Supervised learning
"učenie s učiteľom"
2. časť

Zuzana Rošťáková

18. november 2021

Seminár UM

Učenie s učiteľom - supervised learning

- vstupné údaje:
  - matica X o rozmere $n \times p$
  - **n** pozorovaných objektov
  - pre i-ty objekt pozorujeme **p** premenných v tzv. vektore čŕt

$$\mathbf{x}_i = \left(x_{i_\mathbf{1}}, x_{i_\mathbf{2}}, \ldots, x_{i_p}\right)^T$$

  - premenné môžu byť
    - kardinálne (číselné) - napr. vek, výška, krvný tlak, ...
    - kategorické - napr. krvná skupina, pohlavie, vzdelanie, ...

- výstupné údaje:
  - kardinálne - reálne čísla $\rightarrow y_1, \ldots, y_n$
    $\rightarrow$ **regresná analýza**
  - kategorické - označenie skupiny/triedy $\rightarrow c_1, c_2, \ldots, c_n$
    $\rightarrow$ **klasifikačná analýza**

Klasifikácia - prehľad metód

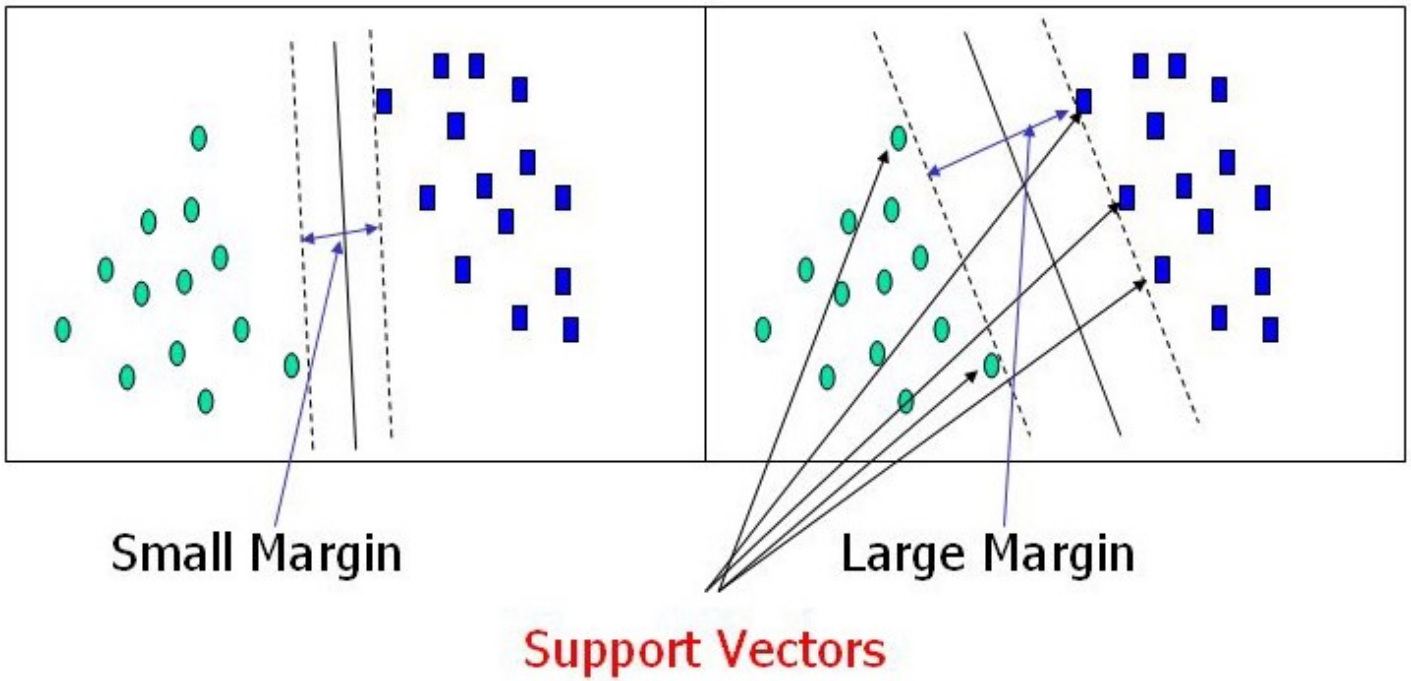
- **k-najbližších susedov (k-nearest neighbours)**
- **klasifikačné stromy**
- **diskriminačná analýza - lineárna, kvadratická**
- **naivná Bayesovská klasifikácia (naïve Bayes)**
- **logistická regresia**
- metóda nosného bodu (support vector machine)
- neurónové siete
- ...

