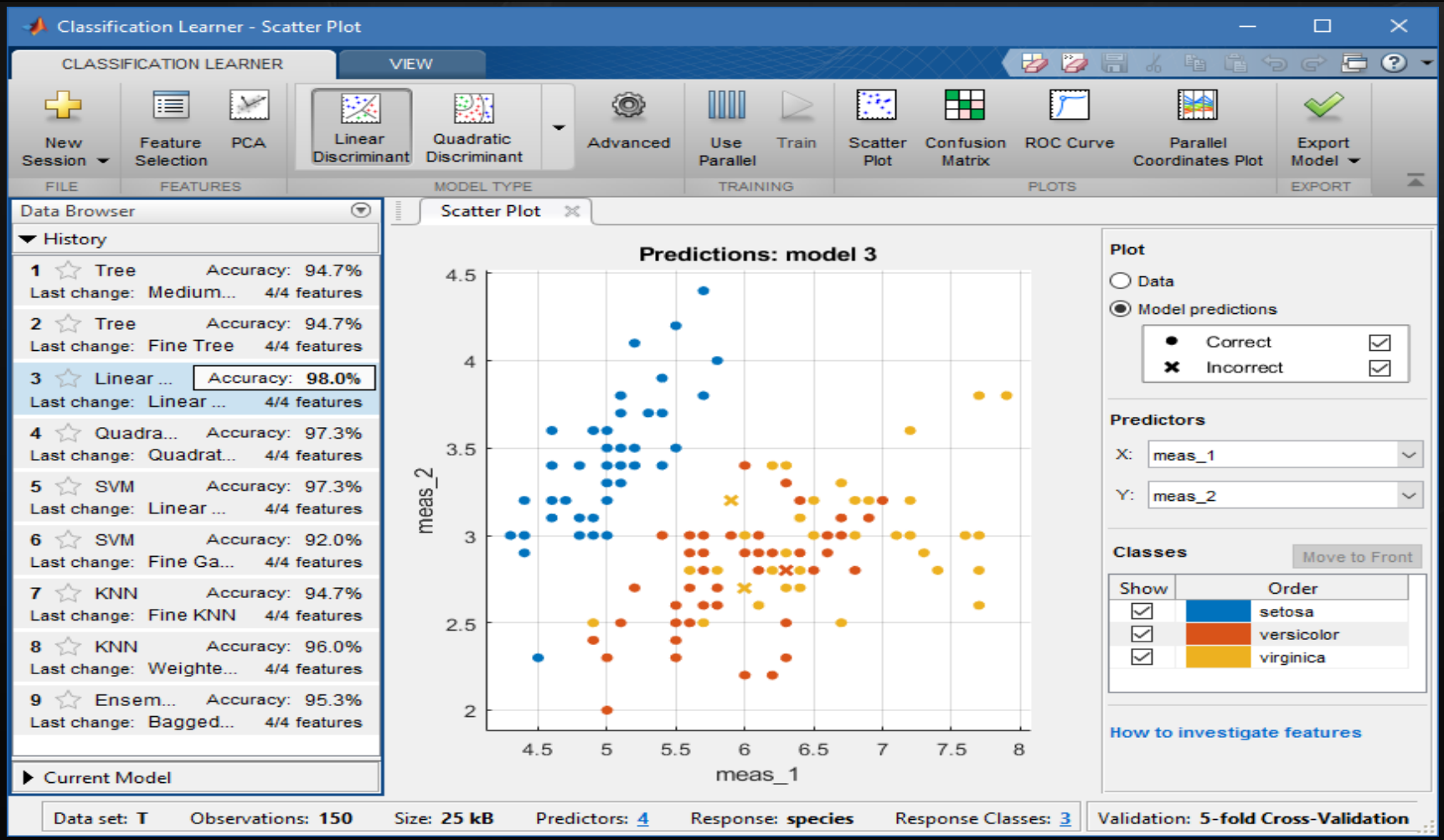
https://vas3k.com/blog/machine_learning/index.html

Příklad metod v Matlabu



https://humusoft.cz/blog/?view=20180817-strojove-uceni

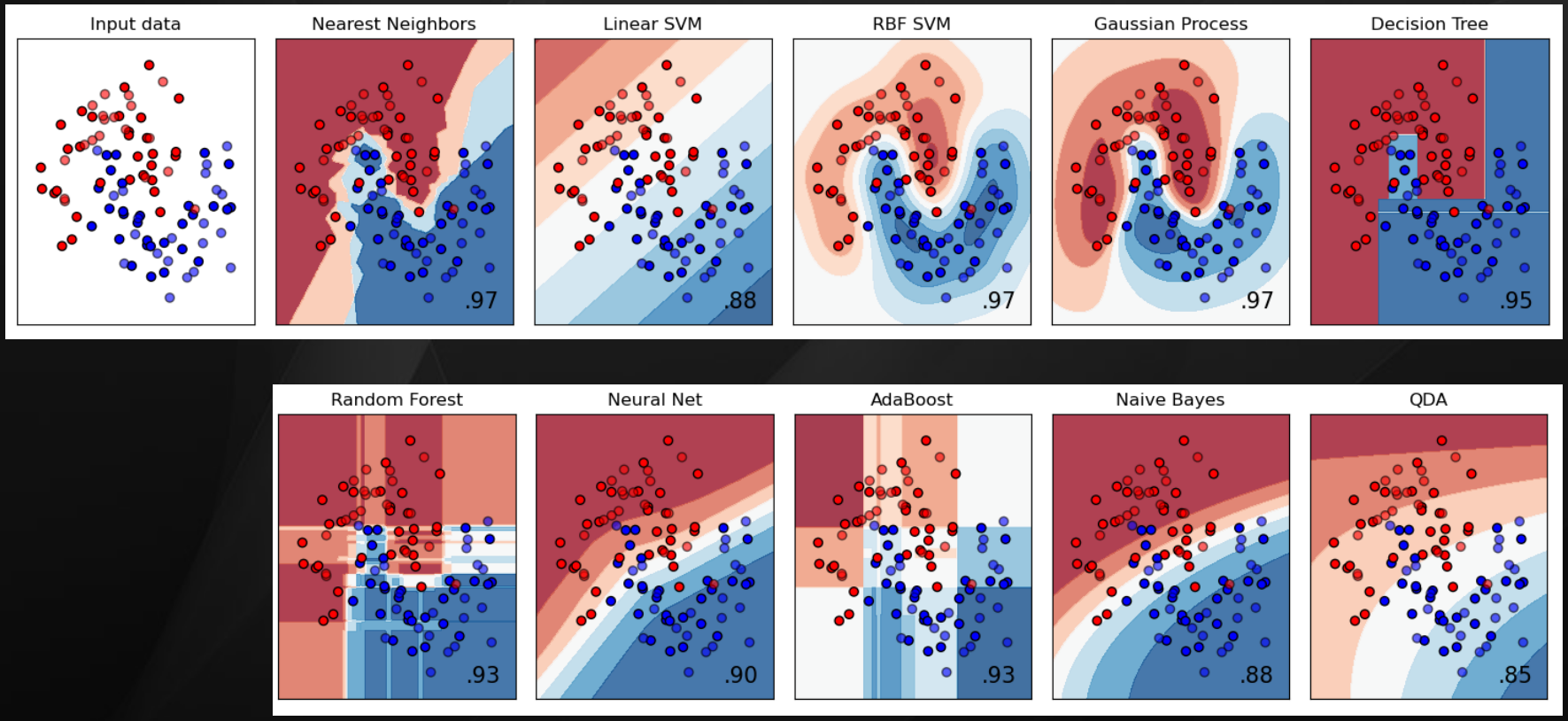
„Classification Learner“ v Matlabu



https://humusoft.cz/blog/?view=20180817-strojove-uceni

Klasifikátory v Pythonu

CIIRC

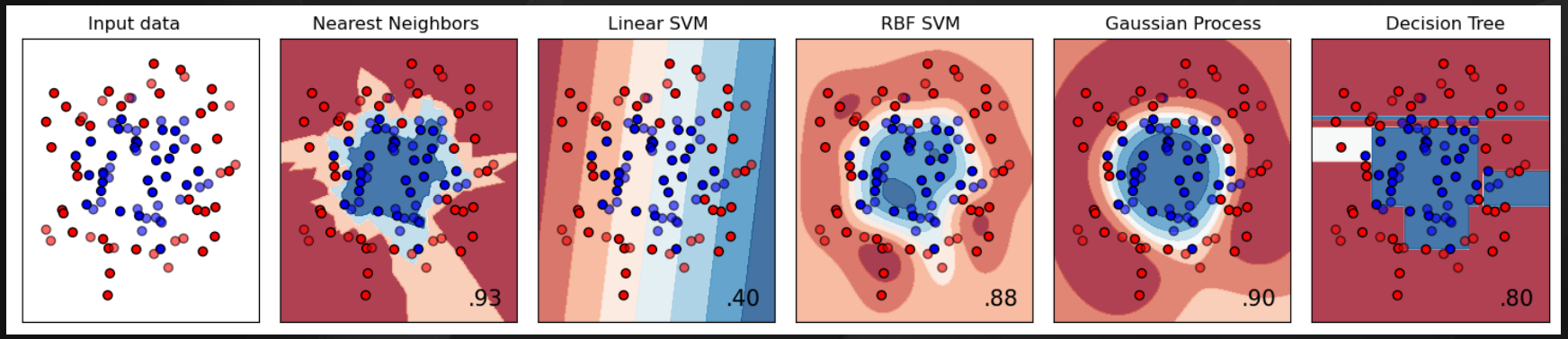


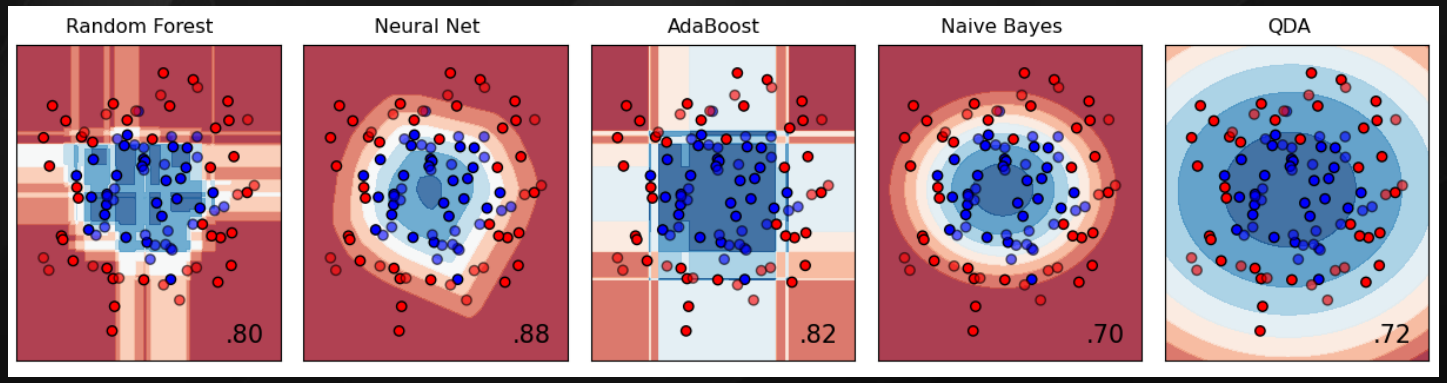
Rozhodovací hranice je různá pro různé klasifikátory

https://scikit-learn.org/

Klasifikátory v Pythonu

CIIRC





https://scikit-learn.org/

Základní typy klasifikátorů

CIIRC

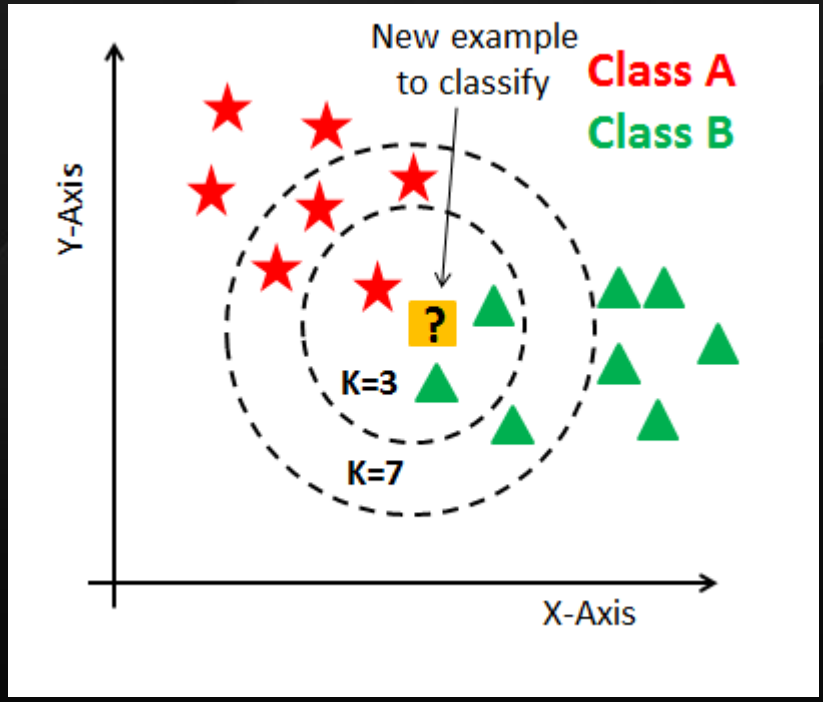
- Nejbližší soused (K-NN)
- Bayesovské klasifikátory
- Rozhodovací stromy
- SVM
- Neuronové sítě (ANN)

Nearest-neighbor klasifikátory

CIIRC

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

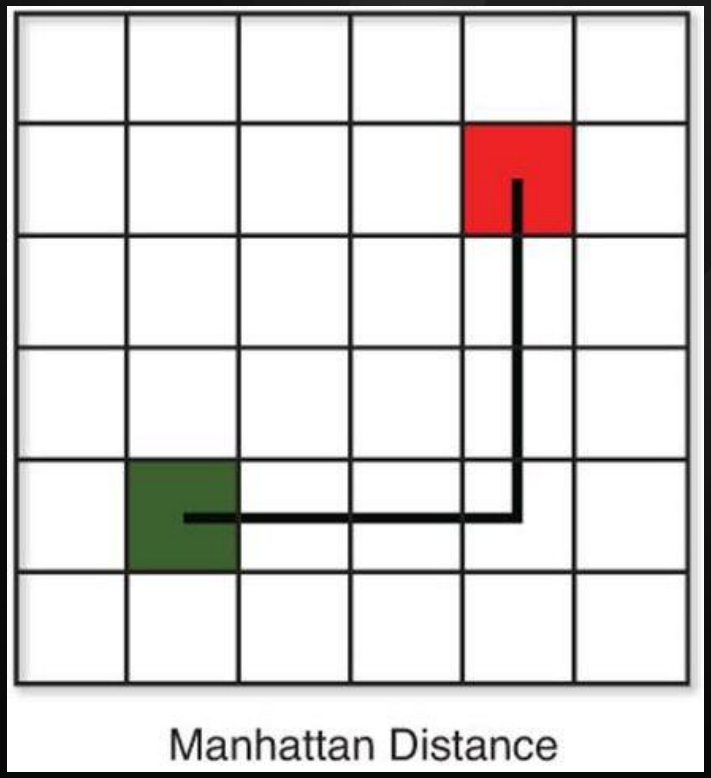
K-NN: Najde se k nejbližších sousedů a rozhodne se podle majoritní třídy



