Řízená klasifikace síťového provozu

Patrik Urban[1]
Školitel: Michal Sofka[2], Ph.D.

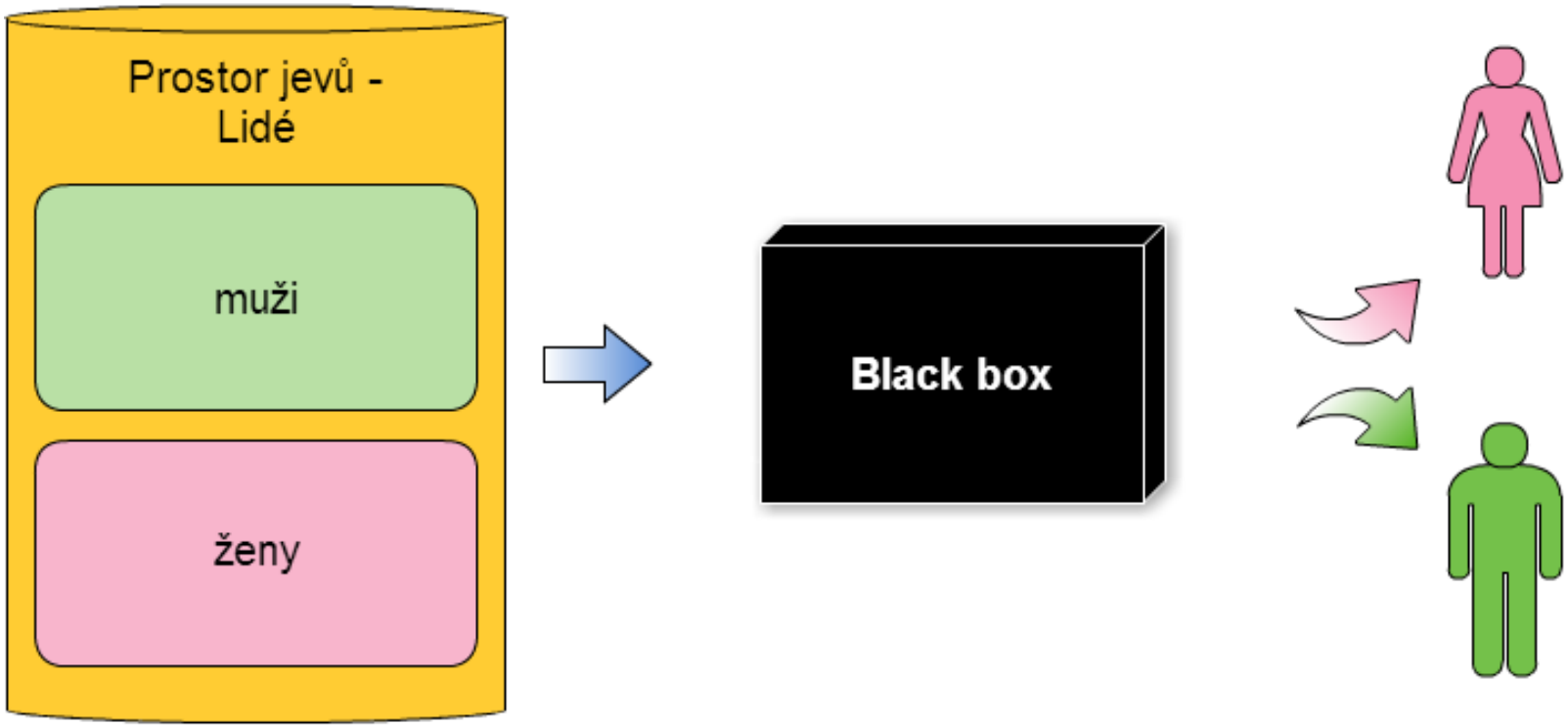
[1]České vysoké učení technické v Praze

[2]Cisco Systems, Inc.

