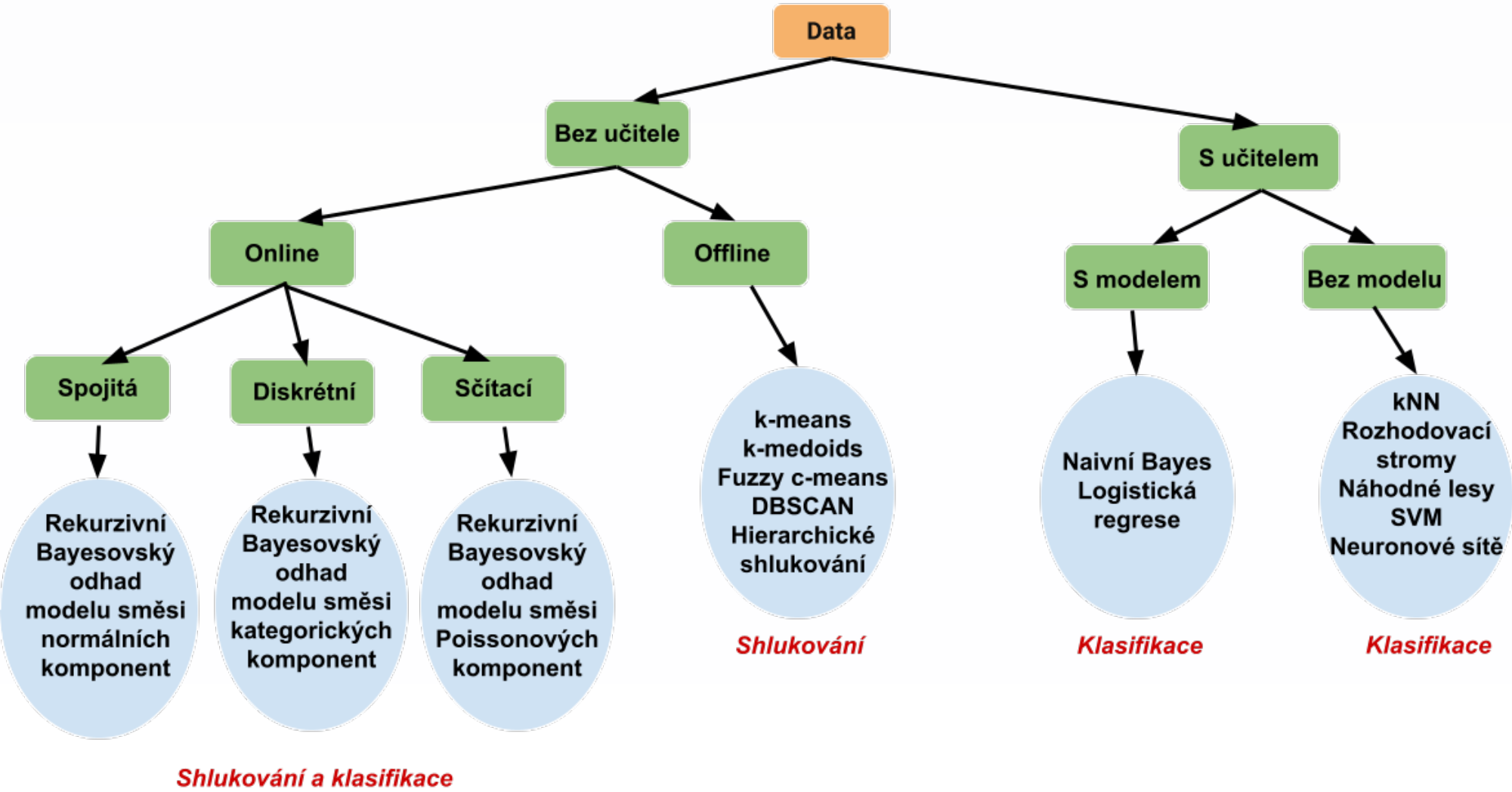
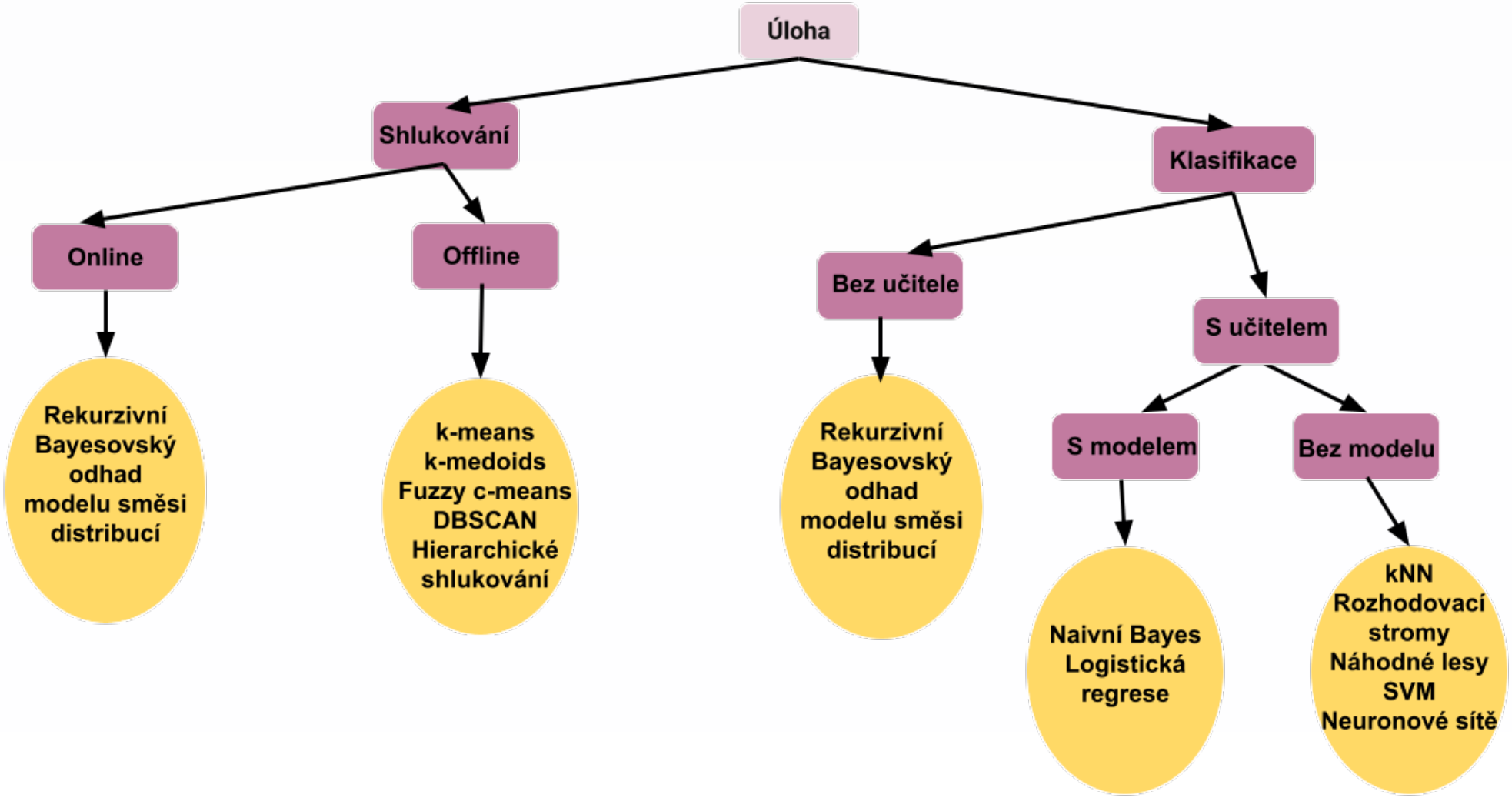
Přednáška 9 – Shrnutí a opakování algoritmů