Metóda nosného bodu

Support vector machine (SVM)

Support vector machine - Metóda nosného bodu

- klasifikácia do dvoch tried $\rightarrow$ -1 a 1
- **idea:** nájsť $(p-1)$-rozmernú nadrovinu v p-rozmernom priestore, ktorá najlepšie separuje tieto dve triedy
  - $\rightarrow$ má maximálnu vzdialenosť od "najbližšieho" bodu ľubovoľnej triedy
  - $\rightarrow$ nosné body (support vectors) - trénovacie dáta z oboch skupín, ktoré sú "najbližšie" k separačnej nadrovine, t.j. definujú ju



https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Support vector machine - lineárna verzia

- **predpoklad:** triedy sú lineárne separovateľné, t.j. $\exists\, \mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}$

$$\mathbf{w}^T \mathbf{x_i} - b > 0, \text{ ak } c_i = 1$$
$$\mathbf{w}^T \mathbf{x_i} - b < 0, \text{ ak } c_i = -1,$$

- potom nadrovina $\mathbf{w}^T \mathbf{x} - b = 0$ separuje triedy -1 a 1
  $\Rightarrow$ chceme nájsť tú, ktorá ich separuje maximálne
- konvexný optimalizačný problém $\rightarrow$ riešenie cez niektorý "solver"

$$\mathbf{w} \in \text{argmin}_{\|\mathbf{w}\|} c_i \left(\mathbf{w}^T \mathbf{x}_i - b \right) \geq 1, \text{ pre všetky } i = 1, \ldots, n$$

- ak triedy nie sú separovateľné $\rightarrow$ penalizácia

$$\mathbf{w} \in \text{argmin}_{\|\mathbf{w}\|} \frac{1}{n} \sum_{i=1}^{n} \max \left(0, 1 - c_i(\mathbf{w}^T \mathbf{x_i} - b) \right) + \lambda \|\mathbf{w}\|^2$$

Support vector machine - nelineárna verzia

- algoritmus je formálne podobný lineárnej verzií SVM
- **ROZDIEL:** skalárny súčin v lineárnej verzií je nahradený nelineárnou jadrovou funkciou
  - Gaussovská (radial basis function - rbf)

$$k(\mathbf{x_i}, \mathbf{x_j}) = e^{-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2}, \ \gamma > 0$$

  - polynomická

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i}^T \mathbf{x_j})^q, q \text{ je rád polynómu}$$

Support vector machine v MATLAB-e

Mdl = fitcsvm(X,C,Name,Value)

- X - $n \times p$ matica pozorovaní
- C - $n \times 1$ vektor zaradenia do tried
- voliteľné parametre
  - KernelFunction - linear (default pre dve triedy), gaussian / rbf (default pre 1 triedu), polynomial (+ špecifikovať rád q)
  - PolynomialOrder - q, ak KernelFunction = polynomial
  - Solver - ISDA, L1QP, SMO $\rightarrow$ algoritmus na hľadanie $\mathbf{w}$
  - OutlierFraction $\in [0, 1)$ - podiel outlierov v trénovacej vzorke
  - IterationLimit - maximálny počet iterácií
  - ...
- *Mdl* - natrénovaný model

Support vector machine v MATLAB-e

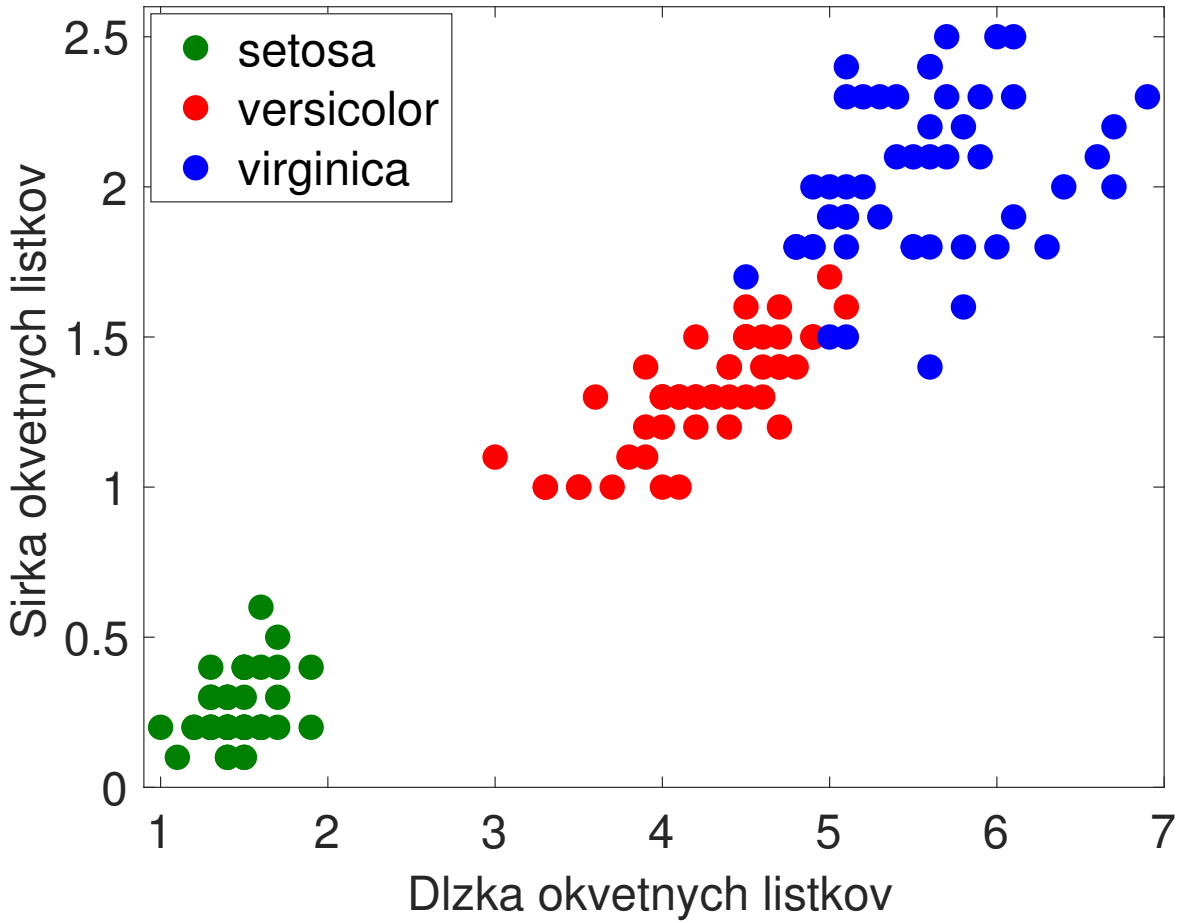
ak počet tried > 2 $\rightarrow$ len lineárna verzia