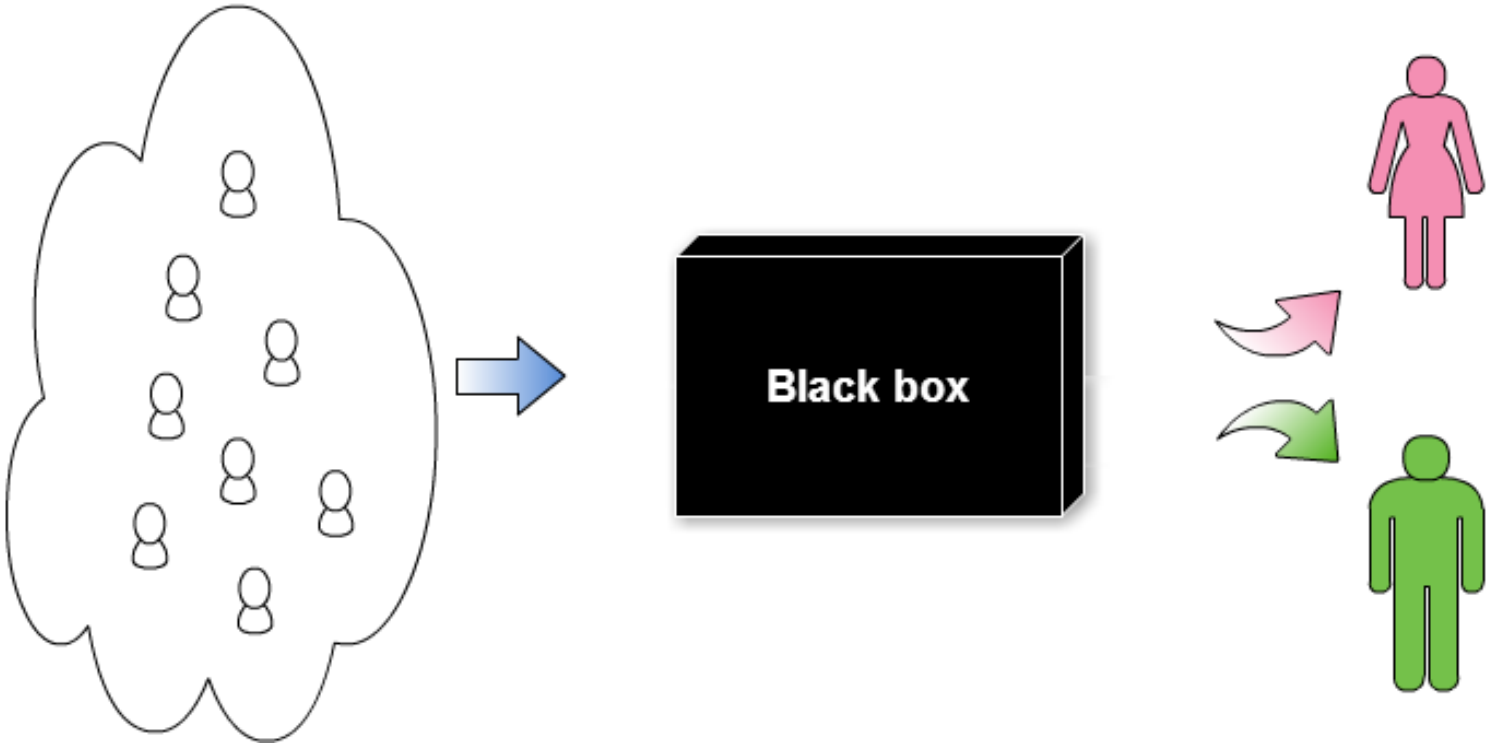
24. srpna 2015

1. Co je to klasifikace
   - Formulace úlohy
   - Přesná matematická formulace
   - Discriminant functions
   - Lineární diskriminátory
2. Motivace práce
   - Cisco, proxy server
   - Klasifikace síťového provozu
3. Repetitivnost
4. Metody klasifikace
   - Naivní klasifikace
   - Single log klasifikace
   - IP klasifikace
   - Kvantifikace výsledků klasifikace
5. Výsledky klasifikace a porovnání metod
6. Robustní metoda klasifikace multimediálního streamingu

Úloha klasifikace



Úloha klasifikace



Matematická formulace:

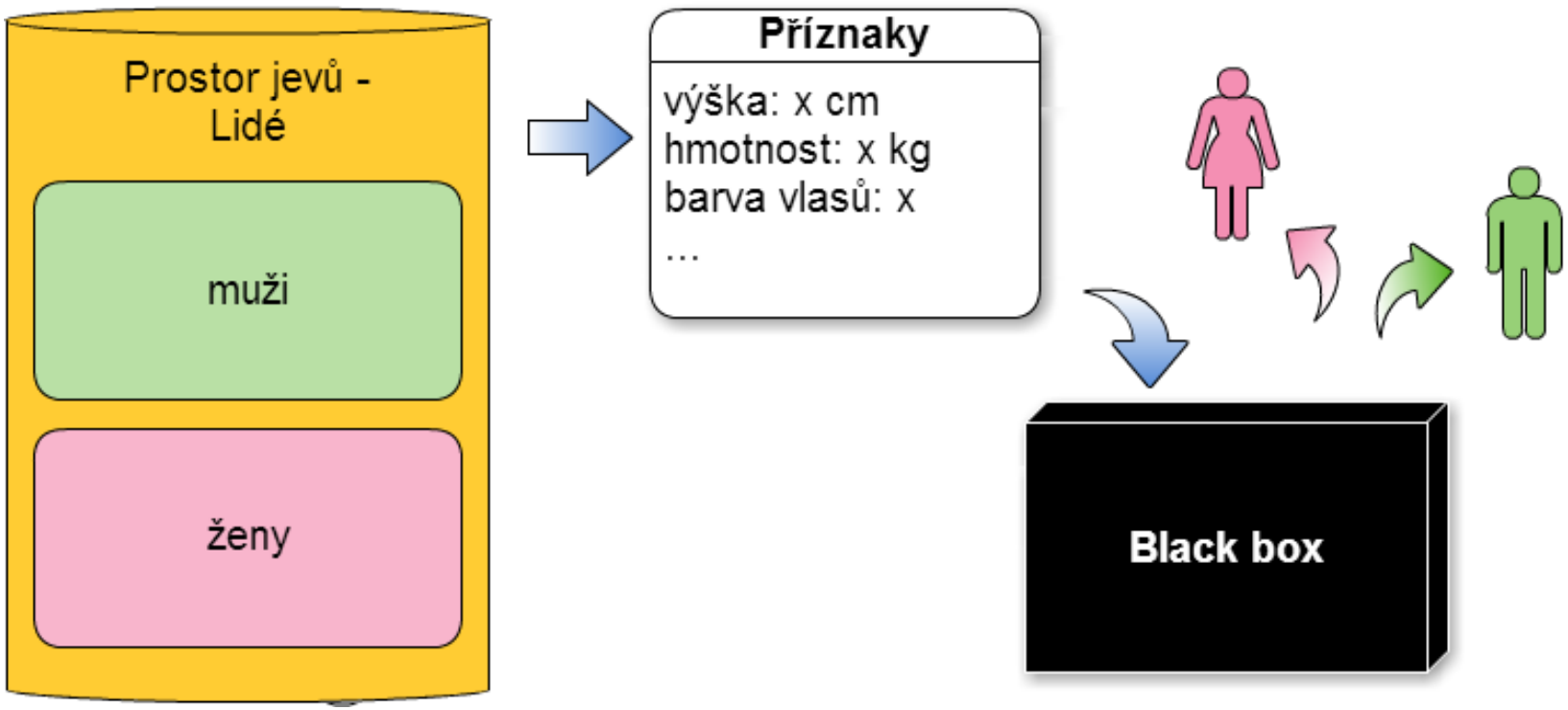
- X prostor jevů
- $\mathcal{G}$ konečná množina kategorií/tříd ($\mathcal{G} = \{"\text{muž}", "\text{žena}"\}$)
- Klasifikátor je zobrazení $f : X \to \mathcal{G} = (\mathcal{G}_k)_{k=1}^{K}$

Matematická formulace:

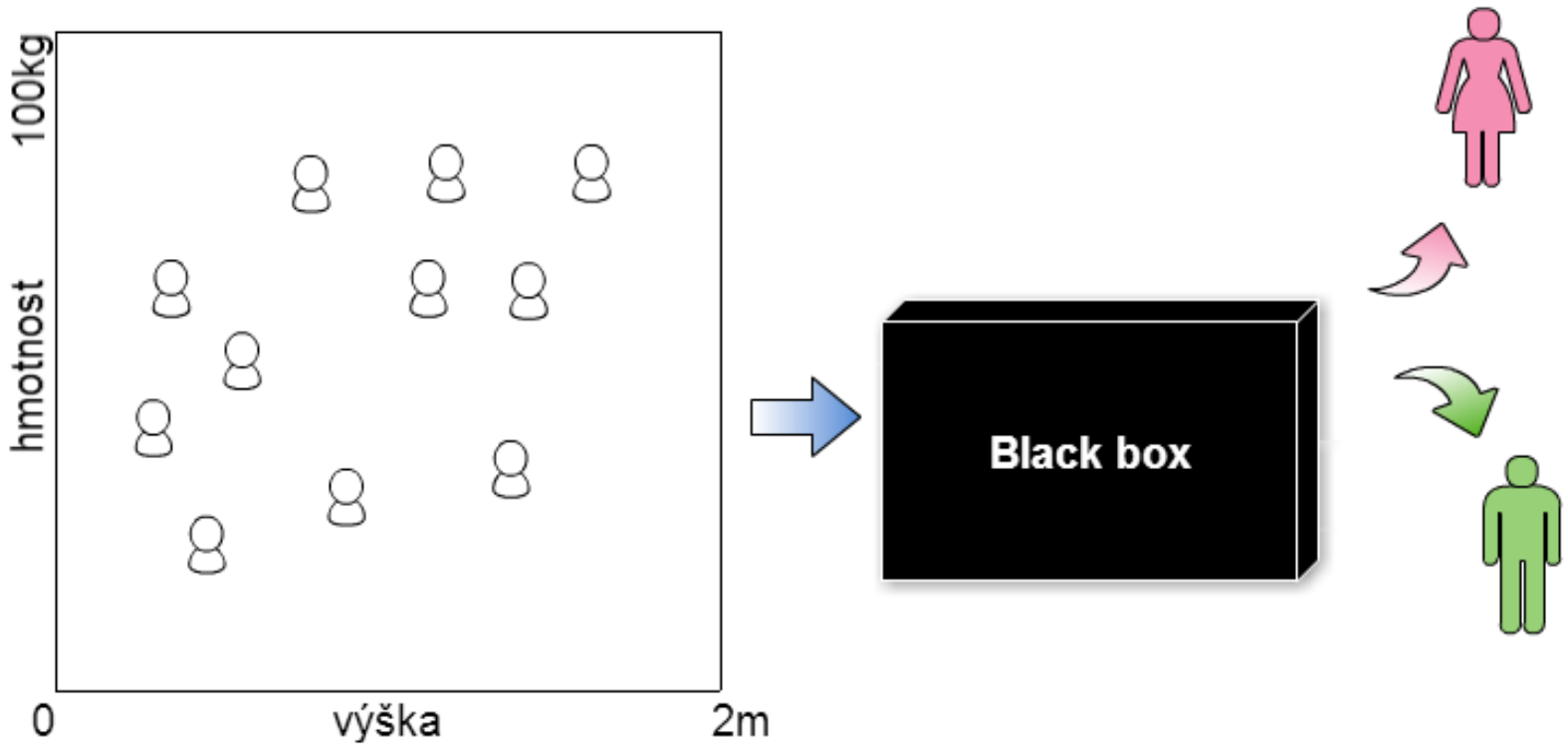
- X prostor jevů
- $\mathcal{G}$ konečná množina kategorií/tříd ($\mathcal{G} = \{"muž", "žena"\}$)
- Klasifikátor je zobrazení $f : X \to \mathcal{G} = (\mathcal{G}_k)_{k=1}^{K}$

- $\tilde{X}$ příznakový prostor (například $\mathbb{R}^n$)
- Jevy $\omega \in X$ zobrazím do $\tilde{X}$, kde mohu požadovat hezké vlastnosti (metrika, lineární prostor....)
- Klasifikátor je zobrazení $\tilde{f} : \tilde{X} \to \mathcal{G} = (\mathcal{G}_k)_{k=1}^{K}$

Úloha klasifikace



Úloha klasifikace



- Označíme $(x_1, \ldots, x_N) =: \mathbf{x} \in \tilde{X}$ příznak (**input**)

- Označíme $(x_1, \ldots, x_N) =: \mathbf{x} \in \tilde{X}$ příznak (**input**)
- Označíme $\mathcal{G}_k \in \mathcal{G}$ kategorii, $k \in \{1, \ldots, K\}$, K počet kategorií

- Označíme $(x_1, \ldots, x_N) =: \mathbf{x} \in \tilde{X}$ příznak (**input**)
- Označíme $\mathcal{G}_k \in \mathcal{G}$ kategorii, $k \in \{1, \ldots, K\}$, K počet kategorií
- Kategorie "přejmenuji" na čísla: $y_k \sim \mathcal{G}_k$ (**output**)

- Označíme $(x_1, \ldots, x_N) =: \mathbf{x} \in \tilde{X}$ příznak (**input**)
- Označíme $\mathcal{G}_k \in \mathcal{G}$ kategorii, $k \in \{1, \ldots, K\}$, K počet kategorií
- Kategorie "přejmenuji" na čísla: $y_k \sim \mathcal{G}_k$ (**output**)
- Uspořádanou dvojici **input-output** označím $(\mathbf{x}, y_k)$

- Označíme $(x_1, \ldots, x_N) =: \mathbf{x} \in \tilde{X}$ příznak (**input**)
- Označíme $\mathcal{G}_k \in \mathcal{G}$ kategorii, $k \in \{1, \ldots, K\}$, K počet kategorií
- Kategorie "přejmenuji" na čísla: $y_k \sim \mathcal{G}_k$ (**output**)
- Uspořádanou dvojici **input-output** označím $(\mathbf{x}, y_k)$
- Supervised learning: máme M dvojic, u kterých známe výsledky - **trénovací množina**$=\{(\mathbf{x}_i, y_{k_i})\}_{i=1}^{M}$

- Označíme $(x_1, \ldots, x_N) =: \mathbf{x} \in \tilde{X}$ příznak (**input**)
- Označíme $\mathcal{G}_k \in \mathcal{G}$ kategorii, $k \in \{1, \ldots, K\}$, K počet kategorií
- Kategorie "přejmenuji" na čísla: $y_k \sim \mathcal{G}_k$ (**output**)
- Uspořádanou dvojici **input-output** označím $(\mathbf{x}, y_k)$
- Supervised learning: máme M dvojic, u kterých známe výsledky - **trénovací množina**$=\{(\mathbf{x}_i, y_{k_i})\}_{i=1}^{M}$
- Zobrazení $\mathbf{x} \rightarrow y_k$ rozdělí příznakový prostor na K oblastí - **decision regions**, které jsou rozděleny hranicemi - **decision boundaries**