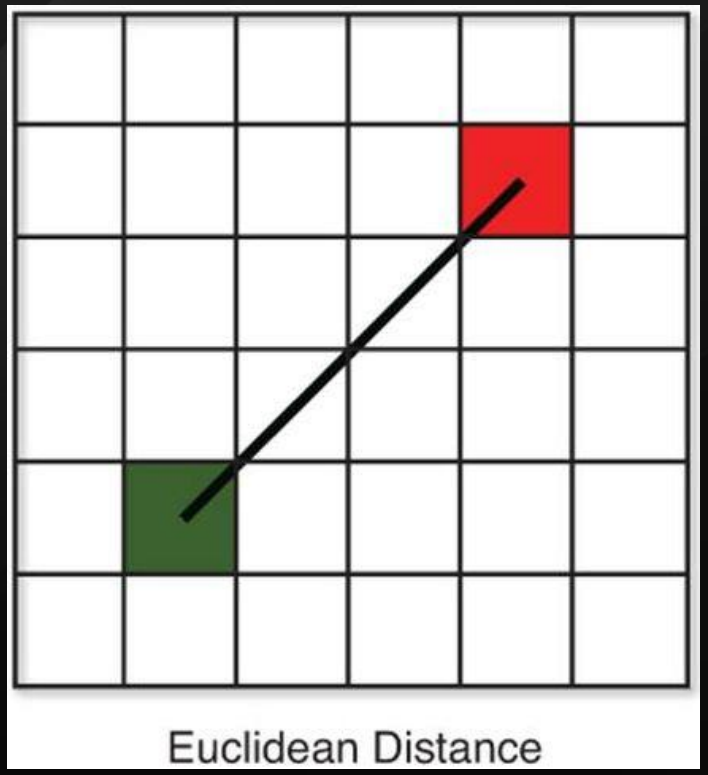
Nearest-neighbor klasifikátory

CIIRC

Jak měřit vzdálenost mezi body?



Manhattan Distance



Euclidean Distance

Nearest-neighbor klasifikátory

CIIRC

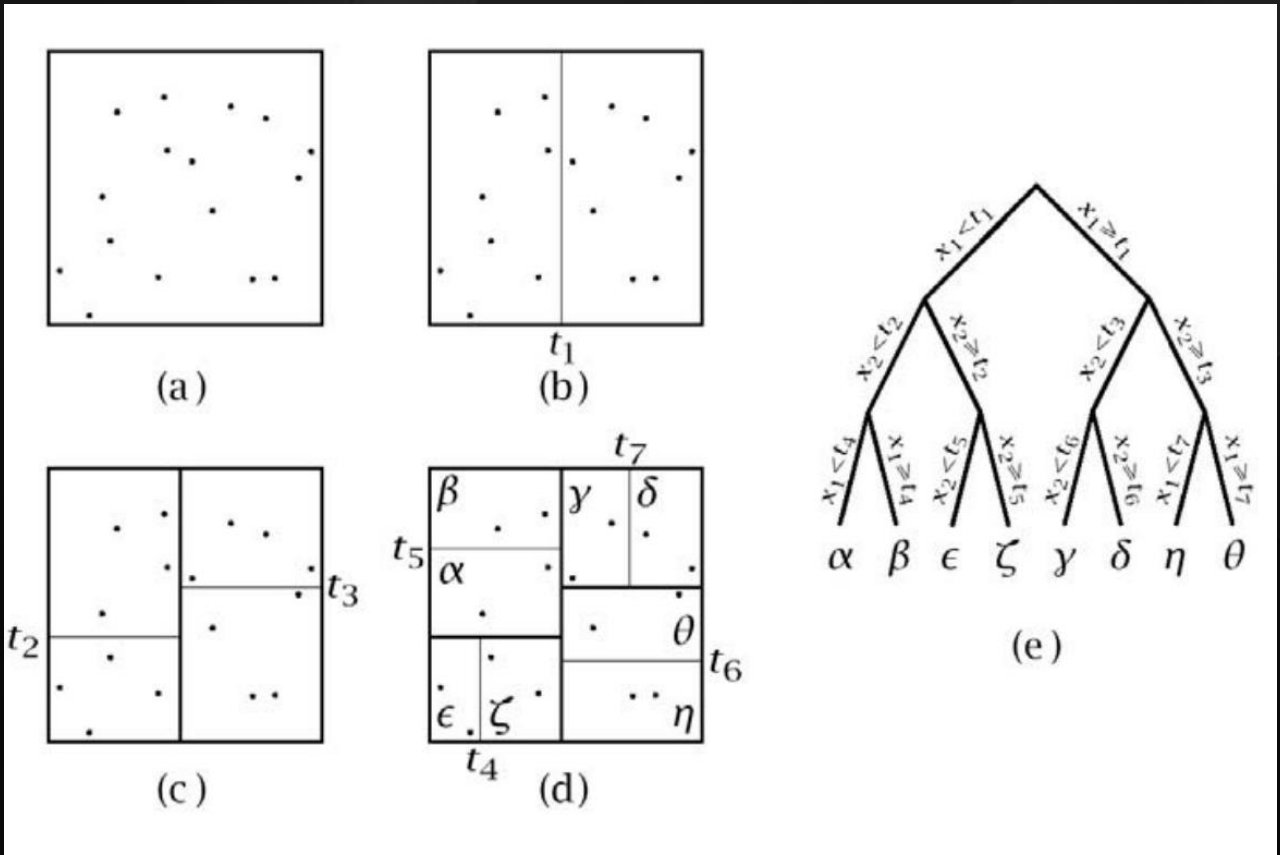
Výhody: Rychlé učení, jednoduchá implementace, snadné pochopení

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

K-D stromy

CIIRC

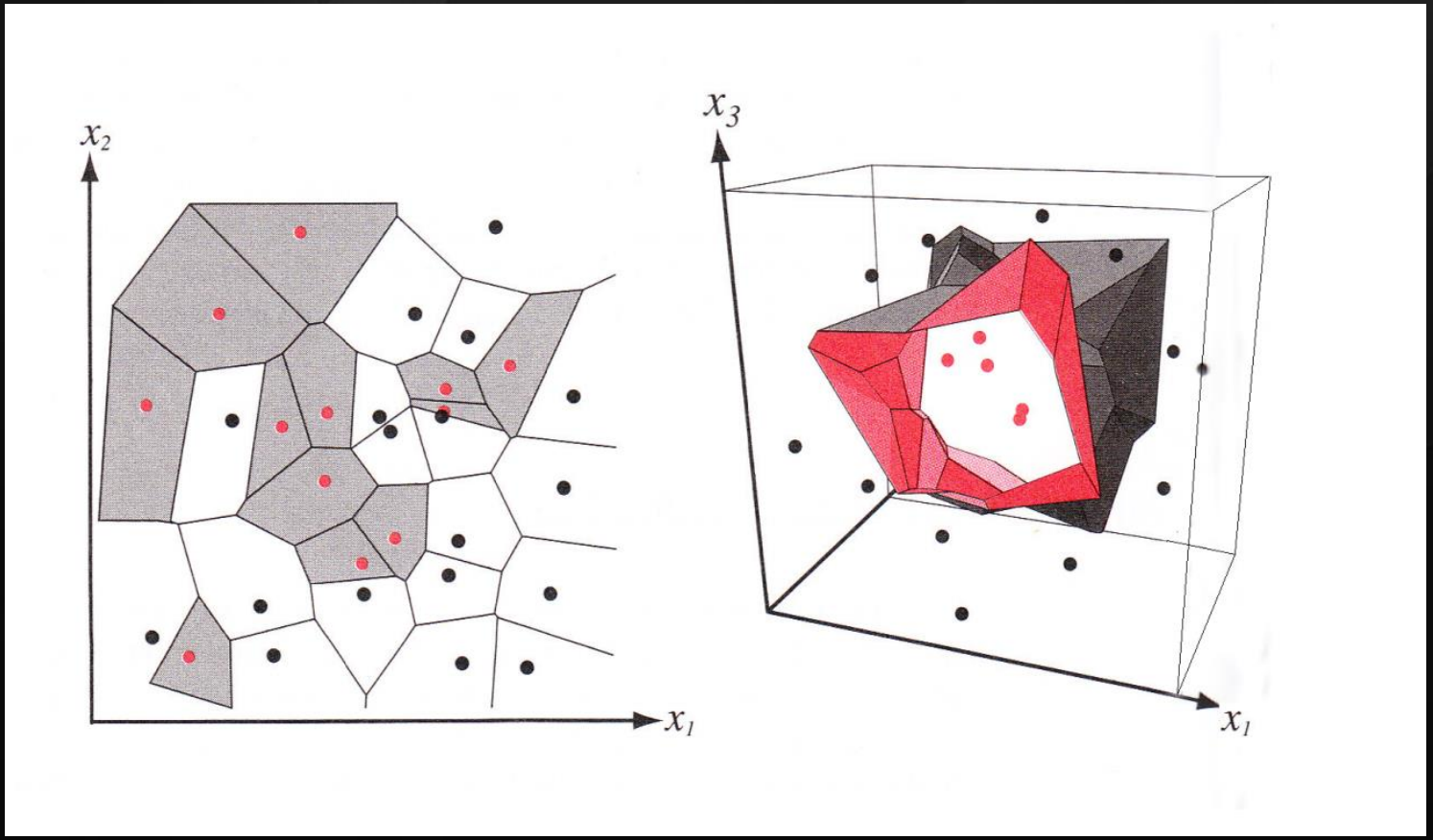
K-D stromy jsou modifikací algoritmu K-NN, která využívá rozdělení trénovací množiny na oblasti.



Voronoiovy diagramy

CIIRC

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



Bayesovské klasifikátory

CIIRC

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

Bayesovské klasifikátory

CIIRC

Bayesovská rozhodovací teorie

If $P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$, přiřaď $\boldsymbol{x}$ do ω_1,
else přiřaď $\boldsymbol{x}$ do ω_2.

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

Minimalizujeme střední pravděpodobnost chyby.

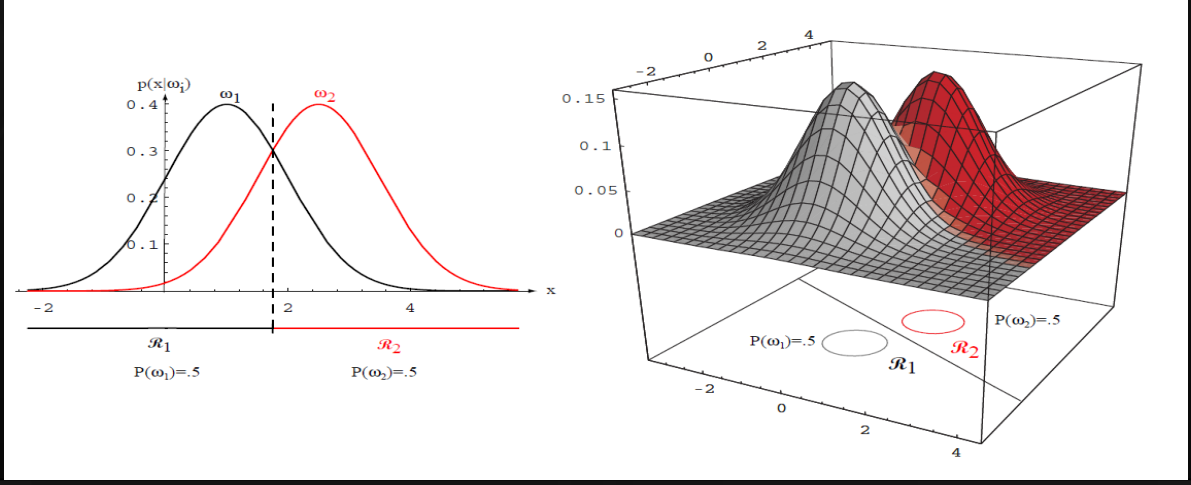
Bayesovské klasifikátory

CIIRC

Pro výpočet aposteriorních pravděpodobností lze využít Bayesův vztah

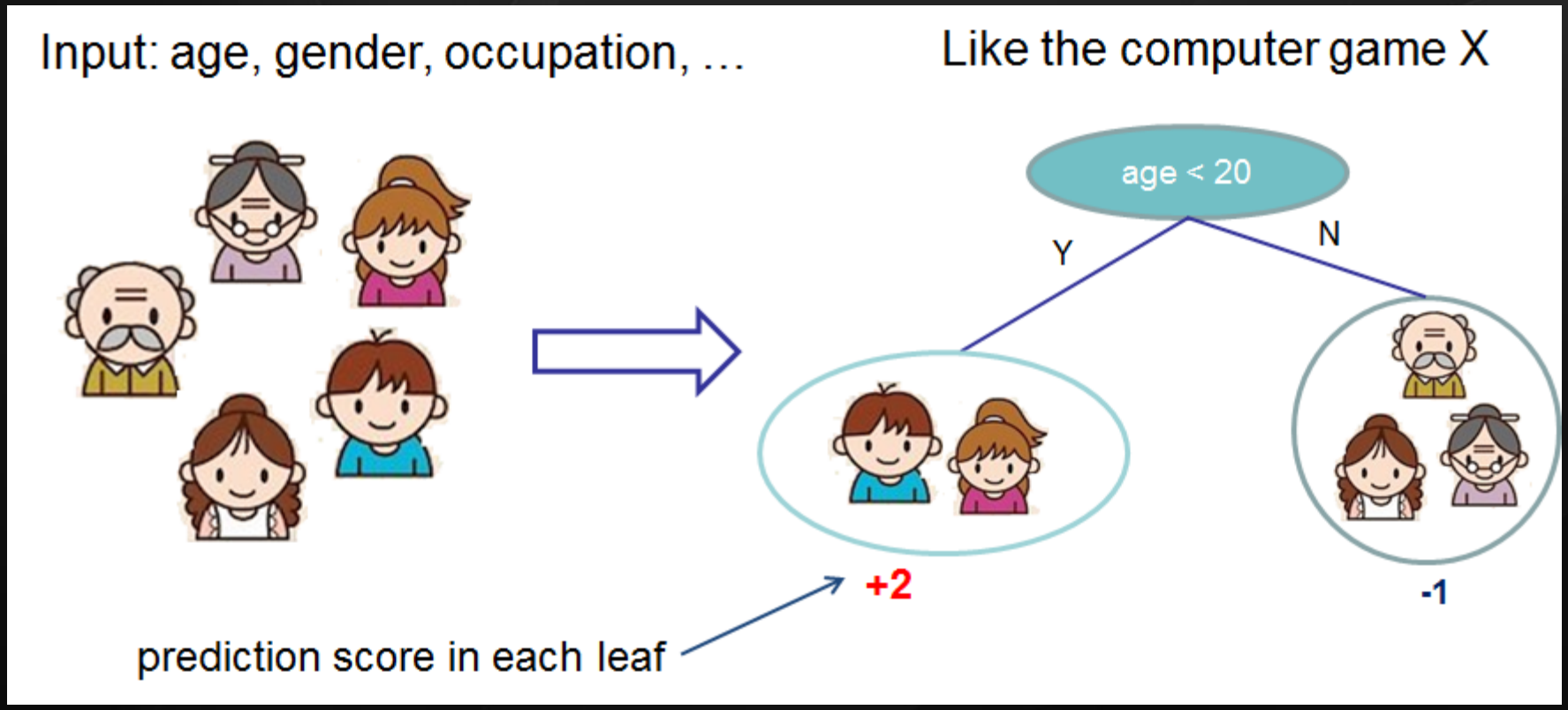
$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})}$$