Jak zvolit vhodný algoritmus podle dat?



Jak zvolit vhodný algoritmus podle úlohy?



Příklady – který algoritmus je vhodný?

1. Měříme: tep a vzdálenost na hodinkách.
   Co chceme: detekovat v reálném case, zda se jedná o klidový režim, chůzi nebo běh.

2. Měříme: hladinu hluku ve sluchátkách (pravé, levé).
   Co chceme: detekovat v reálném case, zda hladina zvuku je bezpečná.

3. Máme data: věk a cena automobilových náhradních dílů zařazených do kategorií "nové originální", "neoriginální", "použité originální".
   Co chceme: zařadit do kategorií další náhradní díly

4. Měříme: počty návštěvníků webu za minutu a jejich lokalitu.
   Co chceme: určit, zda aktuální provoz webu je normální nebo se jedná o zvýšenou aktivitu.

5. Měříme: hodnoty z NOx senzoru před a po NOx katalyzátoru v autě.
   Co chceme: za provozu vozidla vyhodnotit, zda se jedná o správnou funkci katalyzátoru nebo poruchu.

Otázky 1– 26: data, shluky, modely jednotlivých shluků a jejich odhad

Data:

- spojitá • diskrétní
- sčítací – počty náhodných nezávislých událostí za jednotku času count data
- multimodální – naměřená v jednotlivých režimech pozorovaného systému
- tvoří shluky

Modely jednotlivých shluků a rekurzivní Bayesovský odhad

Normální model (spojitá data):

$$y_t = \theta + e_t$$

$$V_t = V_{t-1} + \begin{bmatrix} y_t \\ 1 \end{bmatrix} [y_t' \;\; 1], \quad \kappa_t = \kappa_{t-1} + 1$$

$$V_t = \begin{bmatrix} V_{n\times n} & V_y' \\ V_y & V_n \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \cdots & \cdots \\ \cdots & \cdots \end{bmatrix} & | \\ \hline & . \end{bmatrix}$$

$$\hat{\theta}_t = V_n^{-1} V_y$$

$$\hat{R}_t = \frac{V_{n\times n} - V_y' V_n^{-1} V_y}{\kappa_t}$$

Kategorický model (diskrétní data):

y_t	1	2	...	N
Θ	Θ_1	Θ_2	...	Θ_N

$$\nu_{i;t} = \nu_{i;t-1} + \delta(i;\, y_t)$$

$$\hat{\Theta}_{i;t} = \frac{\nu_{i;t}}{\sum_{j=1}^{n} \nu_{j;t}}$$

Poissonův model (sčítací data):

$$f(y_t|\lambda) = \exp\{-\lambda\} \frac{\lambda^{y_t}}{y_t!}$$

$$S_t = S_{t-1} + y_t, \quad \kappa_t = \kappa_{t-1} + 1$$

$$\hat{\lambda}_t = \frac{S_t}{\kappa_t}$$

Otázky 27 – 29, 33, 34: model směsi distribucí – obecně

- komponenty – modely jednotlivých shluků $f_j(y_t|\Theta), j = 1, \ldots, n_c$
- ukazovátko – diskrétní náhodná veličina $c_t \in \{1, \ldots, n_c\}$,

ukazuje aktivní komponentu

Základní princip – obecně:

- Inicializace komponent

V cyklu:

- Měříme data
- Vzdálenosti od jednotlivých komponent – proximity
- Normalizace proximit – váhy komponent (pravděpodobnost, že je komponenta aktivní)
- Maximální váha – bodový odhad ukazovátka (klasifikace)
- Update statistik s váženými daty
- Přepočet bodových odhadů podle typu komponent
- Jdeme na krok 1

Otázky 30 – 32, 35 – 40: shlukování a klasifikace online s modelem směsi normálních komponent

- spojitá data, každá komponenta – normální model
- generování:

```
c(t)=sum(rand(1,1)>cumsum(al))+1;
y(:,t)=th(:,c(t))+sqrt(r)*randn(2,1);
```