Mdl = fitcecoc(X,C,Name,Value)

- X - $n \times p$ matica pozorovaní
- C - $n \times 1$ vektor zaradenia do tried
- voliteľné parametre
  - CategoricalPredictors - zoznam kategorických premenných
  - ...
- *Mdl* - natrénovaný model

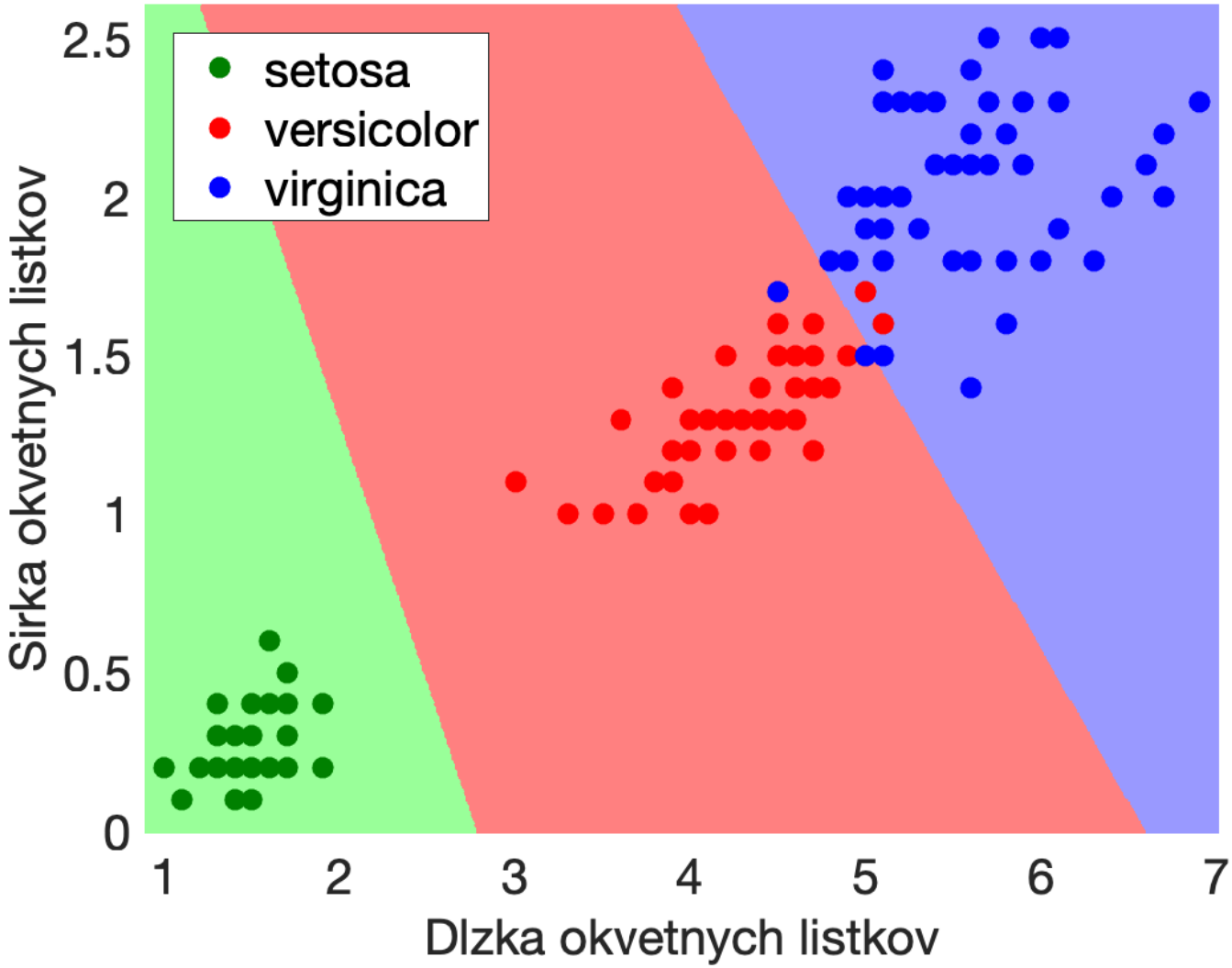
Support vector machine - Príklad: Kosatce