Učení je odhad parametrů pravděpodobnostních rozložení $p(\boldsymbol{x}|\omega_2)$



Rozhodovací stromy (Decision tree)

CIIRC



https://xgboost.readthedocs.io/en/stable/tutorials/model.html

Rozhodovací stromy

CIIRC

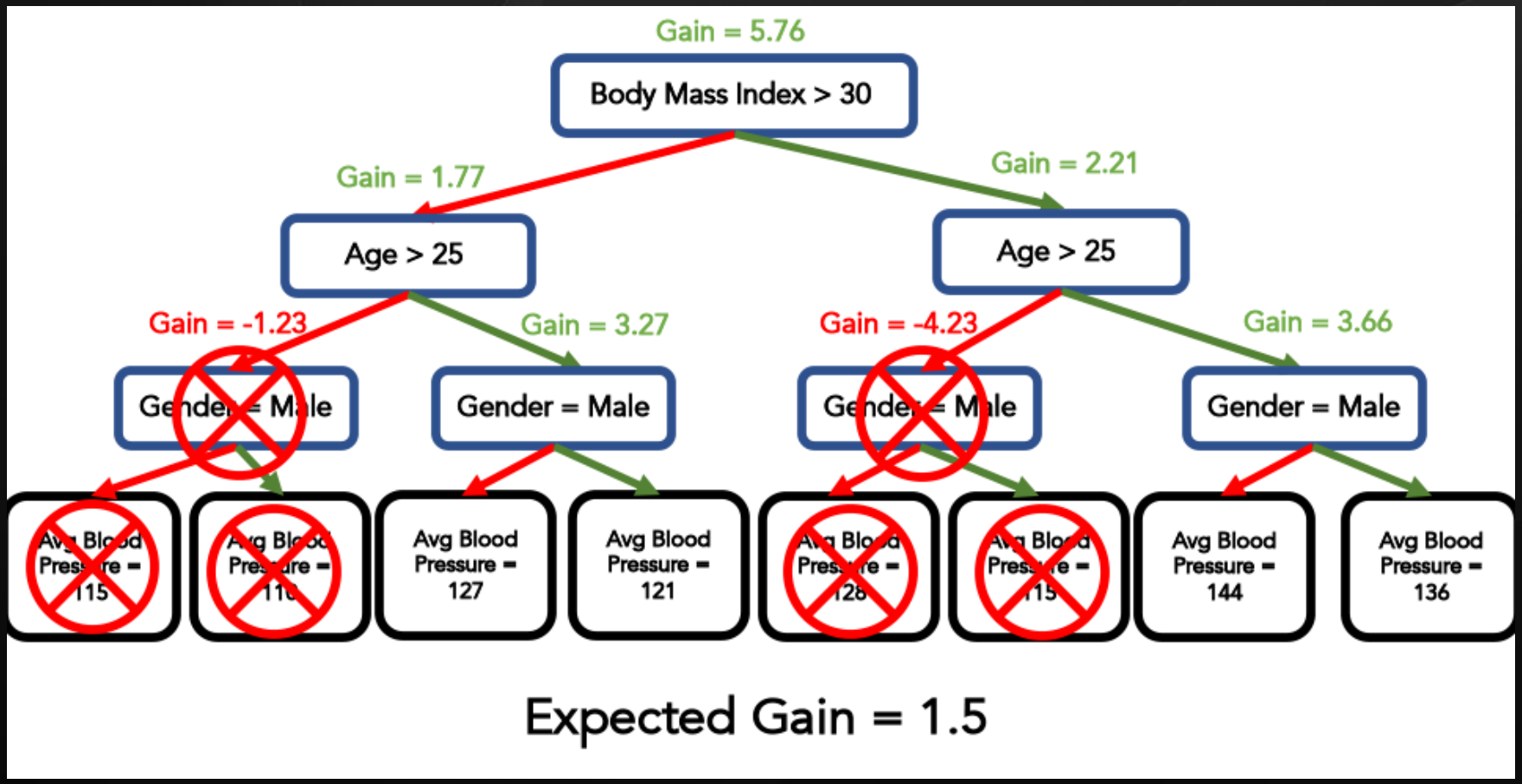
Řešení klasifikačních úloh pomocí rozhodovacích stromů vyžaduje dva kroky:

1. Indukce rozhodovacího stromu (podle trénovacích dat)

2. Použití rozhodovacího stromu a určení třídy

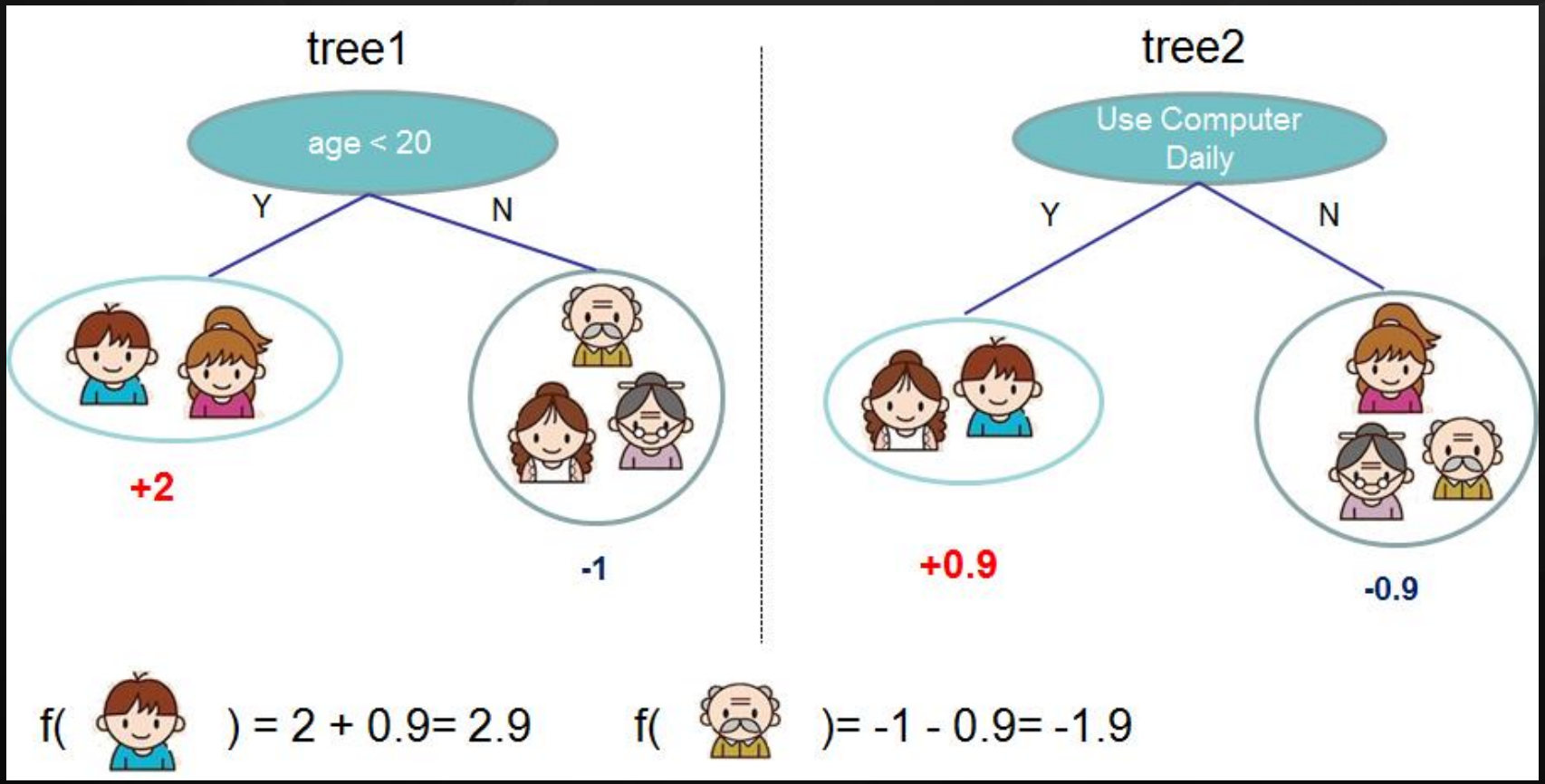
Prořezávání rozhodovacích stromů

CIIRC



Ensemble learning

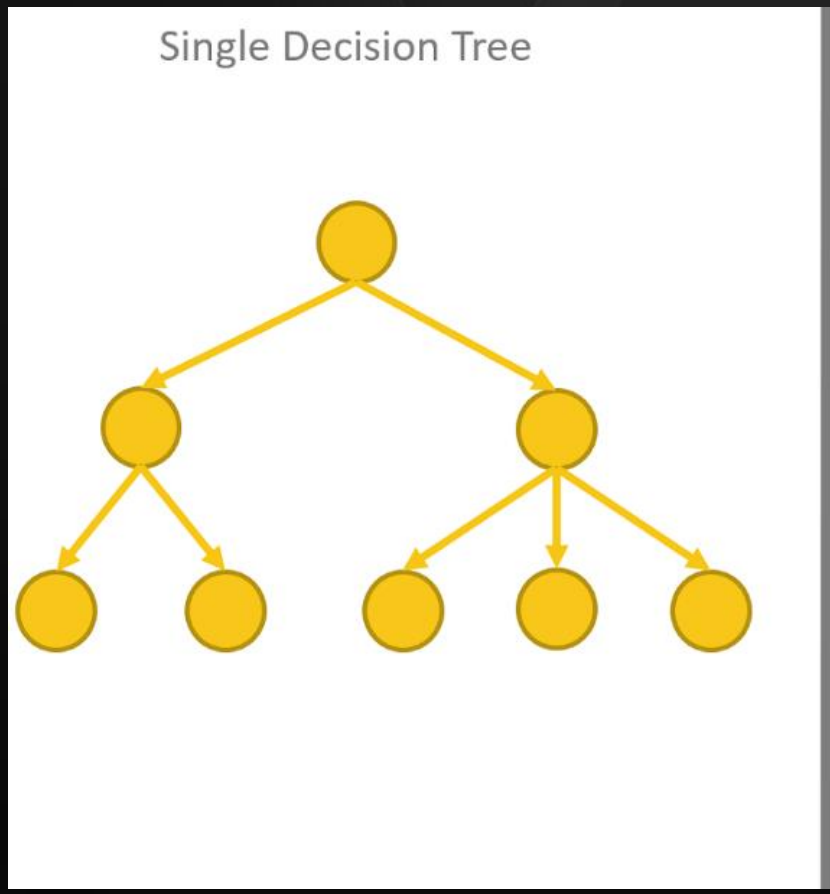
CIIRC

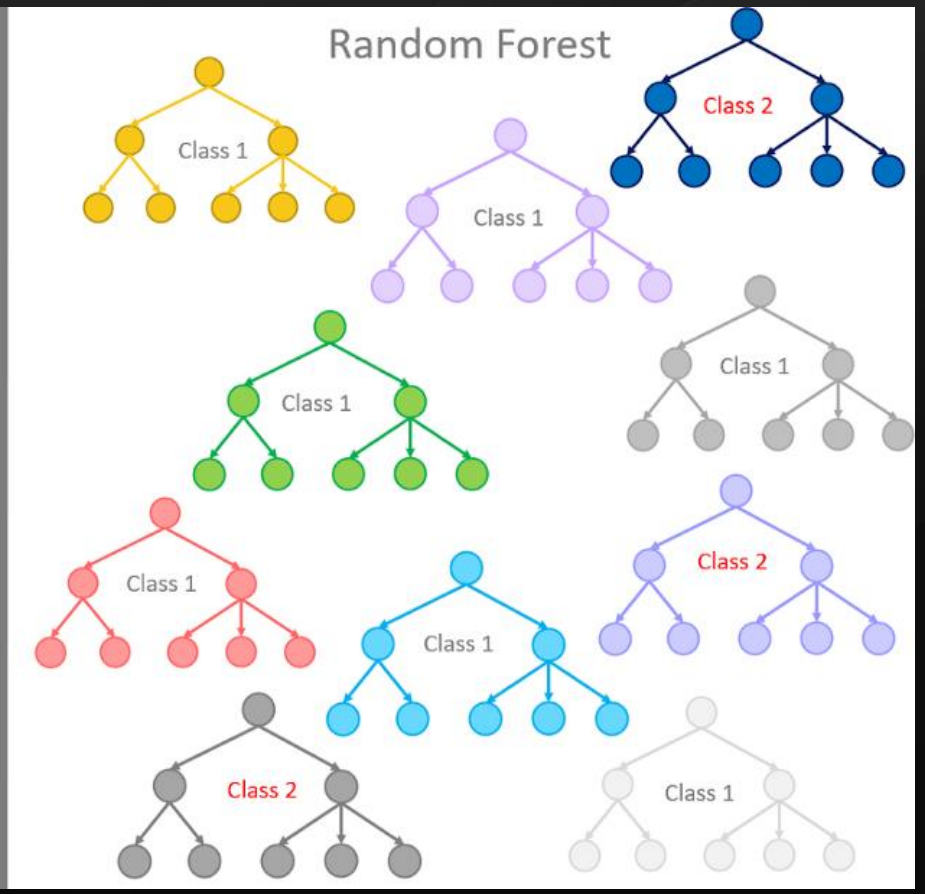


V praxi používáme více stromů a jejich predikci poté kombinujeme.