- vzdálenost – rozptyly
- inicializace – histogram, x-y graf

Pro čas $t = 0$

1. Počet komponent, počáteční bodové odhady $(\hat{\theta}_0)_j$, $(\hat{r}_0)_j$, statistiky

Pro čas $t = 1, 2, \ldots$, pro každou komponentu

1. Měříme nová data y_t
2. Proximity m_j – dosadíme do komponent $(\hat{\theta}_{t-1})_j$, $(\hat{r}_{t-1})_j$ a y_t
3. Váhy komponent $w_{j;t} = \frac{m_j}{\sum_{i=1}^{n_c} m_i}$
4. Bodový odhad ukazovátka: $\hat{c}_t = \arg\max_j w_t$
5. Update statistik s váženými daty:

$$(V_t)_j = (V_{t-1})_j + w_{j;t} \begin{bmatrix} y_t \\ 1 \end{bmatrix} [y_t' \;\; 1], \quad (\kappa_t)_j = (\kappa_{t-1})_j + w_{j;t}$$

6. Přepočet bodových odhadů $(\hat{\theta}_t)_j$, $(\hat{r}_t)_j$

Otázky 41 – 44: shlukování a klasifikace online s modelem směsi kategorických komponent

- diskrétní data, každá komponenta – kategorický model
- generování:

```
c(t)=sum(rand(1,1)>cumsum(al))+1;
pp=cumsum(th{c(t)});
y(t)=sum(rand(1,1)>pp)+1;
```

Algoritmus

Pro čas $t = 0$

1. Počet komponent, počáteční statistiky expertně, bodové odhady $(\hat{\Theta}_0)_j$

Pro čas $t = 1, 2, \ldots$, pro každou komponentu

1. Měříme nová data y_t
2. Proximity m_j – dosadíme do komponent $(\hat{\Theta}_{t-1})_j$ a y_t
3. Váhy komponent $w_{j;t} = \frac{m_j}{\sum_{i=1}^{n_c} m_i}$
4. Bodový odhad ukazovátka: $\hat{c}_t = \arg\max_j w_t$
5. Update statistik s váženými daty: $(\nu_{i;t})_j = (\nu_{i;t-1})_j + w_{j;t}\delta(i;\, y_t)$
6. Přepočet bodových odhadů $(\hat{\Theta}_{i;t})_j = \frac{(\nu_{i;t})_j}{\sum_{k=1}^{n}(\nu_{k;t})_j}$

Otázky 45 – 51: shlukování a klasifikace online s modelem směsi Poissonových komponent

- sčítací data, každá komponenta – Poissonův model
- generování: `c(t)=sum(rand(1,1)>cumsum(al))+1;`
  `for i=1:ny y(i,t)=poissrnd(La{i}(c(t))) end`
- vzdálenost – λ
- inicializace – expertně, histogram, x-y graf

Algoritmus

Pro čas $t = 0$

1. Počet komponent, počáteční bodové odhady $(\hat{\lambda}_0)_j$, statistiky

Pro čas $t = 1, 2, \ldots$, pro každou komponentu

1. Měříme nová data y_t
2. Proximity m_j – dosadíme do komponent $(\hat{\lambda}_{t-1})_j$ a y_t
3. Váhy komponent $w_{j;t} = \frac{m_j}{\sum_{i=1}^{n_c} m_i}$
4. Bodový odhad ukazovátka: $\hat{c}_t = \arg\max_j w_t$
5. Update statistik s váženými daty:
   $(S_t)_j = (S_{t-1})_j + w_{j;t} y_t, (\kappa_t)_j = (\kappa_{t-1})_j + w_{j;t},$
6. Přepočet bodových odhadů $(\hat{\lambda}_t)_j = \frac{(S_t)_j}{(\kappa_{t-1})_j}$

Otázky 52 – 55: Poissonova regrese (predikce sčítacích dat)

y_t – Poissonovo rozdělení

$$f(y_t|\lambda) = \exp\{-\lambda\}\frac{\lambda^{y_t}}{y_t!}$$

Poissonova regrese

$$\ln(\underbrace{\bar{y}_t}_{\lambda}) = x_{1;t}\theta_1 + x_{2;t}\theta_2 + \ldots + k$$