- databáza **Fisher Iris** v MATLAB-e
- 150 pozorovaní, 3 triedy $\rightarrow$ musíme ísť cez funkciu *fitcecoc*
- 2 premenné - dĺžka a šírka okvetného lístka



Support vector machine - Príklad: Kosatce

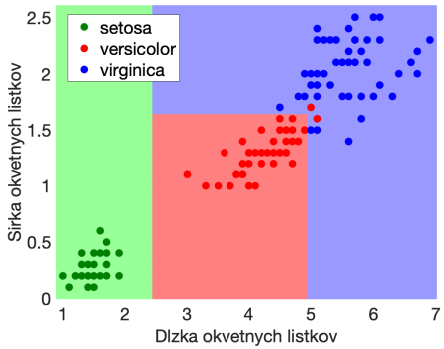
$$mdlSVM = fitcecoc(X, C)$$



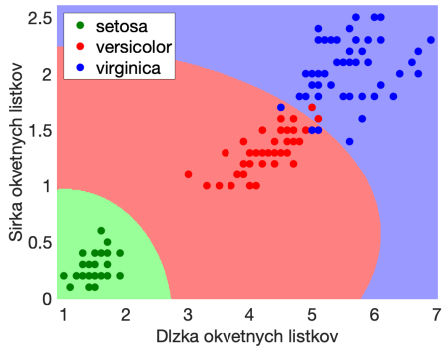
Support vector machine - Príklad: Kosatce