Random decision forest

CIIRC





Rozhodovací stromy

CIIRC

Výhody:

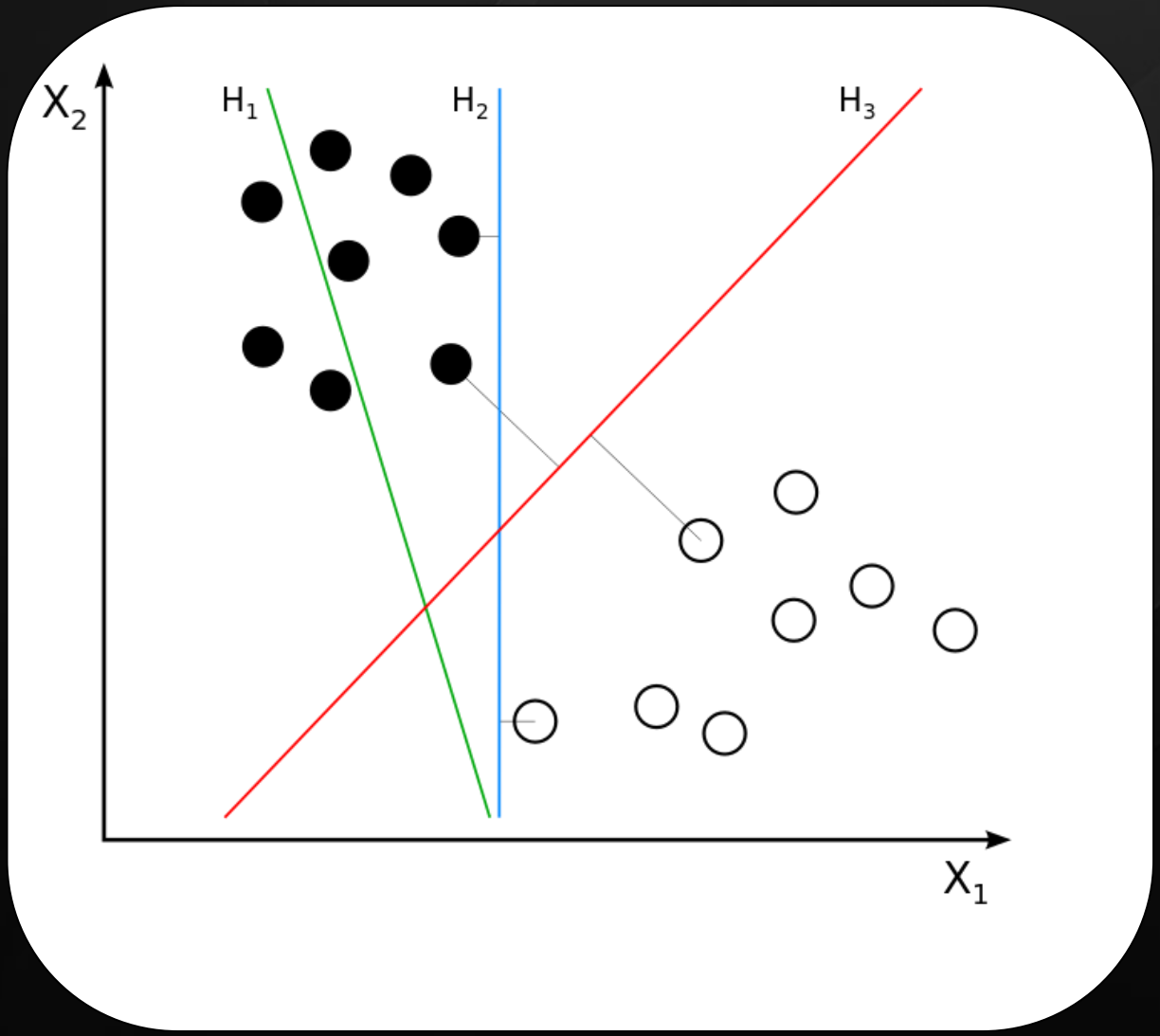
- Jednoduché
- Efektivní
- Extrakce jednoduchých pravidel
- Použitelné i pro velké datasety

Nevýhody:

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

Support-vector machine, SVM

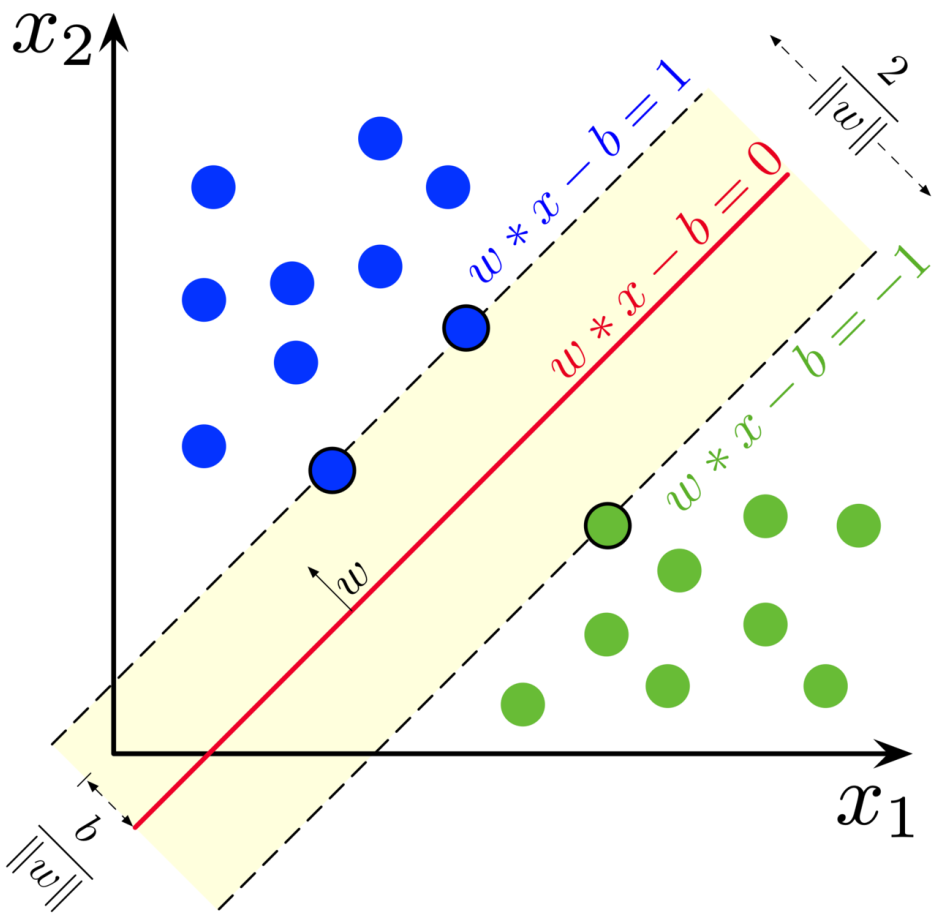
CIIRC



H2? H3?

Support-vector machine, SVM

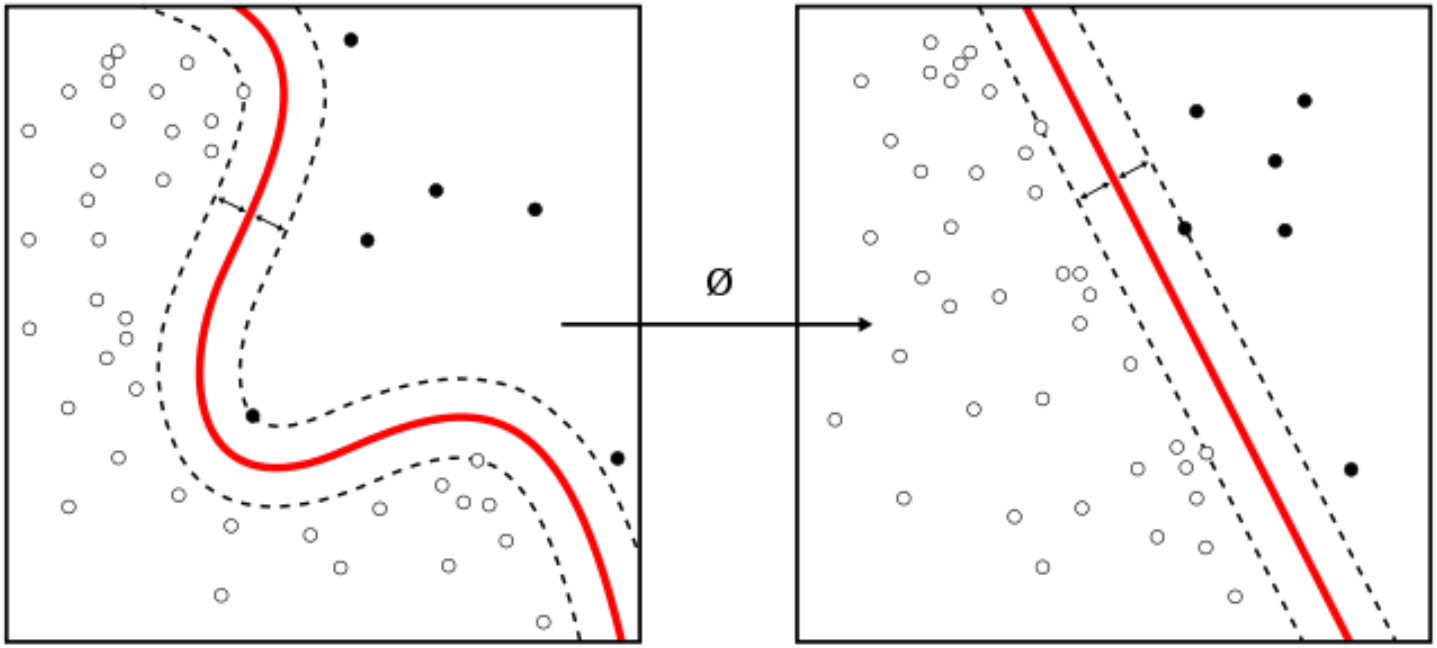
CIIRC



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

SVM, Kernel machine

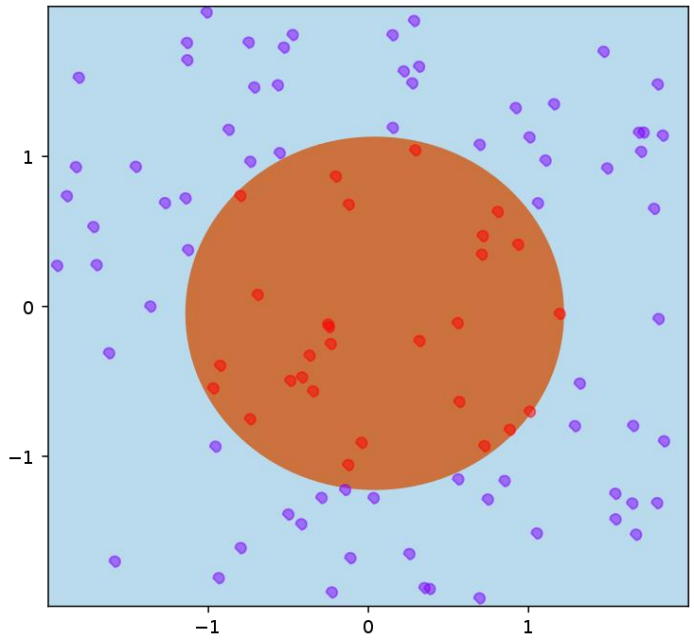
CIIRC

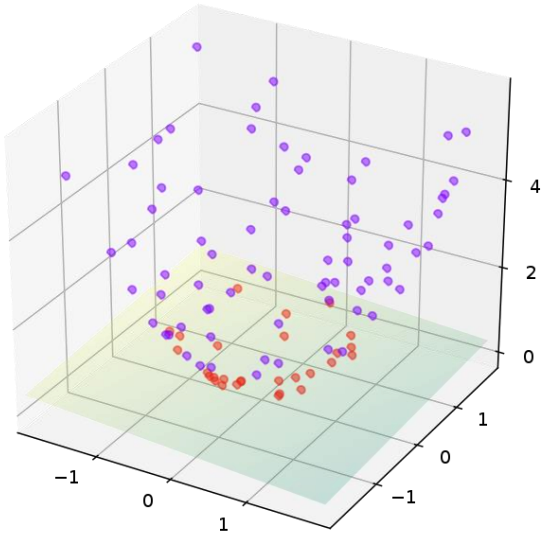


SVM, Kernel machine

CIIRC

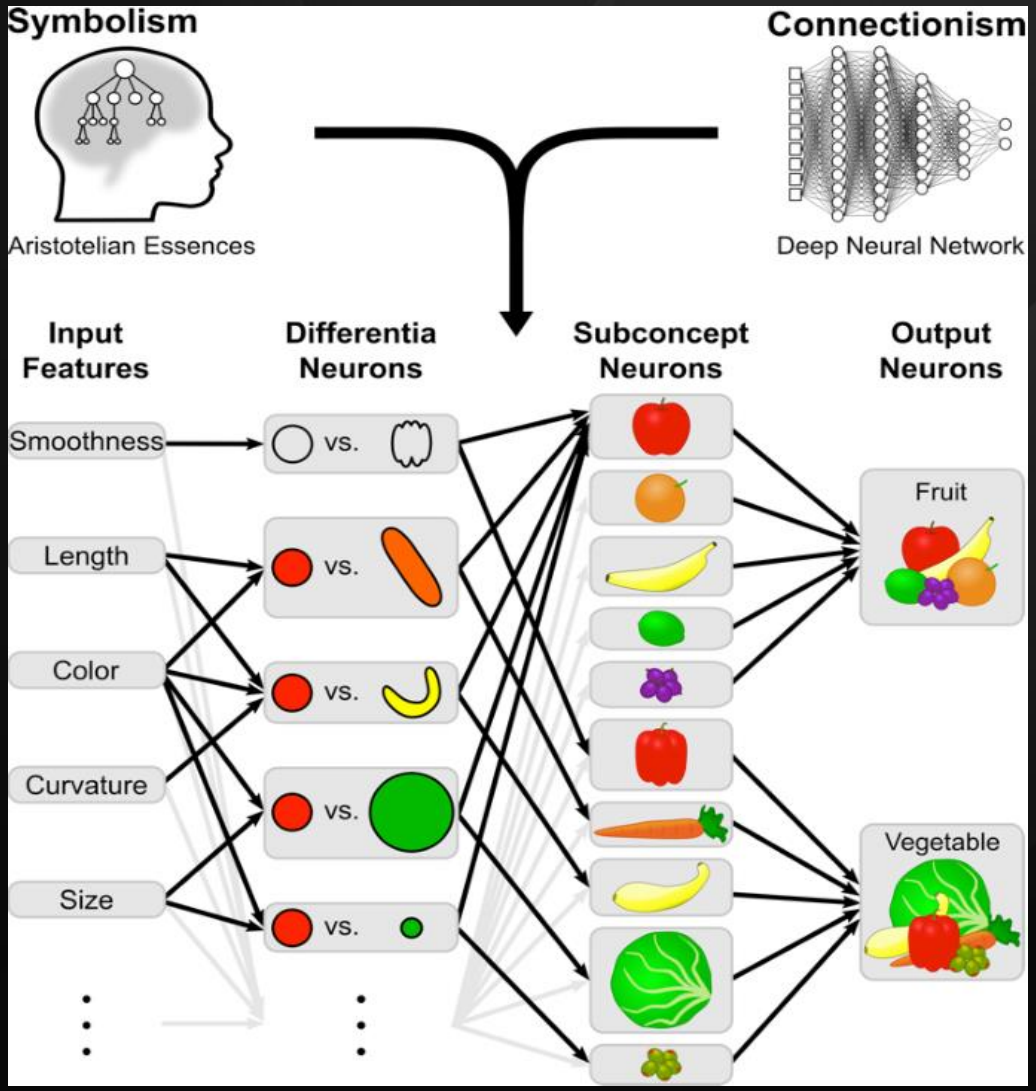
A training example of SVM with kernel given by $\varphi((a, b)) = (a, b, a^2 + b^2)$