- Označíme $(x_1, \ldots, x_N) =: \mathbf{x} \in \tilde{X}$ příznak (**input**)
- Označíme $\mathcal{G}_k \in \mathcal{G}$ kategorii, $k \in \{1, \ldots, K\}$, K počet kategorií
- Kategorie "přejmenuji" na čísla: $y_k \sim \mathcal{G}_k$ (**output**)
- Uspořádanou dvojici **input-output** označím $(\mathbf{x}, y_k)$
- Supervised learning: máme M dvojic, u kterých známe výsledky - **trénovací množina**$=\{(\mathbf{x}_i, y_{k_i})\}_{i=1}^{M}$
- Zobrazení $\mathbf{x} \rightarrow y_k$ rozdělí příznakový prostor na K oblastí - **decision regions**, které jsou rozděleny hranicemi - **decision boundaries**
- Uvažujme $\tilde{X} \times (\mathcal{G}_k)_{k=1}^{K}$ jako pravděpodobnostní prostor se sdruženou pravděpodobnostní hustotou $p(\mathbf{x}, \mathcal{G}_k) \equiv p(\mathbf{x}, y_k)$ splňující

$$1 = \int_{\tilde{X}} \sum_k p(\mathbf{x}, \mathcal{G}_k) d\mathbf{x}$$

- Máme tedy $\tilde{X}$ rozdělený na K oblastí $\mathcal{R}_k$ pro které platí, že $\forall(\mathbf{x}, y) : \mathbf{x} \in \mathcal{R}_k \Rightarrow y \in \mathcal{G}_k$

- Máme tedy $\tilde{X}$ rozdělený na K oblastí $\mathcal{R}_k$ pro které platí, že $\forall(\mathbf{x}, y) : \mathbf{x} \in \mathcal{R}_k \Rightarrow y \in \mathcal{G}_k$
- Úloha klasifikace lze formulovat jako minimalizační úloha - chci minimalizovat chybu ($K = 2$):

$$\begin{aligned} p(\text{chyby}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{G}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{G}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{G}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{G}_1) d\mathbf{x} \end{aligned}$$

- Máme tedy $\tilde{X}$ rozdělený na K oblastí $\mathcal{R}_k$ pro které platí, že $\forall(\mathbf{x}, y) : \mathbf{x} \in \mathcal{R}_k \Rightarrow y \in \mathcal{G}_k$
- Úloha klasifikace lze formulovat jako minimalizační úloha - chci minimalizovat chybu ($K = 2$):

$$\begin{aligned} p(\text{chyby}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{G}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{G}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{G}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{G}_1) d\mathbf{x} \end{aligned}$$

- Tedy: Jestliže $p(\mathbf{x}, \mathcal{G}_1) > p(\mathbf{x}, \mathcal{G}_2)$ pak $\mathbf{x}$ je kategorie $\mathcal{G}_1$

- Máme tedy $\tilde{X}$ rozdělený na K oblastí $\mathcal{R}_k$ pro které platí, že $\forall(\mathbf{x}, y) : \mathbf{x} \in \mathcal{R}_k \Rightarrow y \in \mathcal{G}_k$
- Úloha klasifikace lze formulovat jako minimalizační úloha - chci minimalizovat chybu ($K = 2$):

$$\begin{aligned} p(\text{chyby}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{G}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{G}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{G}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{G}_1) d\mathbf{x} \end{aligned}$$

- Tedy: Jestliže $p(\mathbf{x}, \mathcal{G}_1) > p(\mathbf{x}, \mathcal{G}_2)$ pak $\mathbf{x}$ je kategorie $\mathcal{G}_1$
- Z řetězového pravidla plyne:

$$p(\mathcal{G}_1|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}, \mathcal{G}_1) > p(\mathbf{x}, \mathcal{G}_2) = p(\mathcal{G}_2|\mathbf{x})p(\mathbf{x})$$

$$p(\mathcal{G}_1|\mathbf{x}) > p(\mathcal{G}_2|\mathbf{x})$$

- Bayesův teorém

$$p(\mathcal{G}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{G}_k)p(\mathcal{G}_k)}{p(\mathbf{x})}$$

- Bayesův teorém

$$p(\mathcal{G}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{G}_k)p(\mathcal{G}_k)}{p(\mathbf{x})}$$