KNN

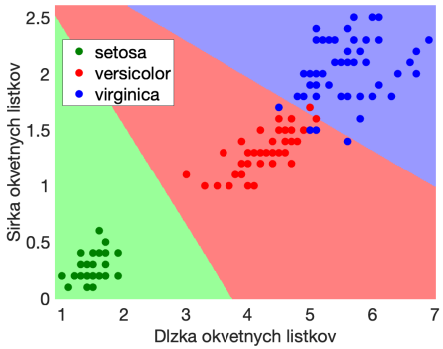
Ctree



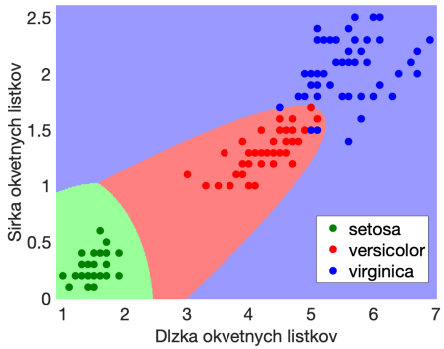
NBC



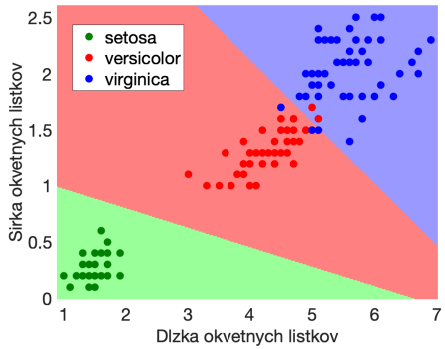
LDA



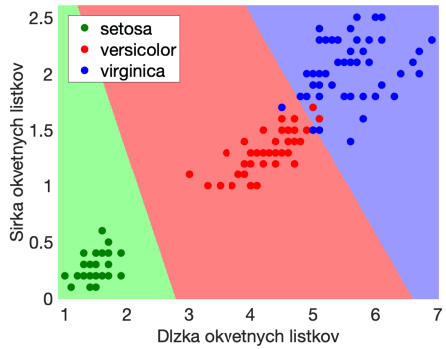
QDA



LogReg



SVM



Support vector machine zhrnutie

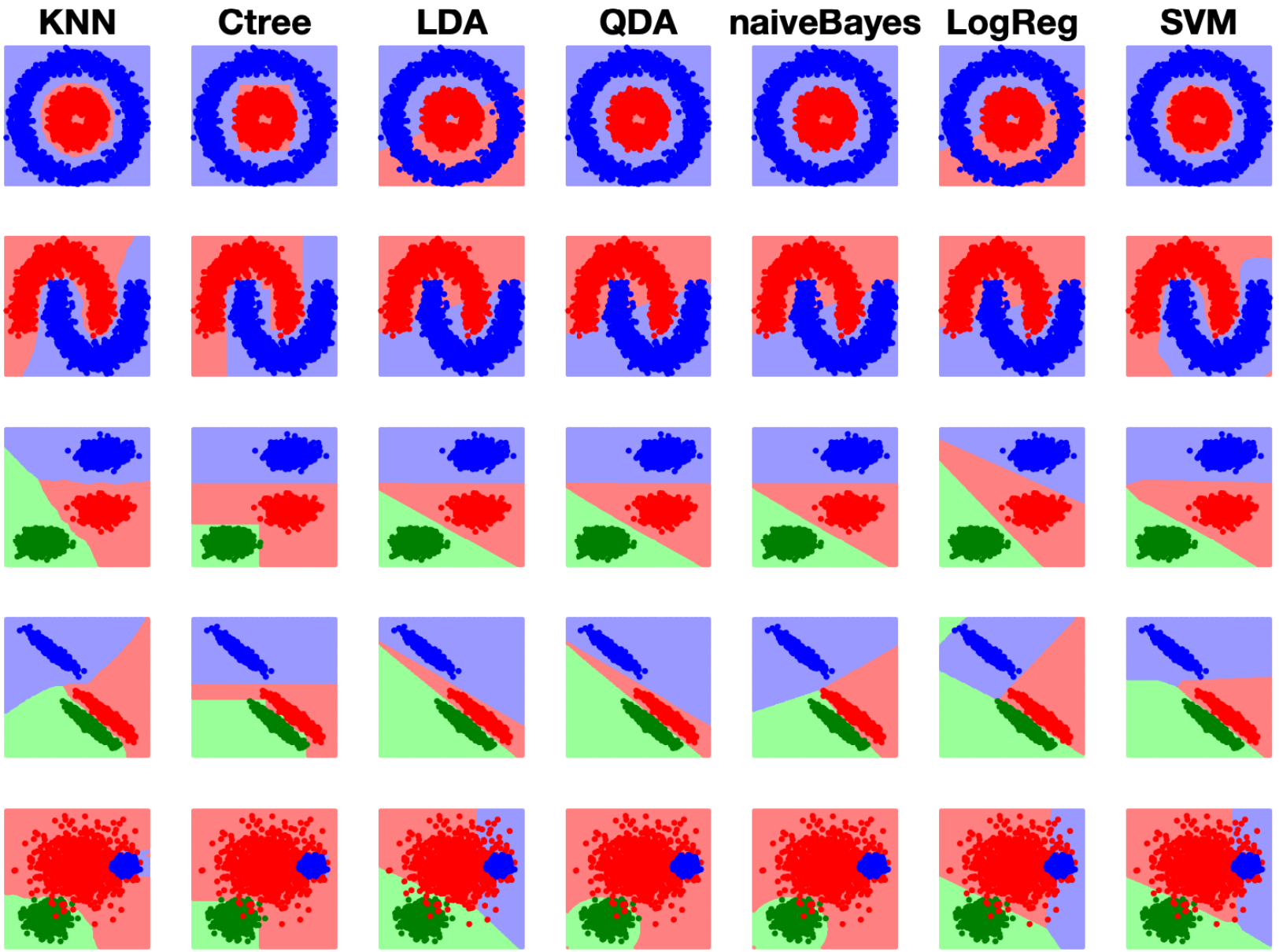
- **výhody:**
  - jedna z najrobustnejších klasifikačných metód
  - komplexná, nelineárna separácia tried
  - efektívne riešenie optimalizačného problému pomocou kvadratickej konvexnej optimalizácie
  - vie pracovať aj s kategorickými premennými
- **nevýhody:**
  - *fitcsvm* pracuje len s 1 alebo dvoma triedami $\rightarrow$ v prípade viactriedovej klasifikácie treba použiť funkciu *fitcecoc* alebo inú metódu
  - *fitcsvm* dokáže pracovať len s dátami "vhodnej" dimenzie $\rightarrow$ pre vysokodimenzionálne dáta sa odporúča použiť rutinu *fitclinear*

Voľba klasifikátora

Voľba klasifikátora

- vhodnosť pre "naše" dáta
- rýchlosť natrénovania modelu
- náročnosť trénovania - odhad parametrov, programovanie, ...
- rýchlosť klasifikácie nového pozorovania
- náročnosť porozumenia, čo sa deje "vnútri" klasifikátora
  - problém interpretovateľnosti natrénovaného modelu
  - **black-box classifiers** - model funguje, ale nie je úplne priamočiare pochopiť ako
- podiel prvkov zaradených do nesprávnej triedy (missclassification)