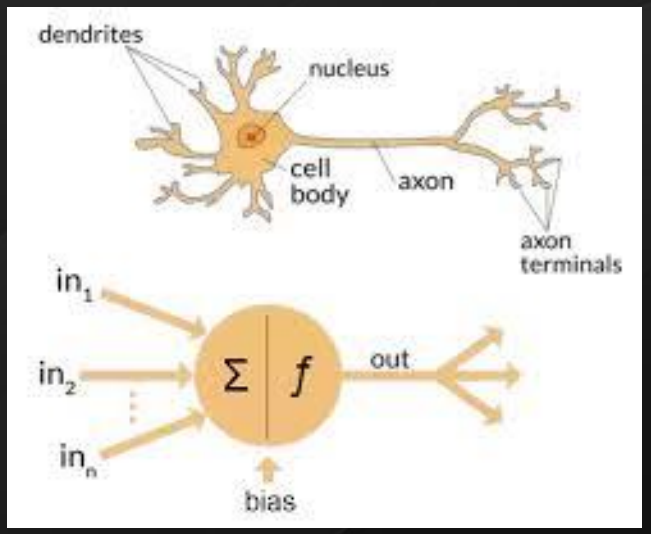


Neuronové sítě

CIIRC



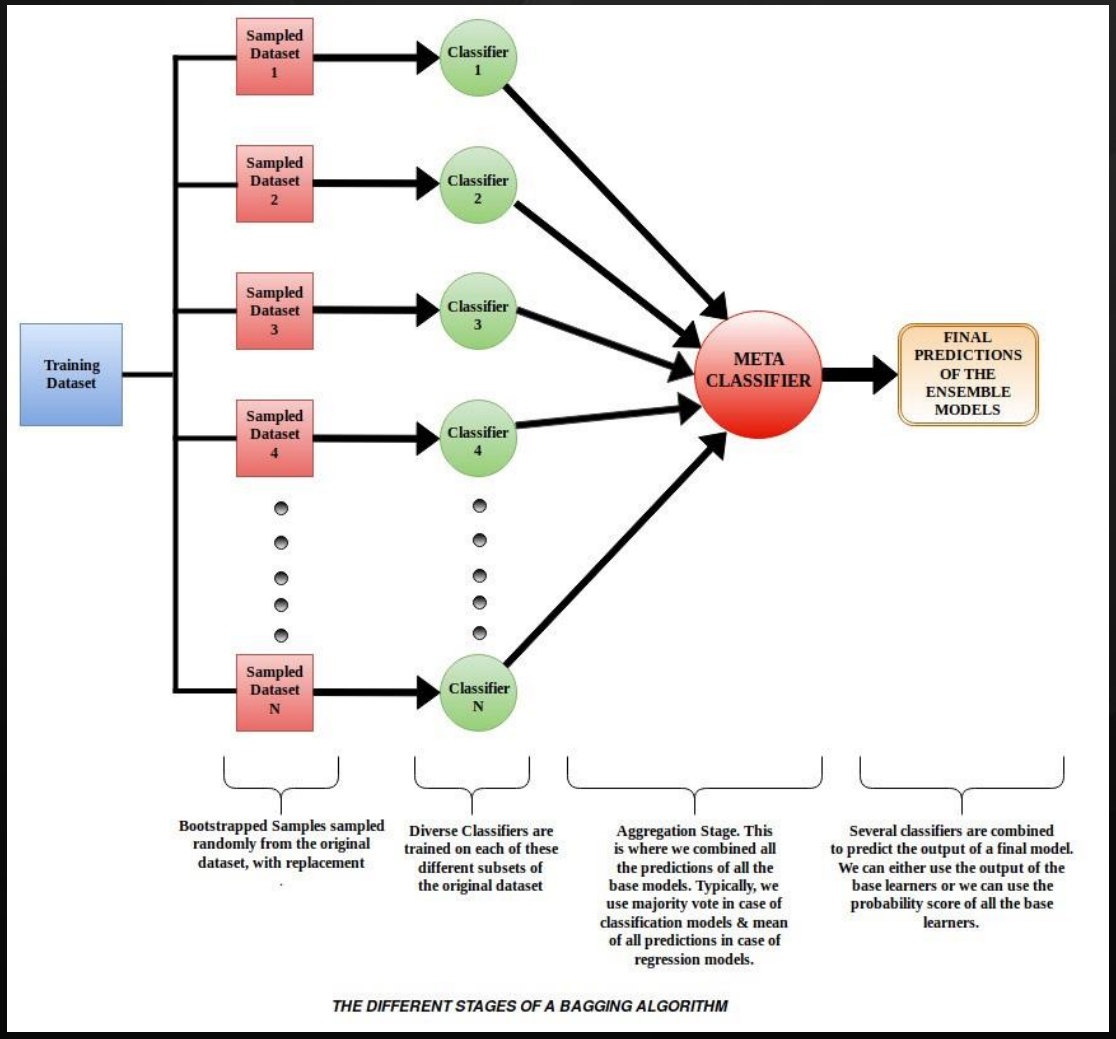
Paul J. Blazek, Milo M. Lin: A neural network model of perception and reasoning



Richard Nagyfi: The differences between Artificial and Biological Neural Networks

Meta klasifikátory

CIIRC



- Bagging
- Boosting
- Stacking

Předzpracování dat - odstranění outlierů

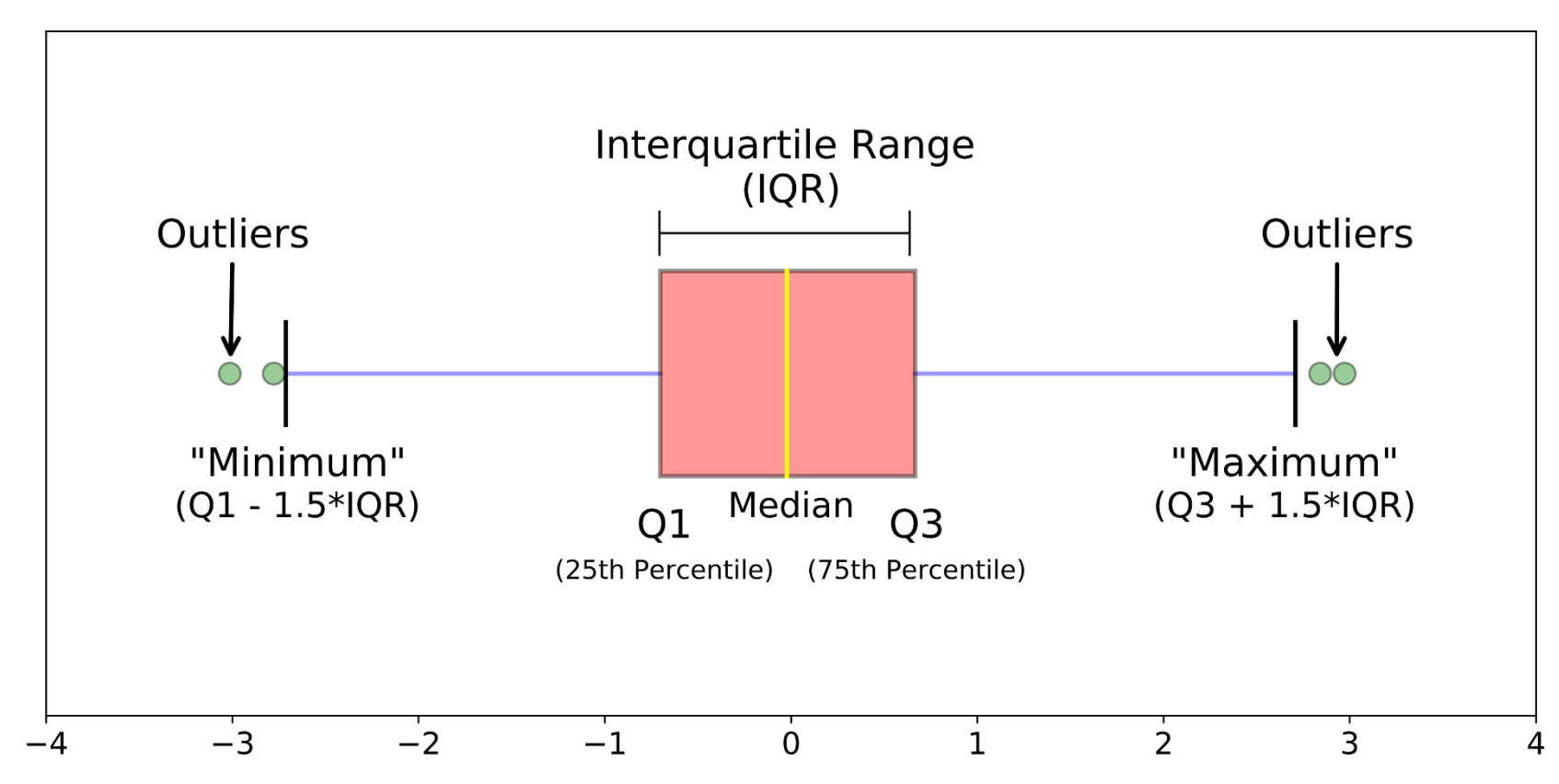
CIIRC

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



Normalizace příznaků

CIIRC

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

Normalizace, základní dvě metody:

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

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

Některé typy klasifikátorů vyžadují normalizaci. Např. KNN klasifikátor.
U jiných to není potřebné, např. rozhodovací stromy.

Selekce příznaků

CIIRC

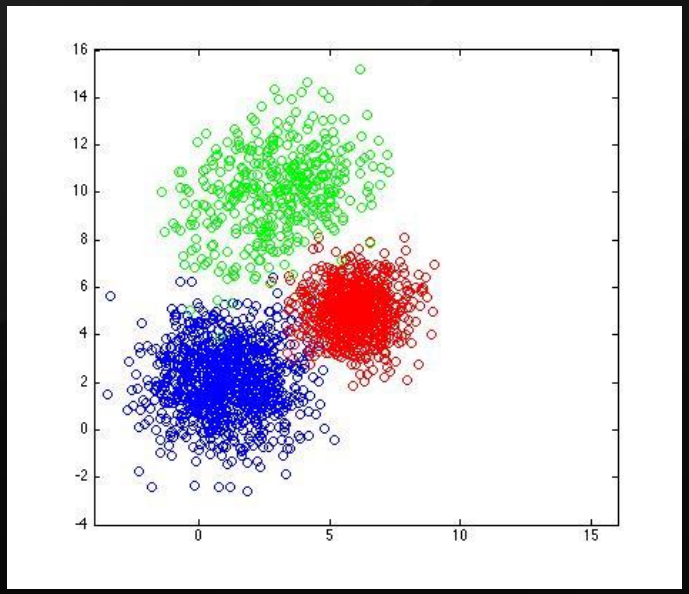
Hledání podsouboru příznaků, který maximalizuje nějaké kritérium.

Kritérium

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

Prohledávání

- Individual ranking
- Sequential Forward Selection
- Sequential Backward Elimination
- Evolutionary algorithms