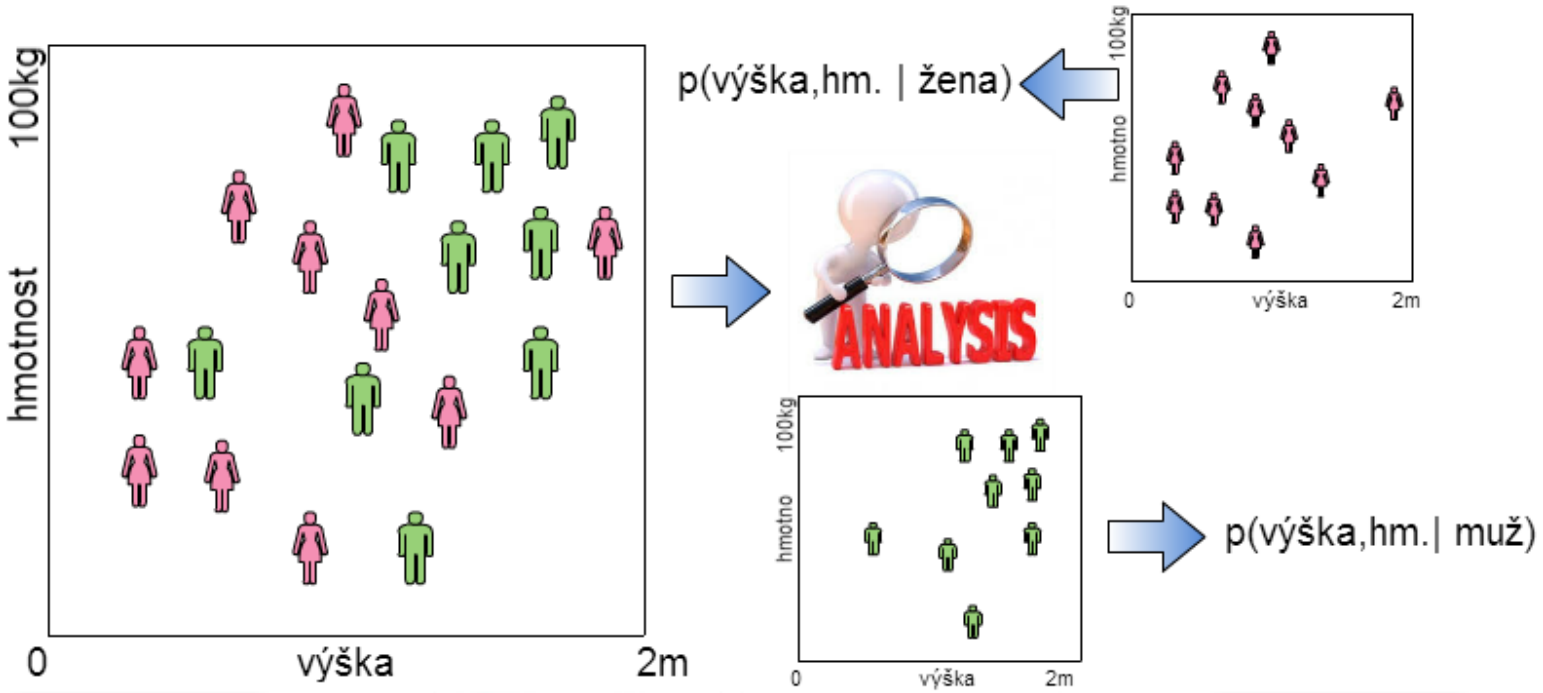
- Úloha je tedy ekvivalentní úloze nalezení distribucí podmíněných pravděpodobností $p(\mathbf{x}|\mathcal{G}_k)$ - získáme z trénovací množiny

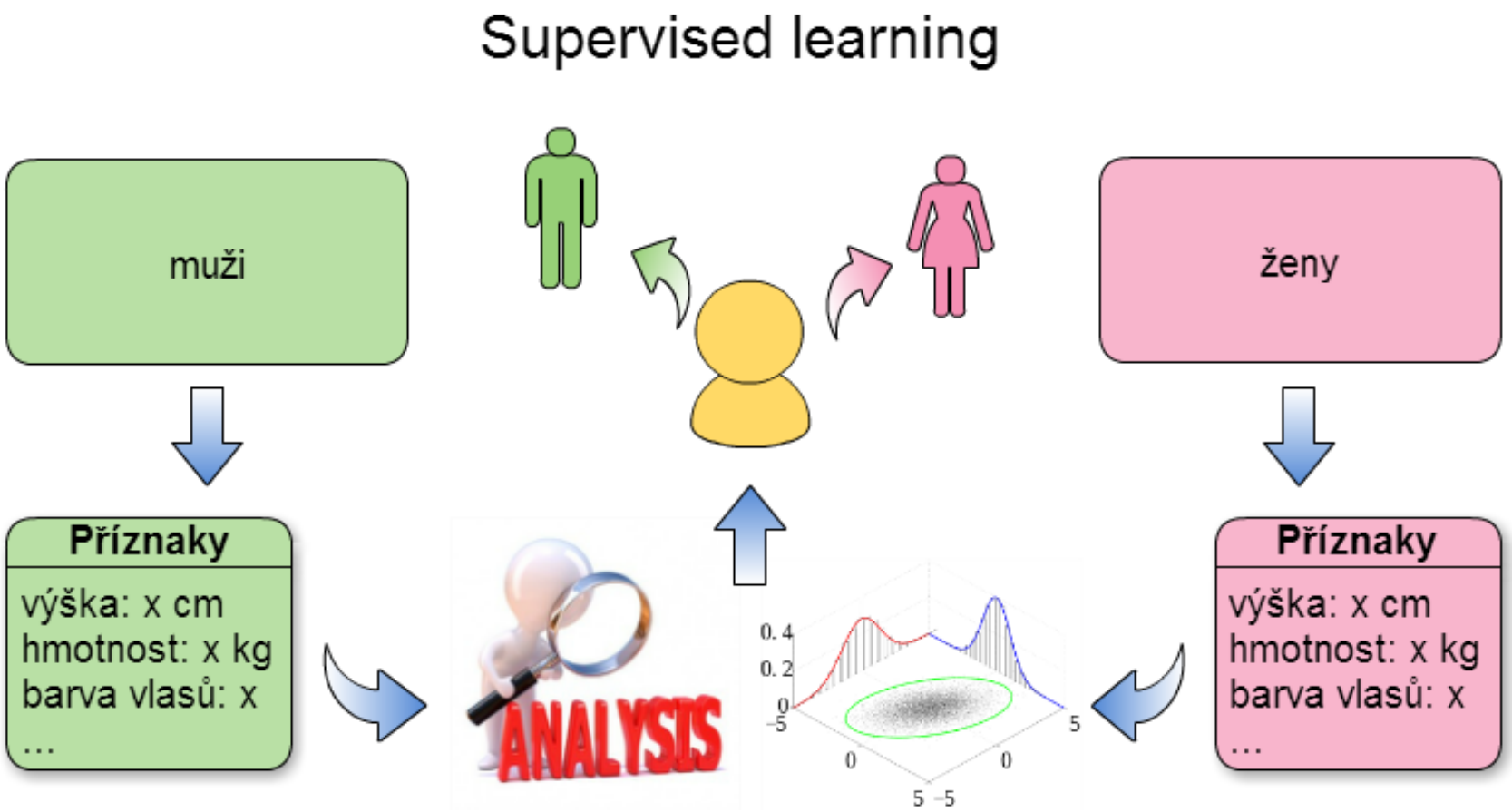
- Bayesův teorém

$$p(\mathcal{G}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{G}_k)p(\mathcal{G}_k)}{p(\mathbf{x})}$$

- Úloha je tedy ekvivalentní úloze nalezení distribucí podmíněných pravděpodobností $p(\mathbf{x}|\mathcal{G}_k)$ - získáme z trénovací množiny
- Důležitost příznaků a trénovací množiny

Úloha klasifikace




Supervised learning
muži
ženy
Příznaky
výška: x cm
hmotnost: x kg
barva vlasů: x
...
ANALYSIS
Příznaky
výška: x cm
hmotnost: x kg
barva vlasů: x
...