Porovnanie klasifikátorov



Porovnanie klasifikátorov - čas trénovania

- n = 3000 pozorovaní
- pre data1 a data2 bola použitá nelineárna verzia SVM
- pre data3, data4 a data5 bola použitá lineárna verzia SVM

data	KNN	Ctree	LDA	QDA	NB	LogReg	SVM
1	0.0096	0.0099	0.0061	0.0088	0.0068	0.0270	0.2839
2	0.0182	0.0189	0.0091	0.0135	0.0072	0.0939	0.1453
3	0.0107	0.0159	0.0186	0.0132	0.0056	1.0579	0.1949
4	0.0135	0.0116	0.0069	0.0141	0.0054	0.9526	0.0593
5	0.0098	0.0243	0.0077	0.0148	0.0059	0.8157	0.0641

Porovnanie klasifikátorov - čas predikcie

- n = 35616 nových pozorovaní
- pre data1 a data2 bola použitá nelineárna verzia SVM
- pre data3, data4 a data5 bola použitá lineárna verzia SVM

data	KNN	Ctree	LDA	QDA	NB	LogReg	SVM
1	60.5741	0.5609	0.8090	0.9459	0.7024	0.2530	15.0400
2	21.7024	0.2692	0.3456	0.3554	0.3005	0.1144	2.6937
3	0.3969	0.0069	0.0085	0.0090	0.0075	0.0037	0.0632
4	0.1849	0.0029	0.0043	0.0038	0.0041	0.0011	0.0086
5	0.5837	0.0044	0.0078	0.0070	0.0068	0.0028	0.0113

Podiel prvkov zaradených do nesprávnej triedy

$$r = \frac{\text{počet prvkov zaradených do nesprávnej triedy}}{n}$$

- ďalšou pomôckou je tzv. confusion (missclassification) matrix

$$M \in \mathbb{R}^{L \times L}$$

M_{ij} = pravdepodobnosť, že objekt z triedy i je zaradený do triedy j

$\Rightarrow$ v optimálnom prípade je M (skoro) diagonálna

$\rightarrow$ Ako odhadnúť r a M ?

Odhady r, M - typy validácie

1. in-sample

- model natrénovaný na všetkých pozorovaniach $\mathbf{x_1}, \ldots, \mathbf{x_n}$
- pomocou modelu predikujeme zaradenie do tried pre $\mathbf{x_1}, \ldots, \mathbf{x_n}$

$$\rightarrow c_1^\star, \ldots, c_n^\star$$

- porovnanie skutočných $c_1, \ldots, c_n$ a predikovaných zaradení $c_1^\star, \ldots, c_n^\star$
- **výhoda:** pri trénovaní použijeme všetky dostupné pozorovania
- **nevýhoda:** nadhodnocuje r, M !

Odhady r, M - typy validácie

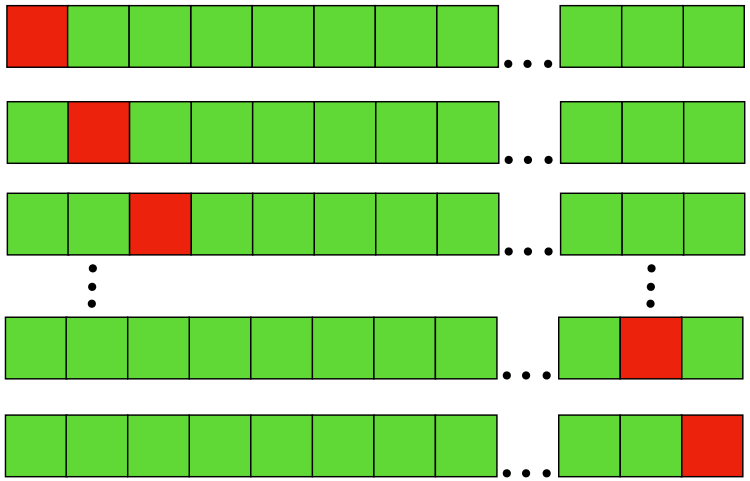
2. validačná metóda

- dáta rozdelíme na dve časti
  - trénovacia vzorka
    - $\rightarrow$ natrénovanie klasifikátora (modelu)
    - $\rightarrow$ väčšia časť dát
  - validačná (testovacia) vzorka
    - $\rightarrow$ na ňu aplikujeme klasifikátor
    - $\rightarrow$ porovnanie skutočných a predikovaných zaradení do tried
- **výhoda:** dobrý odhad r, M
- **nevýhoda:** model trénovaný len na časti dostupných dát
  $\rightarrow$ použitím celej databázy by sme mohli dostať lepší klasifikátor

Odhady r, M - typy validácie

3. krížová validácia (cross - validation)

- **leave-one-out**
  - validačnú metódu opakujeme n-krát
  - **i-ty krok:** $\mathbf{x_i}$ je vo validačnej vzorke, ostatných $n-1$ pozorovaní tvorí trénovaciu vzorku