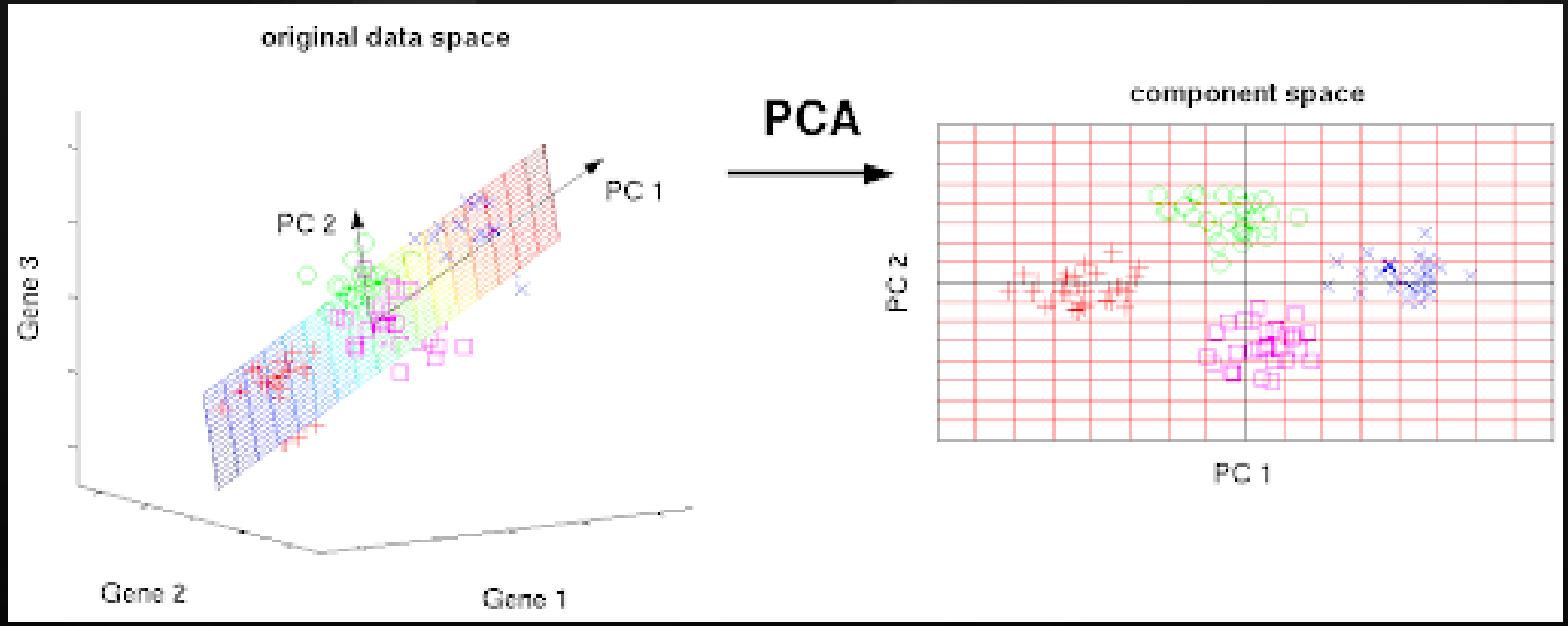


Redukce příznakového prostoru

CIIRC

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

Ukázka snížení dimenze z 3D do 2D

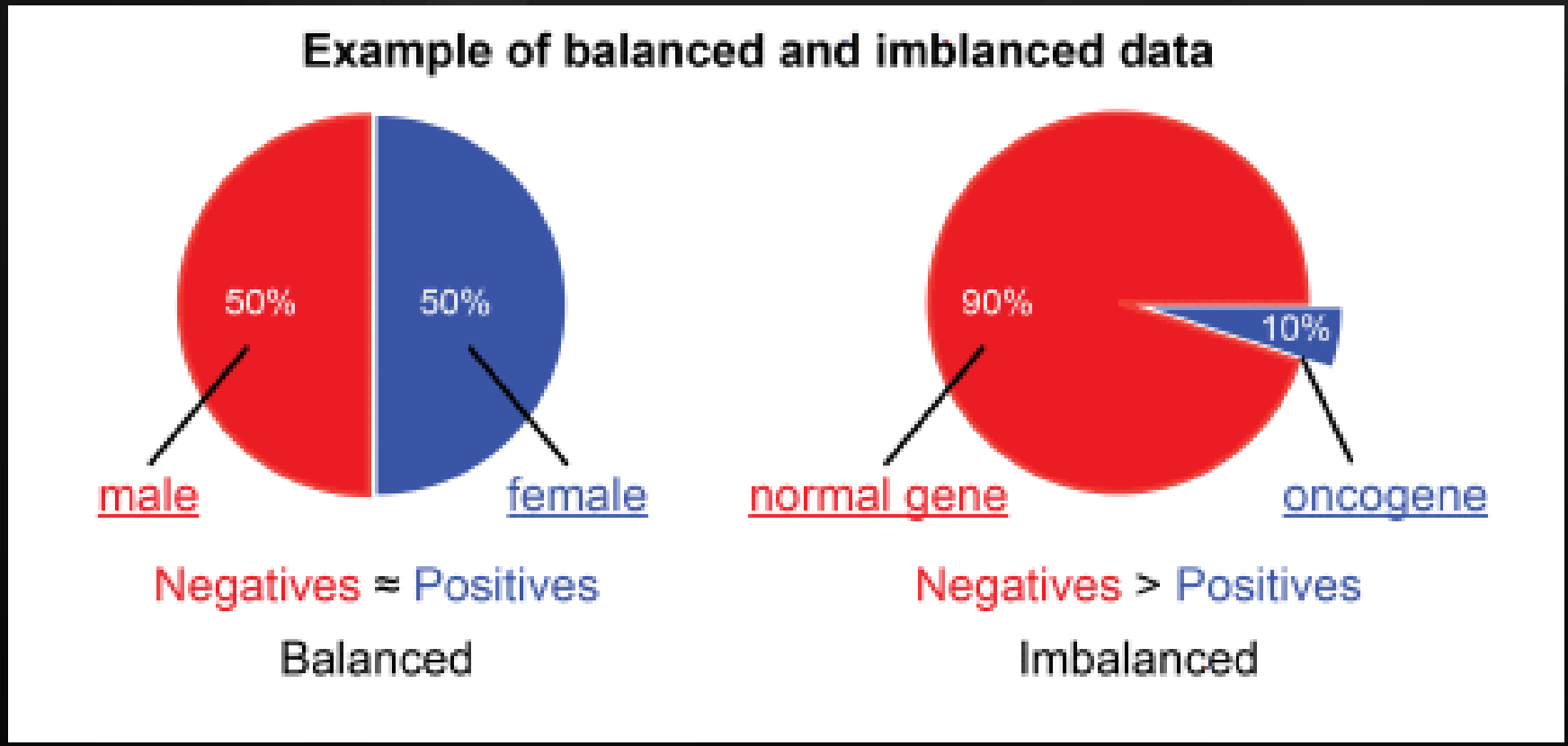


Nerovnoměrné zastoupení tříd

CIIRC

Nerovnoměrné zastoupení tříd

CIIRC

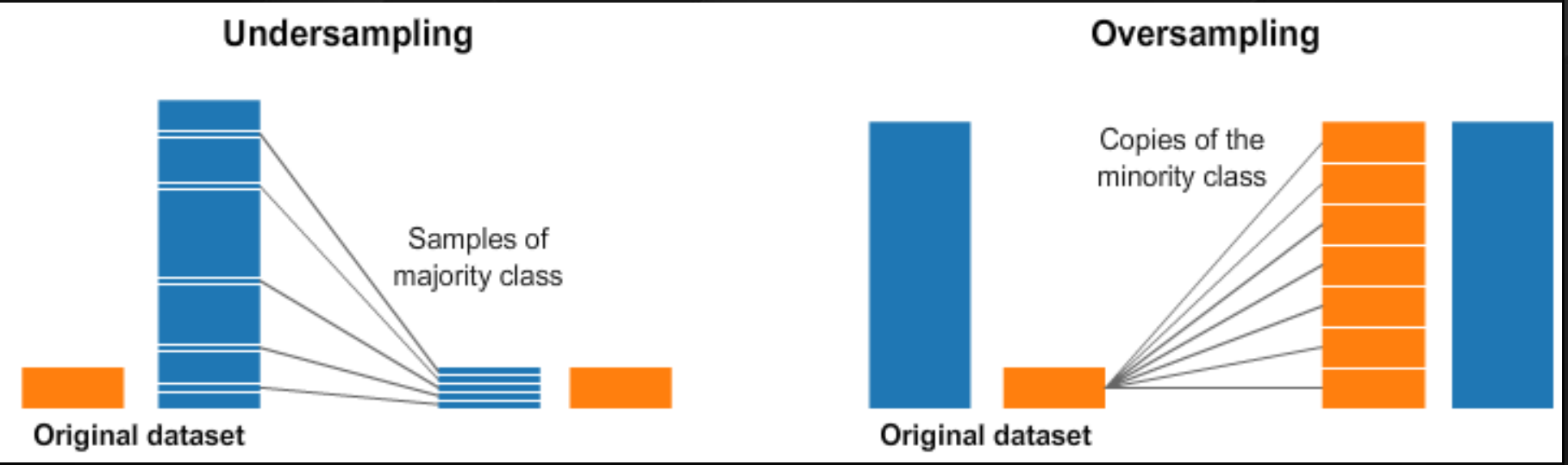


Nerovnoměrné zastoupení tříd

CIIRC

Pokud jsou třídy nevyvážené, může docházat k chybému trénování klasifikátoru.

Řešení – např. Undersampling, nebo Oversampling.



Hodnocení klasifikátoru

CIIRC

Chceme predikovat, zda pacient trpí např. Wilsonovou chorobou.

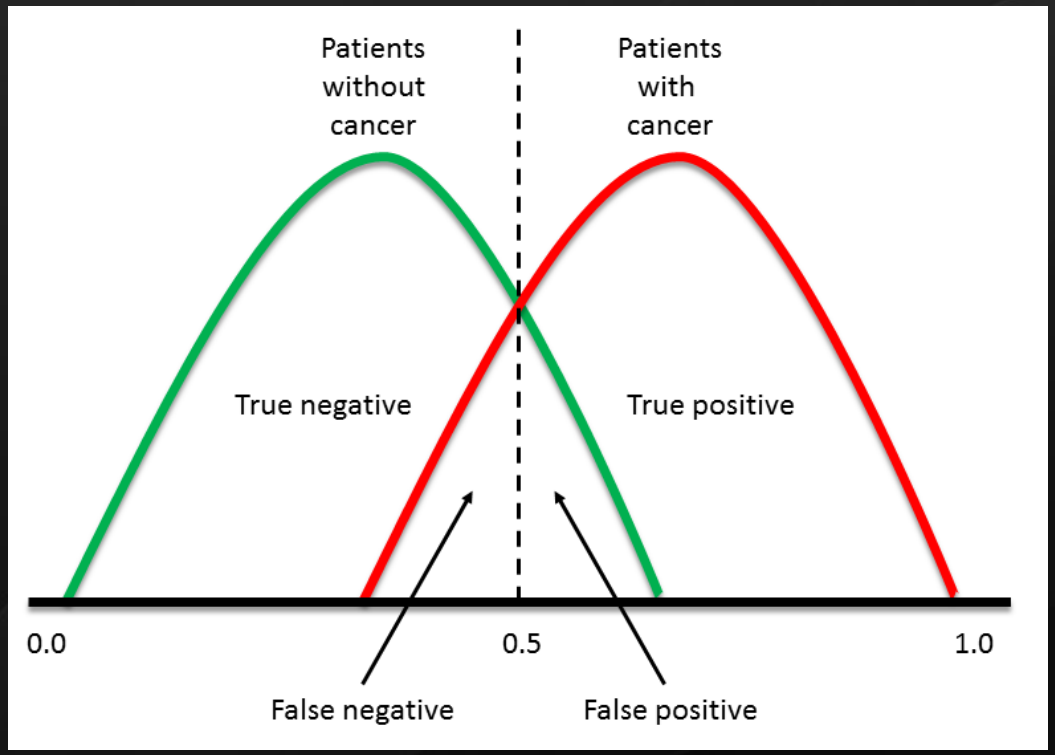
Prevalence choroby je 1:25.000
Budu-li stále predikovat negativní výsledek, klasifikuju 99.996% případů správně!!!!!!!!!!

ACCURACY (PŘESNOST) KLASIFIKÁTORU NEMUSÍ BÝT VHODNÁ MÍRA

Hodnocení klasifikátoru

CIIRC

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



ACCURACY = (TP+TN)/(TP+FP+FN+TN)
SENSITIVITY = TP/P=TP/(TP+FN)
SPECIFICITY = TN/N=TN/(TN+FP)

Kontingenční tabulka

CIIRC

		Predicted 0	Predicted 1
Actual	0	TN	FP
	1	FN	TP

$$Precision = \frac{TP}{TP + FP}$$

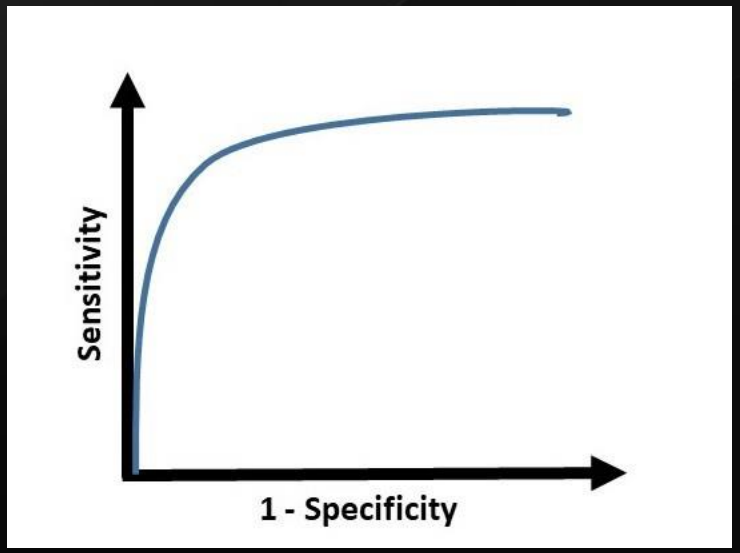
$$Recall = \frac{TP}{TP + FN}$$

ROC křivka

CIIRC

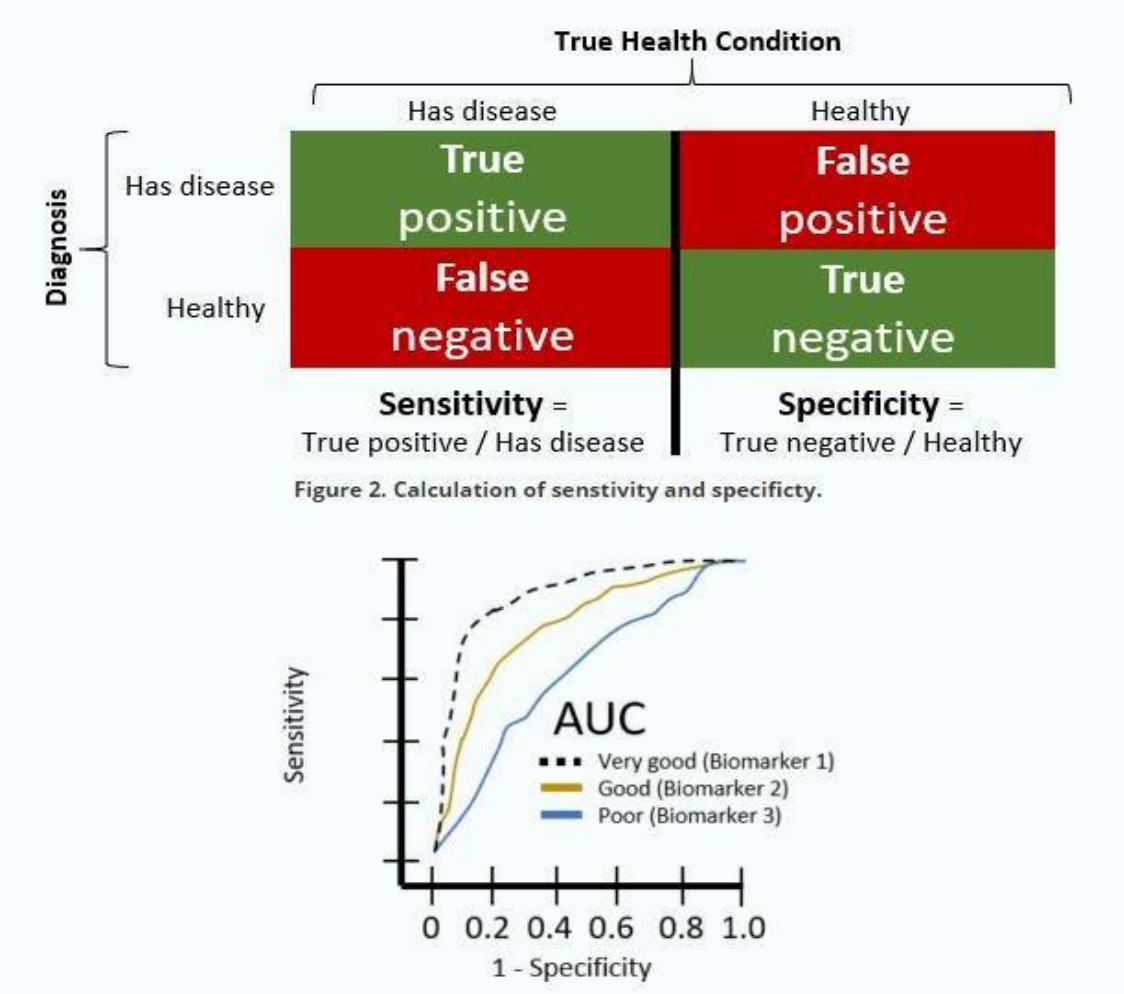
ROC (Receiver Operating Characteristic) křivka je nástroj pro hodnocení a optimalizaci binárního klasifikačního systému (testu), který ukazuje vztah mezi specificitou a senzitivitou pro všechny přípustné hodnoty prahu.