- Předpokádejme existenci diskriminujících funkcí:

$$(\forall \mathcal{G}_k)(\exists \delta_k) : \Big(\big(\delta_k(\mathbf{x}) > \delta_j(\mathbf{x})\big)(\forall j \neq k) \Rightarrow \mathbf{x} \text{ je třídy } \mathcal{G}_k \Big)$$

- Předpokádejme existenci diskriminujících funkcí:

$$(\forall \mathcal{G}_k)(\exists \delta_k) : \Big(\big(\delta_k(\mathbf{x}) > \delta_j(\mathbf{x})\big)(\forall j \neq k) \Rightarrow \mathbf{x} \text{ je třídy } \mathcal{G}_k \Big)$$

- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k = \{\mathbf{x} | \delta_k(\mathbf{x}) = \delta_j(\mathbf{x})\}$

- Předpokádejme existenci diskriminujících funkcí:

$$(\forall \mathcal{G}_k)(\exists \delta_k) : \Big(\big(\delta_k(\mathbf{x}) > \delta_j(\mathbf{x})\big)(\forall j \neq k) \Rightarrow \mathbf{x} \text{ je třídy } \mathcal{G}_k \Big)$$

- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k = \{\mathbf{x} | \delta_k(\mathbf{x}) = \delta_j(\mathbf{x})\}$
- Lze uvažovat speciální případy takovýchto funkcí:

- Předpokádejme existenci diskriminujících funkcí:

$$(\forall \mathcal{G}_k)(\exists \delta_k) : \Big(\big(\delta_k(\mathbf{x}) > \delta_j(\mathbf{x})\big)(\forall j \neq k) \Rightarrow \mathbf{x} \text{ je třídy } \mathcal{G}_k \Big)$$

- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k = \{\mathbf{x} | \delta_k(\mathbf{x}) = \delta_j(\mathbf{x})\}$
- Lze uvažovat speciální případy takovýchto funkcí:
  - $p(\mathbf{x}|\mathcal{G}_k)$ sama o sobě

- Předpokádejme existenci diskriminujících funkcí:

$$(\forall \mathcal{G}_k)(\exists \delta_k) : \Big(\big(\delta_k(\mathbf{x}) > \delta_j(\mathbf{x})\big)(\forall j \neq k) \Rightarrow \mathbf{x} \text{ je třídy } \mathcal{G}_k \Big)$$

- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k = \{\mathbf{x} | \delta_k(\mathbf{x}) = \delta_j(\mathbf{x})\}$
- Lze uvažovat speciální případy takovýchto funkcí:
  - $p(\mathbf{x}|\mathcal{G}_k)$ sama o sobě
  - Lineární diskriminanty: $\delta_k(\mathbf{x}) = \mathbf{w}_k^T \cdot \mathbf{x} + w_{k0}$

- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k$ je afinní prostor (nadrovina)

$$\{\mathbf{x} \in X | (\mathbf{w}_k - \mathbf{w}_j) \cdot \mathbf{x} + (w_{k0} - w_{j0}) = 0\}$$

- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k$ je afinní prostor (nadrovina)

$$\{\mathbf{x} \in X | (\mathbf{w}_k - \mathbf{w}_j) \cdot \mathbf{x} + (w_{k0} - w_{j0}) = 0\}$$

- Uvažujme monotónní transformaci příznakového prostoru $\mathbf{\Phi}$ (existuje inverze)

- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k$ je afinní prostor (nadrovina)

$$\{\mathbf{x} \in X | (\mathbf{w}_k - \mathbf{w}_j) \cdot \mathbf{x} + (w_{k0} - w_{j0}) = 0\}$$

- Uvažujme monotónní transformaci příznakového prostoru $\mathbf{\Phi}$ (existuje inverze)
- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k$ je opět afinní prostor (nadrovina)

$$\{\mathbf{x} \in X | \mathbf{\Phi}(\delta_k(x)) = \mathbf{\Phi}(\delta_j(x))\} = \{\mathbf{x} \in X | \delta_k(x) = \delta_j(x)\}$$

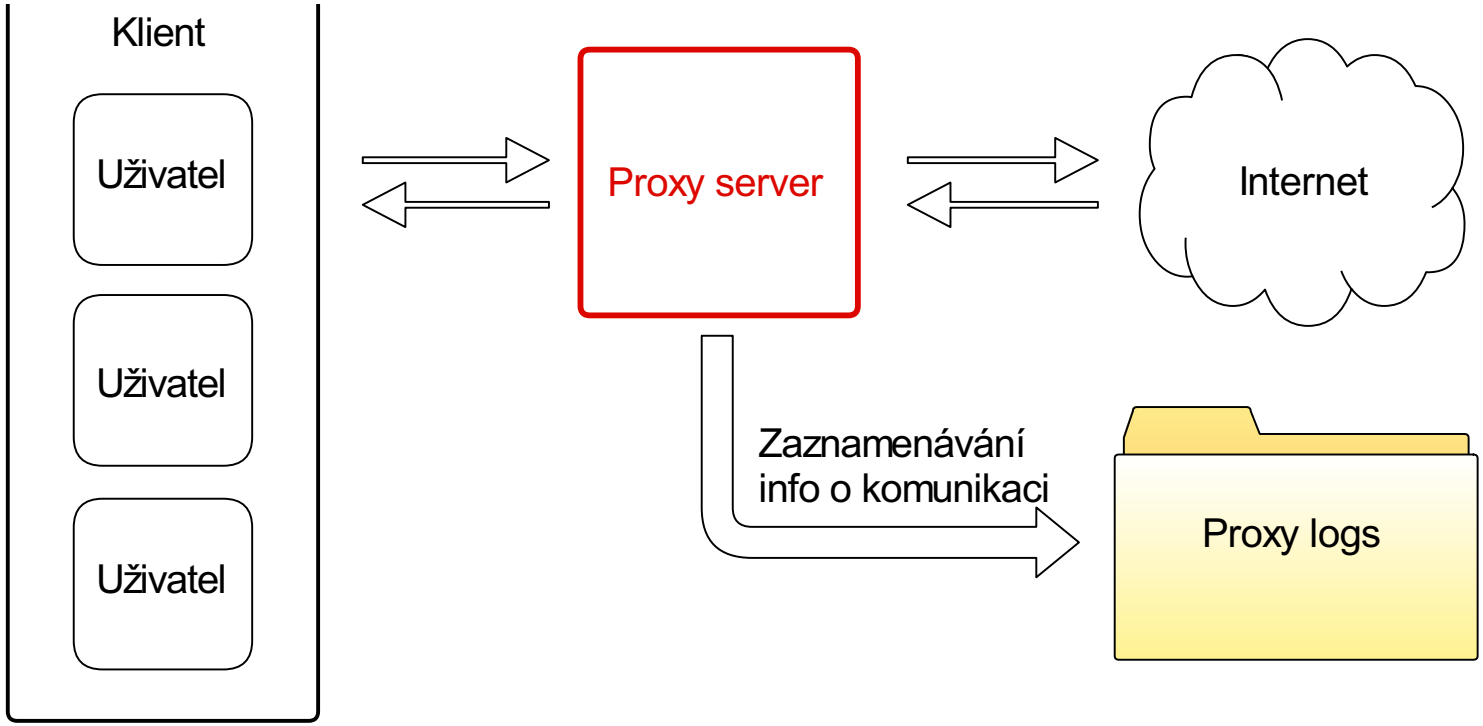
- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k$ je afinní prostor (nadrovina)