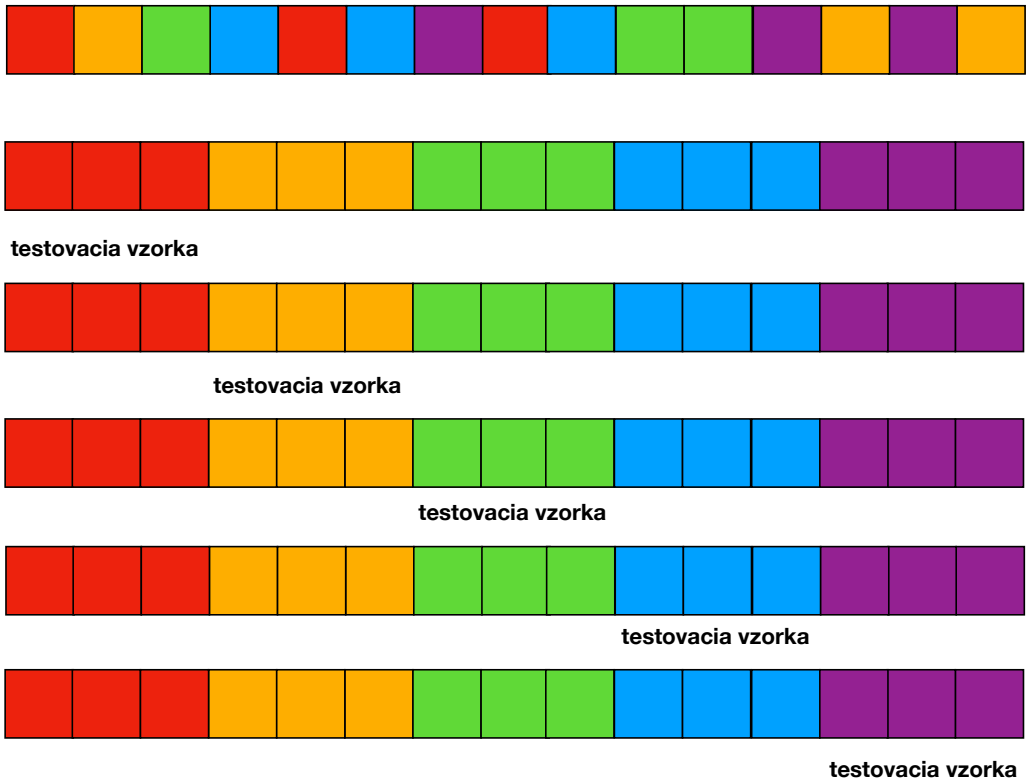


- **leave-p-out**
  - p pozorovaní je vo validačnej vzorke, $n-p$ v trénovacej
  - opakujeme validačnú metódu

Odhady r, M - typy validácie

3. krížová validácia (cross - validation)

- k-fold cross-validation
  - rozdelenie databázy na k častí
  - **i-ty krok:** i-ta časť použitá ako validačná vzorka, zvyšok ako trénovacia vzorka
  - ak $k = n \rightarrow$ leave-one-out cross-validation



In-sample odhady r, M v MATLAB-e

- **r = resubLoss(mdl)**

  mdl - natrénovaný klasifikátor, model

  r - podiel dát zaradených do nesprávnej triedy (in-sample)

  - $r \in [0, 1]$
  - $r \geq 0.5 \rightarrow$ model nefuguje dobre ani pre dáta, na ktorých bol trénovaný

- $C^\star$ = **resubPredict(mdl)**
  M = confusionmat(C,$C^\star$)

  - odhad matice M pomocou in-sample validácie (nadhodnotené)

  mdl - natrénovaný klasifikátor, model

  C - zaradenie do tried pôvodných dát

  $C^\star$ - predikované zaradenie do tried pomocou *mdl*

  M - odhad confusion matrix

Krížová validácia v MATLAB-e

- **[train, test] = crossvalind(Method, . . .)**
- **ind = crossvalind('Kfold',n,k)**

Method - metóda použitá pri krížovej validácií

- 'Kfold', n, k - n je počet pozorovaní, k je počet častí
- 'HoldOut', n, p - $p.n$ pozorovaní tvorí testovaciu vzorku
- 'LeaveMout', n, m - m pozorovaní tvorí testovaciu vzorku