Z ROC je možné spočítat AUC (Area Under Curve)



Plocha pod křivkou - Area under the curve (AUC)

CIIRC



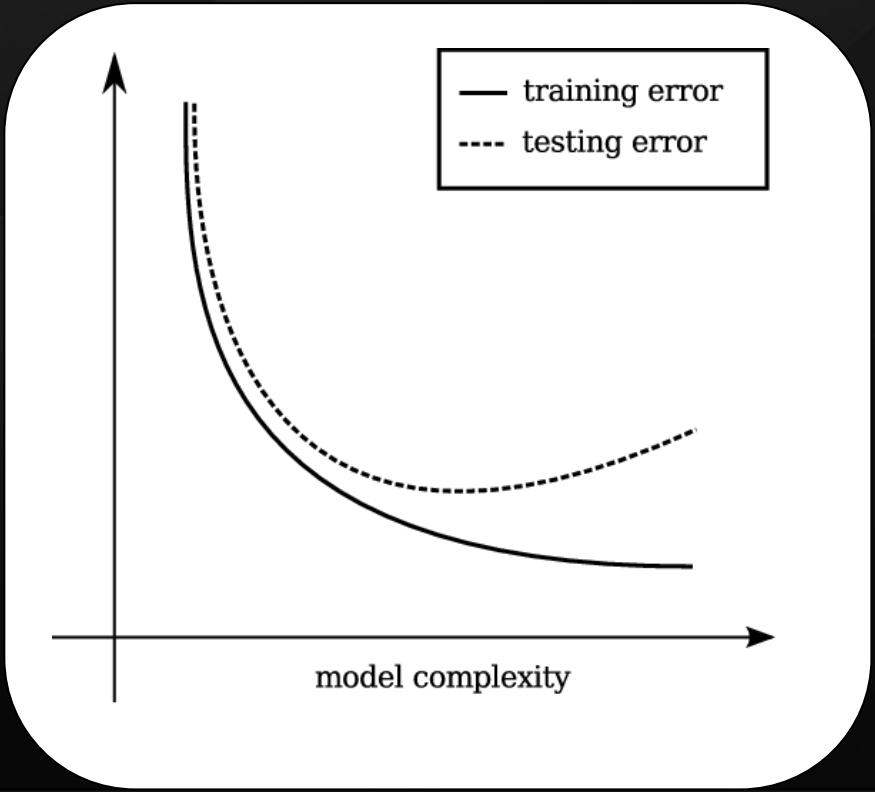
https://medium.com/computer-architecture-club/what-is-the-auc-roc-curve-47fbdcbf7a4a

Generalizace klasifikátoru

CIIRC

Zjištění generalizační schopnosti klasifikátoru

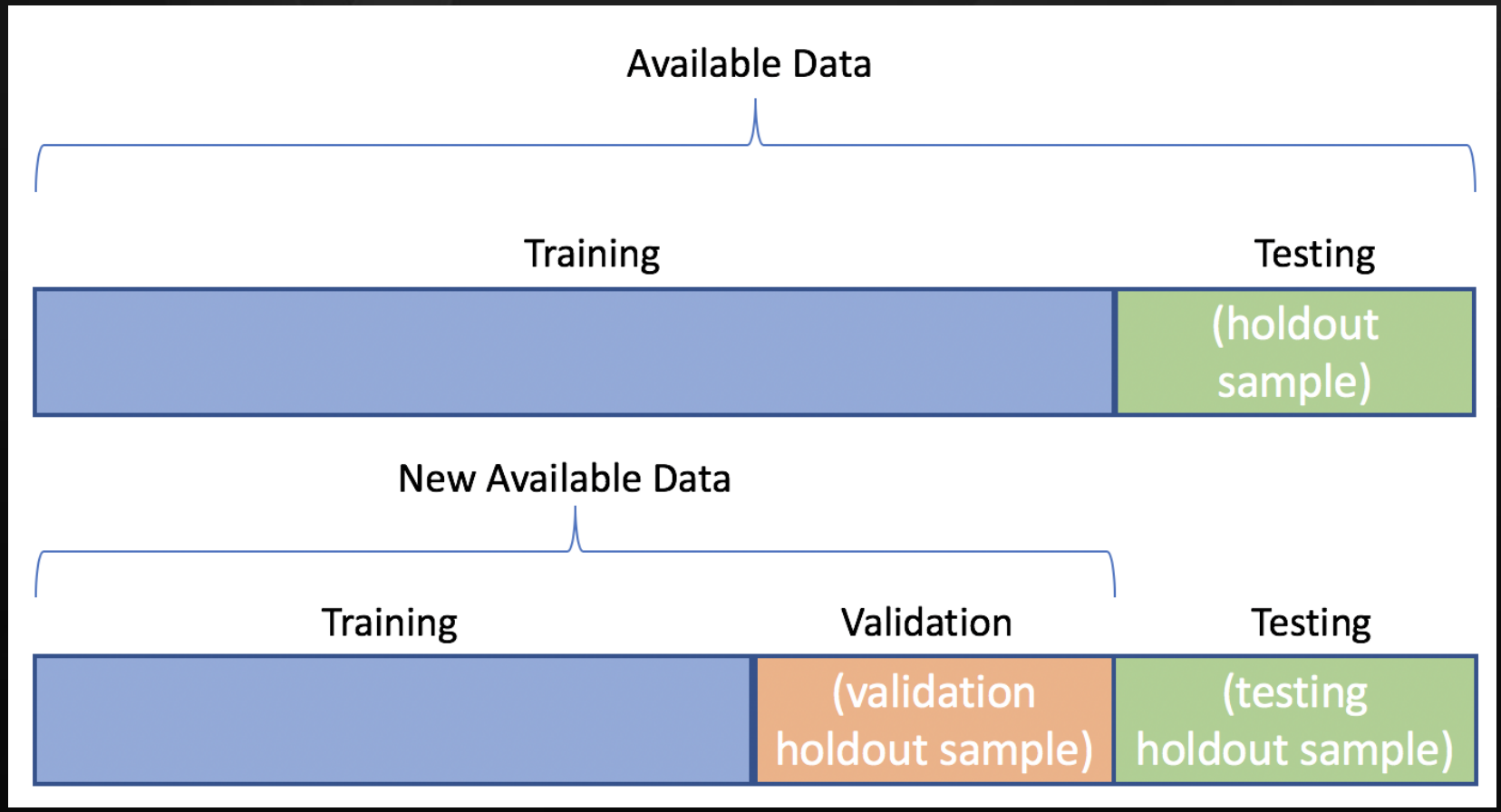
Tj. zda klasifikátor funguje i na datech, která nikdy neviděl



Testování klasifikátoru

CIIRC

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

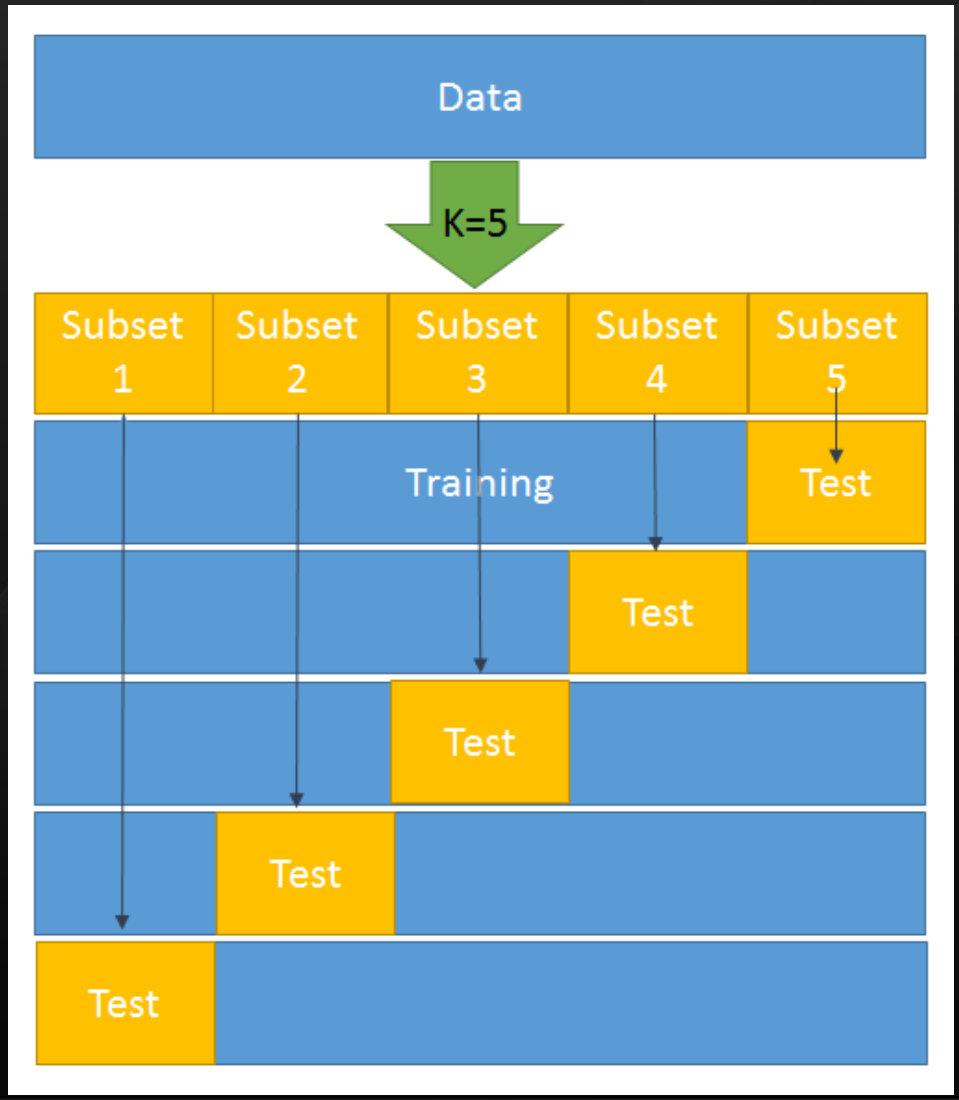


Testování klasifikátoru

CIIRC

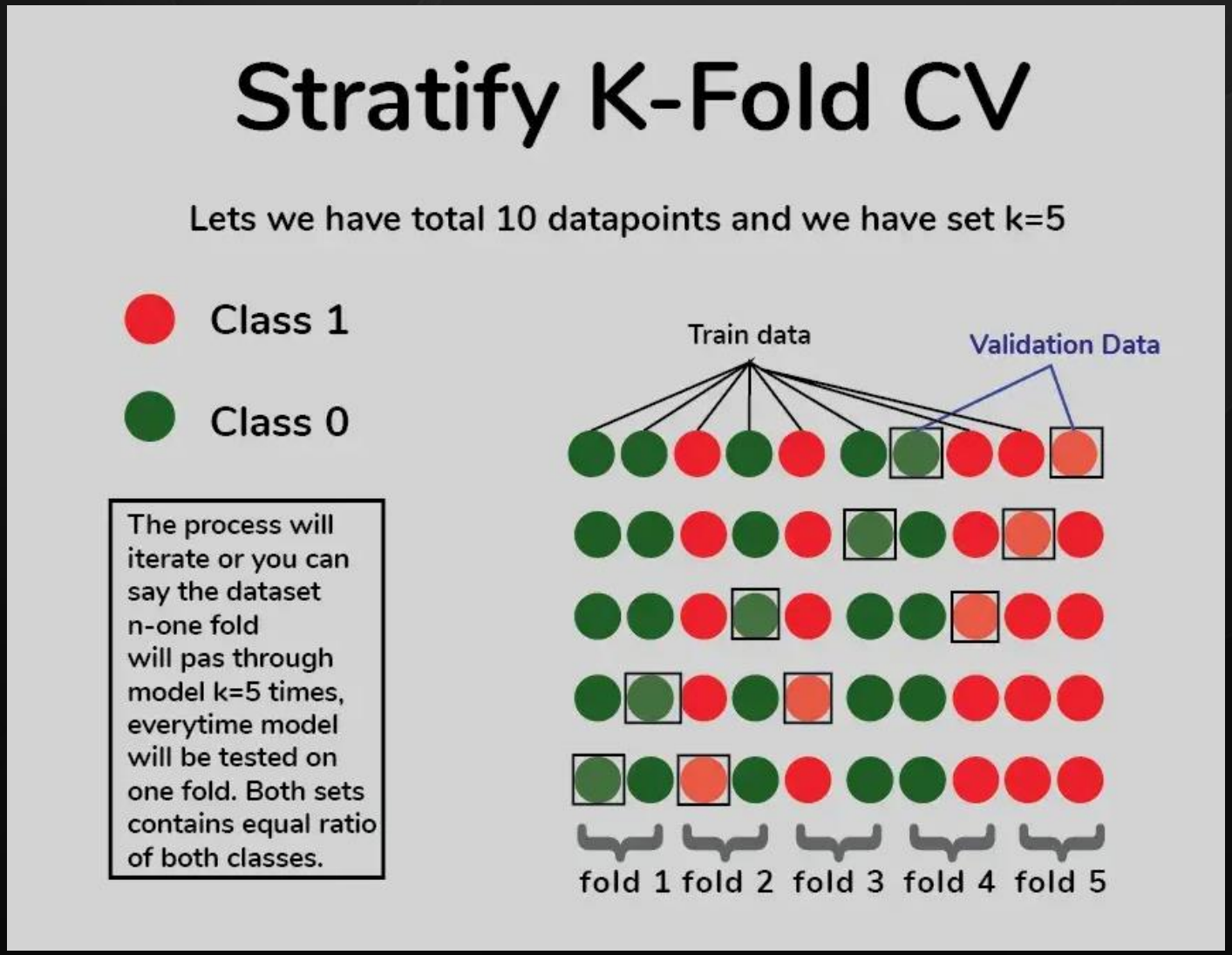
Křížová validace – rozdělení na části, iterativní použití vždy jen jedné části jako testovací a zbytku jako trénovací

Např. 5 skupin (foldů).