$$\{\mathbf{x} \in X | (\mathbf{w}_k - \mathbf{w}_j) \cdot \mathbf{x} + (w_{k0} - w_{j0}) = 0\}$$

- Uvažujme monotónní transformaci příznakového prostoru $\mathbf{\Phi}$ (existuje inverze)
- Decision boundary mezi $\mathcal{R}_j$ a $\mathcal{R}_k$ je opět afinní prostor (nadrovina)

$$\{\mathbf{x} \in X | \mathbf{\Phi}(\delta_k(x)) = \mathbf{\Phi}(\delta_j(x))\} = \{\mathbf{x} \in X | \delta_k(x) = \delta_j(x)\}$$

- Lineární diskriminant - funkce, které je po nějaké monotónní transformaci lineární


Klient
Uživatel
Uživatel
Uživatel
Proxy server
Internet
Zaznamenávání
info o komunikaci
Proxy logs

Tabulka: Příkad proxy logu

timestamp (ms)	1382058540001
http status	200
sc Bytes	180102
cs Bytes	0
url	http://r13—sn-aigeznez.c.youtube.com/videoplayback?...
user ID	8b16fefbe5877034de6707a3751c5c92
elapsed time	92
server IP	173.194.5.242
client IP	důvěrné
RS content type	audio/mp4
referer	http://www.youtube.com/watch?v=Iw8OyE3_nv4
user agent	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 ...

Záznamů je hodně (1 milion logů za 10 minut) $\Rightarrow$ není možné projít a zkontrolovat vše $\Rightarrow$

- Analýza proxy logů
- Rozpoznávání komunikace
- Klasifikace na:
  - Bezpečné
  - Nebezpečné/nechtěné
  - Neznámé
- Audio streaming: internetová rádia, Soundcloud.com, Spotify.com
- Video streaming: Youtube.com, Stream.cz, Netflix.com

Tabulka: Příkad proxy logu

timestamp (ms)	1382096460175
http status	0
sc Bytes	1075
cs Bytes	2408
url	https://31.13.66.49/
user ID	06a3c2f922cbf5f00213298723be88b0
elapsed time	63548
server IP	31.13.66.49
client IP	důvěrné
RS content type	[]
referer	[]
user agent	[]

Tabulka: Ukázka repetitivního přenostu 1 (Youtube)

sc Bytes	cs Bytes	url
1781747	0	http://r17—kul06s01.c.youtube.com/...
1781760	0	http://r17—kul06s01.c.youtube.com/...
1781760	0	http://r17—kul06s01.c.youtube.com/...
1781760	0	http://r17—kul06s01.c.youtube.com/...
1781760	0	http://r17—kul06s01.c.youtube.com/...
1781760	0	http://r17—kul06s01.c.youtube.com/...
1781760	0	http://r17—kul06s01.c.youtube.com/...
1781760	0	http://r17—kul06s01.c.youtube.com/...
1781760	0	http://r17—kul06s01.c.youtube.com/...
1781760	0	http://r17—kul06s01.c.youtube.com/...

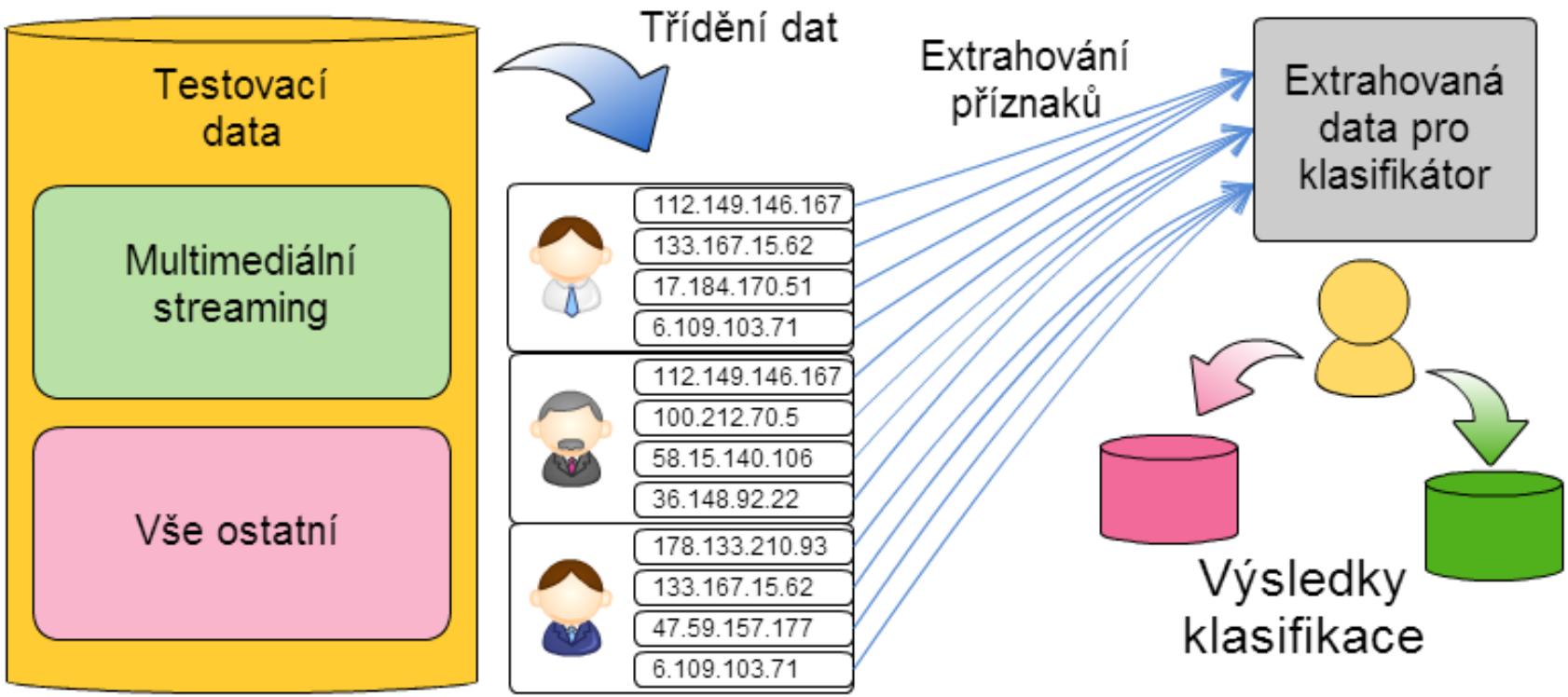
Tabulka: Ukázka repetitivního přenostu 2 (Youtube)