λ = průměr = rozptyl

$x_t = [x_{1;t} \quad x_{2;t} \quad \ldots \quad x_{n;t}]$ – vysvětlující data

Linearizace + nejmenší čtverce

metoda 1

$$\underbrace{\begin{bmatrix} \ln(y_1) \\ \ln(y_2) \\ \ldots \\ \ln(y_{nd}) \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} x_{1;1} & x_{2;1} & \ldots & 1 \\ x_{1;2} & x_{2;2} & \ldots & 1 \\ \ldots & & & \\ x_{1;nd} & x_{2;nd} & \ldots & 1 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \theta_1 \\ \theta_2 \\ \ldots \\ k \end{bmatrix}}_{\theta}$$

Odhad offline:

trénovací data

$$\hat{\theta} = (X'X)^{-1}X'Y$$

Predikce online:

testovací data

$$\hat{Y} = \exp\{X\hat{\theta}\}$$

Metoda maximální věrohodnosti

metoda 2

$$\ln L_t(\theta) = \ln \prod_{\tau=1}^{t} \text{model}(\text{trénovací data}_\tau), \quad \hat{\theta} = \arg\max_{\theta} \ln L_t(\theta)$$

Otázky 56 – 61: Klasifikace offline na základě modelu s učitelem (NB, LR)

- s učitelem – známé shluky (ukazovátko)

Naivní Bayesův klasifikátor:

- vektor veličin – jednotlivé veličiny – podmíněná nezávislost
- model:

$$f([y_{1;t}, y_{2;t}]|\Theta, c_t) = \prod_{i=1}^{N=2} f(y_{i;t}|\Theta_i, c_t)$$

- Postup:

1. Trénování modelů – trénovací data (s učitelem – měřené ukazovátko)
   - odhad parametrů modelů (filtrace každé veličiny + offline odhad)

2. Testování (klasifikace) – testovací data (bez ukazovátka)
   - odhad ukazovátka z odhadnutých modelů

$$f(c_t|y_t) = \underbrace{\text{normovaný histogram ukazovátka}}_{f(c_t)} \times \underbrace{\text{součin modelů}}_{\prod_{i=1}^{N=2} f(y_{i;t}|\hat{\Theta}_i, c_t)}$$

Logistická regrese – binární / multinomiální

- model:
  - pro $c_t \in \{0, 1\}$:

$$\frac{\exp\{c_t z_t\}}{1 + \exp\{z_t\}}, \quad z_t = y_{1;t}\theta_1 + y_{2;t}\theta_2 + \ldots + k$$

  - pro $c_t \in \{0, 1, \ldots, N\}$:

$$\ln \frac{p_1}{p_0} = z_{t(1)}, \ln \frac{p_2}{p_0} = z_{t(2)} \ \ldots \ , \ln \frac{p_N}{p_0} = z_{t(N)}$$

Postup:

1. Trénování modelů – trénovací data (s učitelem – měřené ukazovátko)
   - odhad regresních koeficientů – maximální věrohodnost

$$\ln L_t(\theta) = \ln \prod_{\tau=1}^{t} \text{model}(\text{trénovací data}_\tau), \quad \hat{\theta} = \arg\max_\theta \ln L_t(\theta)$$

2. Testování (klasifikace) – testovací data (bez ukazovátka)
   - odhad ukazovátka

Otázky 62 – 63: vyhodnocení přesnosti klasifikace

Příklad: binární klasifikace (pozitivní, negativní)

- Chybová matice (confusion matrix)

	Skutečně pozitivní	Skutečně negativní
Pozitivní predikce	TP	FP
Negativní predikce	FN	TN

- Přesnost (accuracy)

Procento správných predikcí ze všech dat

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

Pro nevyvážená data:

- Preciznost (precision)

Procento správných pozitivních predikcí mezi pozitivními predikcemi

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

- Výtěžnost (recall)