ind - označenie skupiny, do ktorej sú pozorovania zaradené

train - indexy prvkov v trénovacej vzorke

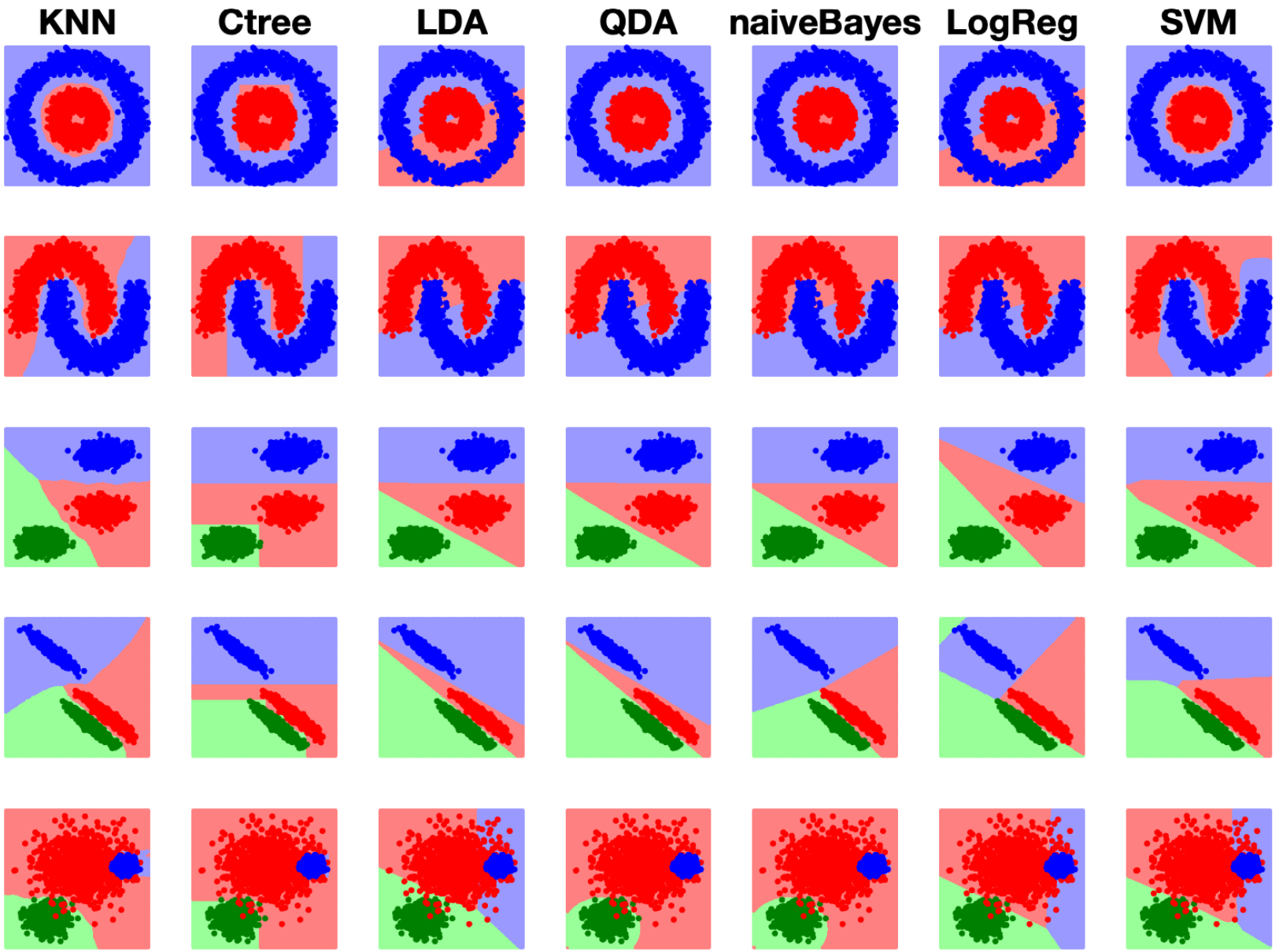
test - indexy prvkov v testovacej vzorke

Porovnanie klasifikátorov - in-sample validácia

- n = 3000 pozorovaní
- pre data1 a data2 bola použitá nelineárna verzia SVM
- pre data3, data4 a data5 bola použitá lineárna verzia SVM

data	KNN	Ctree	LDA	QDA	NB	LogReg	SVM
1	0	0	0.4787	0.0017	0.0017	0.4787	0
2	0	0	0.1300	0.1310	0.1307	0.1353	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0.0003	0	0
5	0.0123	0.0073	0.0677	0.0170	0.0163	0.0310	0.0350

Porovnanie klasifikátorov



Porovnanie klasifikátorov - 10-fold cross-validation

- $k = 10 \rightarrow ind = crossvalind('Kfold', 3000, 10)$
  - 2700 pozorovaní v trénovacej vzorke
  - 300 pozorovaní v testovacej vzorke
- pre data1 a data2 bola použitá nelineárna verzia SVM
- pre data3, data4 a data5 bola použitá lineárna verzia SVM

data	KNN	Ctree	LDA	QDA	NB	LogReg	SVM
1	0	0.0027	0.4873	0.0027	0.0027	0.4870	0
2	0	0.0027	0.1293	0.1303	0.1297	0.1257	0.0003
3	0	0.0003	0	0	0	0	0
4	0	0.0040	0	0	0.0003	0	0
5	0.0200	0.0277	0.0683	0.0170	0.0160	0.0317	0.0360

Otázky ?