Testování klasifikátoru

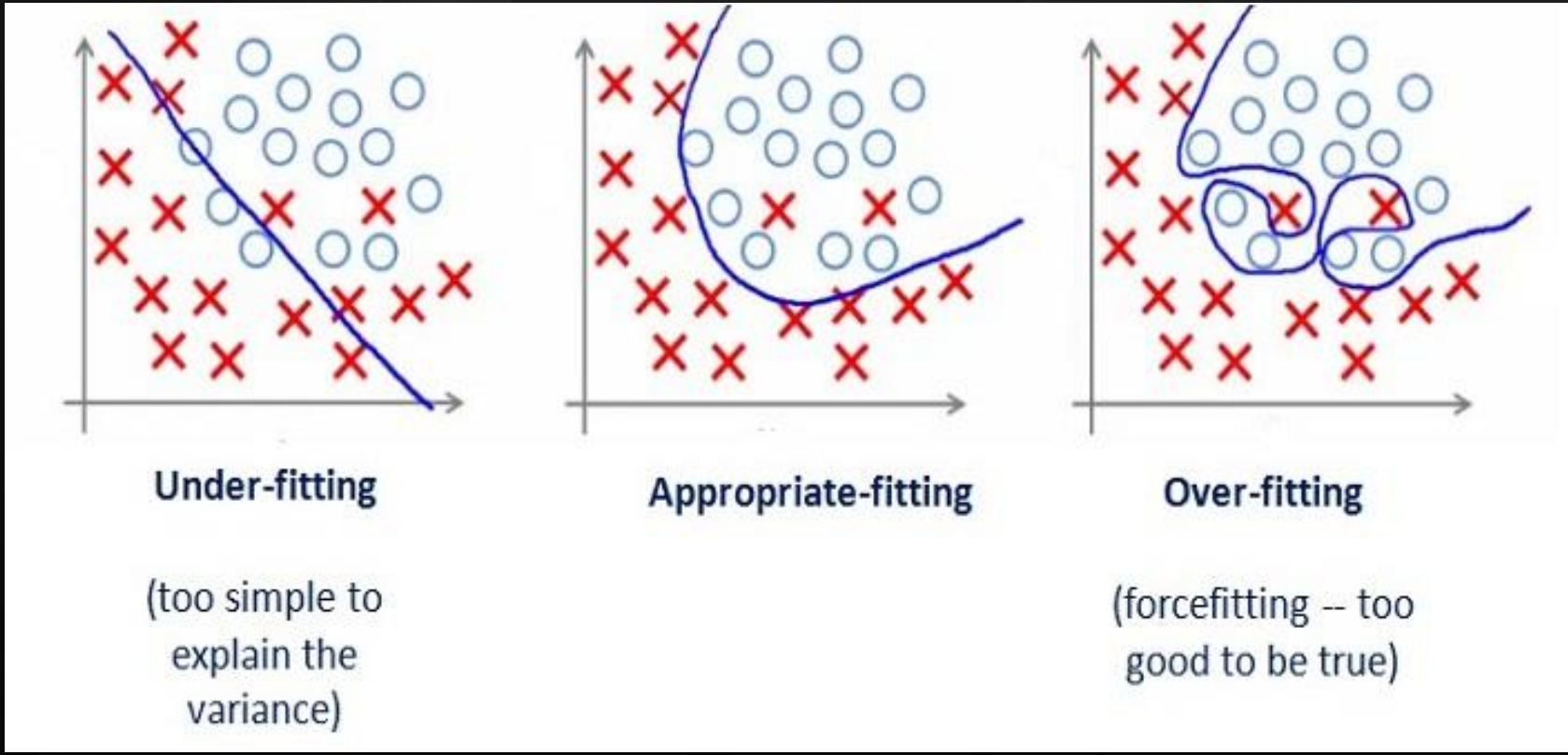
CIIRC



Složitost klasifikátoru

CIIRC

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



Shluková analýza

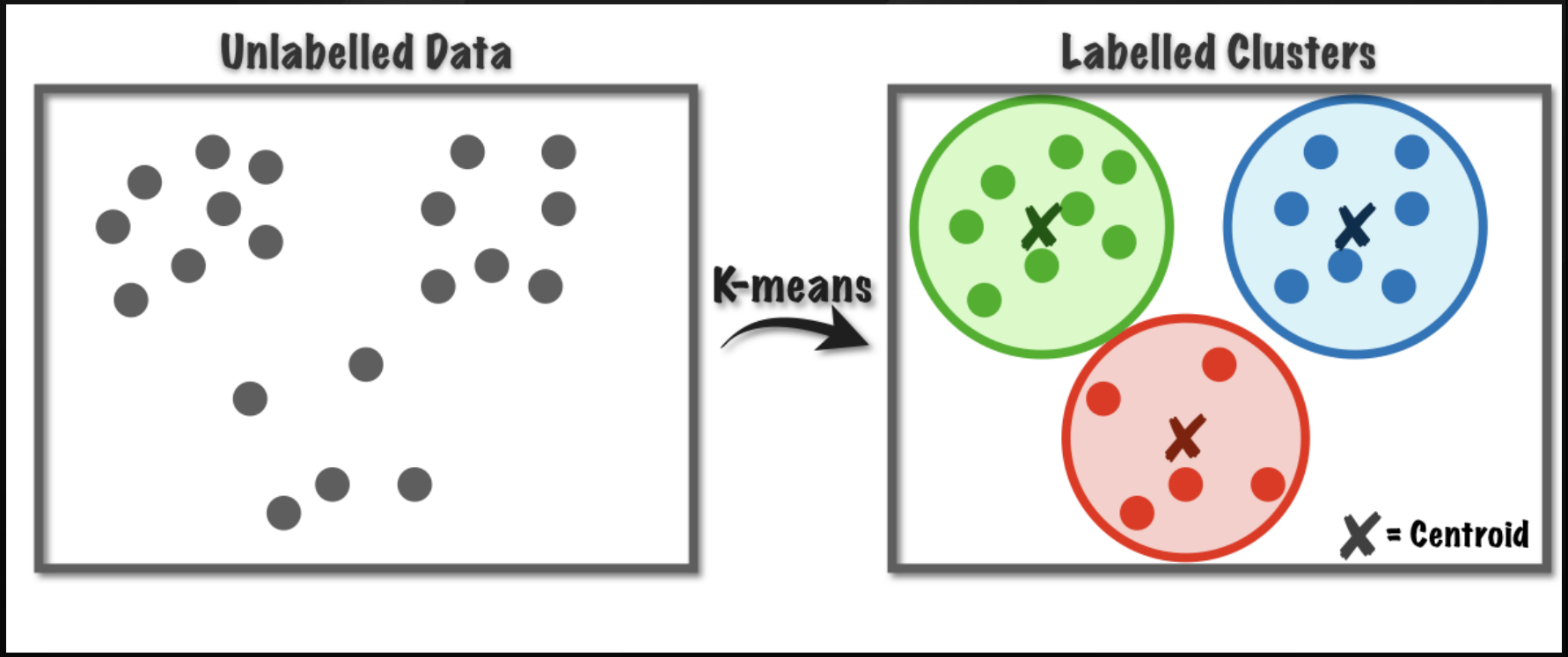
CIIRC

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

K-means

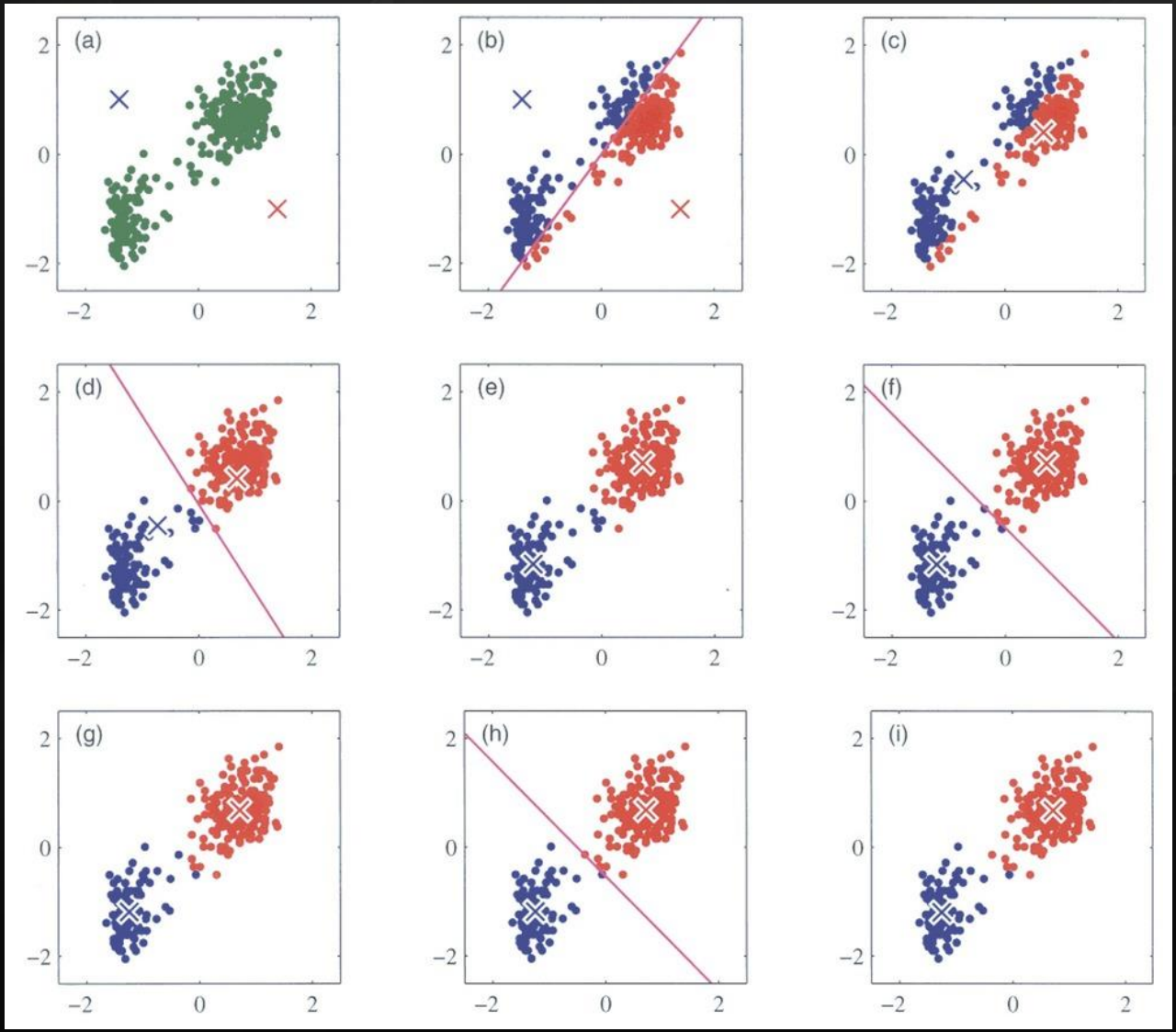
CIIRC

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



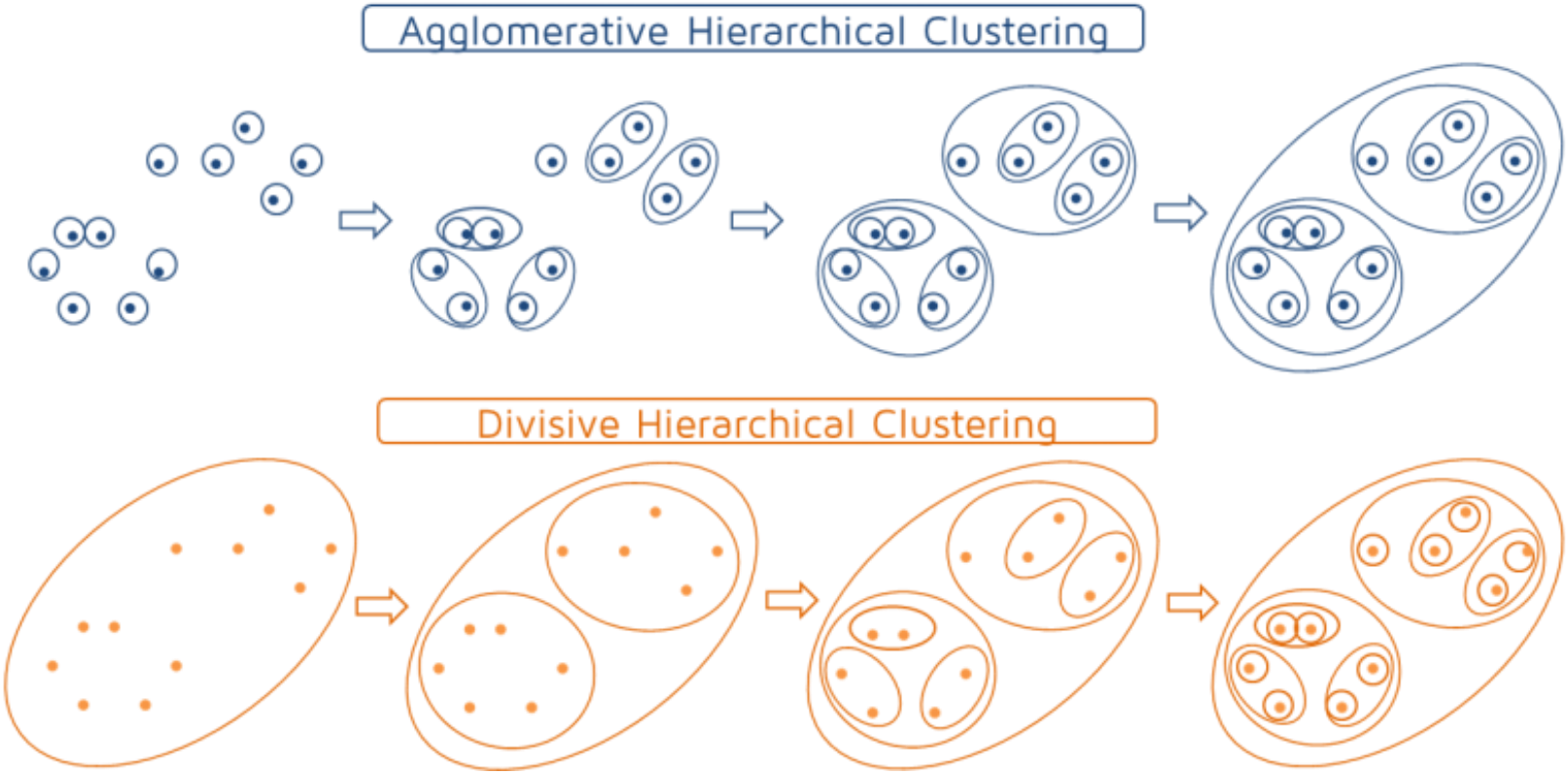
K-means algoritmus

CIIRC



Hiearachické shlukování

CIIRC



Shrnutí

CIIRC

- Kvalitní příznaky (jejich počet << 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