Procento správných pozitivních predikcí mezi skutečně pozitivními

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

- Skóre F1 $\in (0, 1)$

$$F1 = 2\frac{Pre \times Rec}{Pre + Rec}$$

Otázky 64 – 69: shlukovací metody offline bez učitele (k-means, k-medoids, fuzzy, dbscan, hierarchické)

- bez učitele – hledámé shluky (nemáme ukazovátko)
- na základě vzdálenosti mezi datovými body

Euklidovská: $m_E(x, y) = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2}$

k-means

- vzdálenost (bod, středy)
- nejbližší střed – shluk
- průměr bodů ve shluku = nový střed

k-medoids

- vzdálenost (bod, medoidy)
- nejbližší medoid – shluk
- vzdálenost uvnitř shluků
- medoid za ne-medoid

Fuzzy c-means

- středy shluků
- matice příslušnosti

DBSCAN

- ϵ-okolí, minimální počet bodů
- shluk: vnitřní bod + všechny z něj dosažitelné
- hledáme hustě propojený shluk

Hierarchické shlukování

- aglomerativní – "zdola nahoru", divizivní – "shora dolů"
- propojení – min/max/průměrná vzdálenost

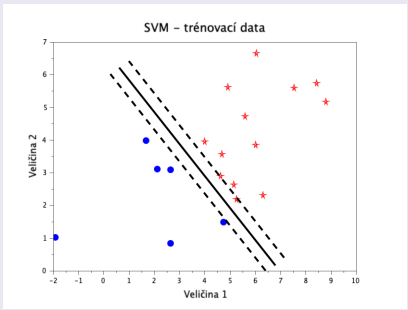
Otázky 70 – 75: klasifikace offline s učitelem (kNN, stromy, lesy, SVM, neuronové sítě)

- s učitelem – mámé shluky (máme ukazovátko)

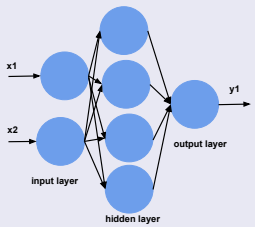
kNN

- vzdálenost (nový bod, body)
- k nejbližších bodů
- nový bod $\in$ do shluku většiny nejbližších

Podpůrné vektorové stroje (SVM)

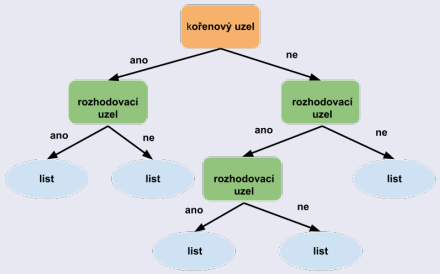


Neuronové sítě



- aktivační funkce
- vícevrstvé perceptrony
- zpětné šíření chyby

Rozhodovací stromy



- IG$\rightarrow$1 – lepší strom

Náhodné lesy

- Bootstrap Aggregation
- 10 – 100 náhodných vzorků
- les = 10 – 100 stromů
- každý strom v lese se učí
- kombinace výsledků stromů

Literatura:

Rekurzivní Bayesovské odhadování modelu směsi distribucí:

- I. Nagy, E. Suzdaleva. Algorithms and Programs of Dynamic Mixture Estimation. Unified Approach to Different Types of Components, Springer, 2017.
- Kárný, M., Böhm, J., Guy, T. V., Jirsa, L., Nagy, I., Nedoma, P., Tesař, L., 2006. Optimized Bayesian Dynamic Advising: Theory and Algorithms. Springer-Verlag, London.

Shlukování a klasifikace offline:

- D. T. Larose. Discovering Knowledge in Data. An Introduction to Data Mining. Willey, 2005.
- J. Han, M. Kamber, J. Pei. Data Mining: Concepts and Techniques, 3rd Edition. Morgan Kaufmann, 2011.
- M. J. Zaki, W. Meira Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014.
- G.James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor. An Introduction to Statistical Learning with Applications in Python. Springer Texts in Statistics. Springer Cham, 2023.