sc Bytes	cs Bytes	url
237568	0	http://r7—sn-npo7enee.c.youtube.com/...
1134592	0	http://r7—sn-npo7enee.c.youtube.com/...
237568	0	http://r7—sn-npo7enee.c.youtube.com/...
237568	0	http://r7—sn-npo7enee.c.youtube.com/...
1134592	0	http://r7—sn-npo7enee.c.youtube.com/...
1134592	0	http://r7—sn-npo7enee.c.youtube.com/...
237568	0	http://r7—sn-npo7enee.c.youtube.com/...
1138688	0	http://r1—sn-npo7en7r.c.youtube.com/...
237568	0	http://r1—sn-npo7en7r.c.youtube.com/...
237568	0	http://r7—sn-npo7enee.c.youtube.com/...

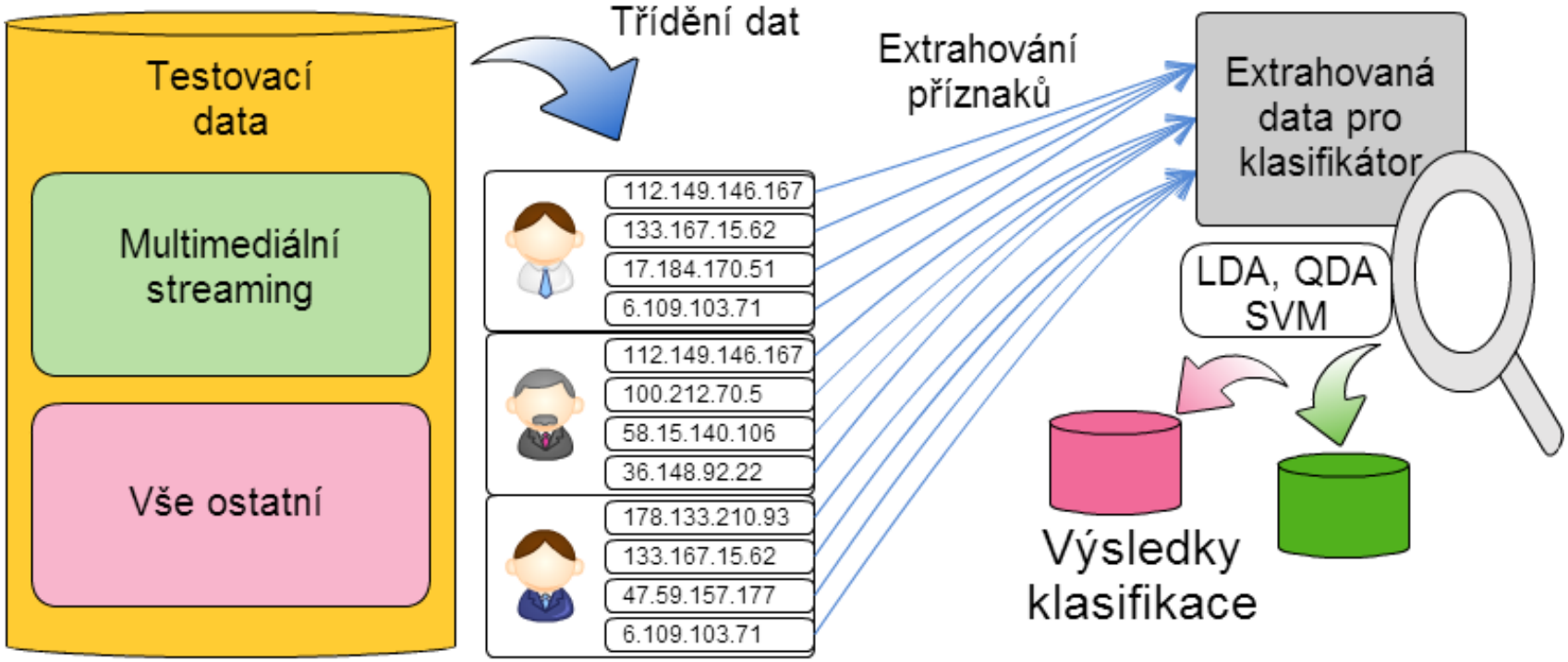
Tabulka: Ukázka repetitivního přenostu 3 (Netflix)

sc Bytes	cs Bytes	url
424495	0	http://198.45.55.163/300734852.ismv/...
441814	0	http://198.45.55.163/300734852.ismv/...
444283	0	http://198.45.55.163/300734852.ismv/...
452532	0	http://198.45.55.163/300734852.ismv/...
389477	0	http://198.45.55.163/300734852.ismv/...
432231	0	http://198.45.55.163/300734852.ismv/...
455956	0	http://198.45.55.163/300734852.ismv/...
544827	0	http://198.45.55.163/300734852.ismv/...
498620	0	http://198.45.55.163/300734852.ismv/...
509960	0	http://198.45.55.163/300734852.ismv/...

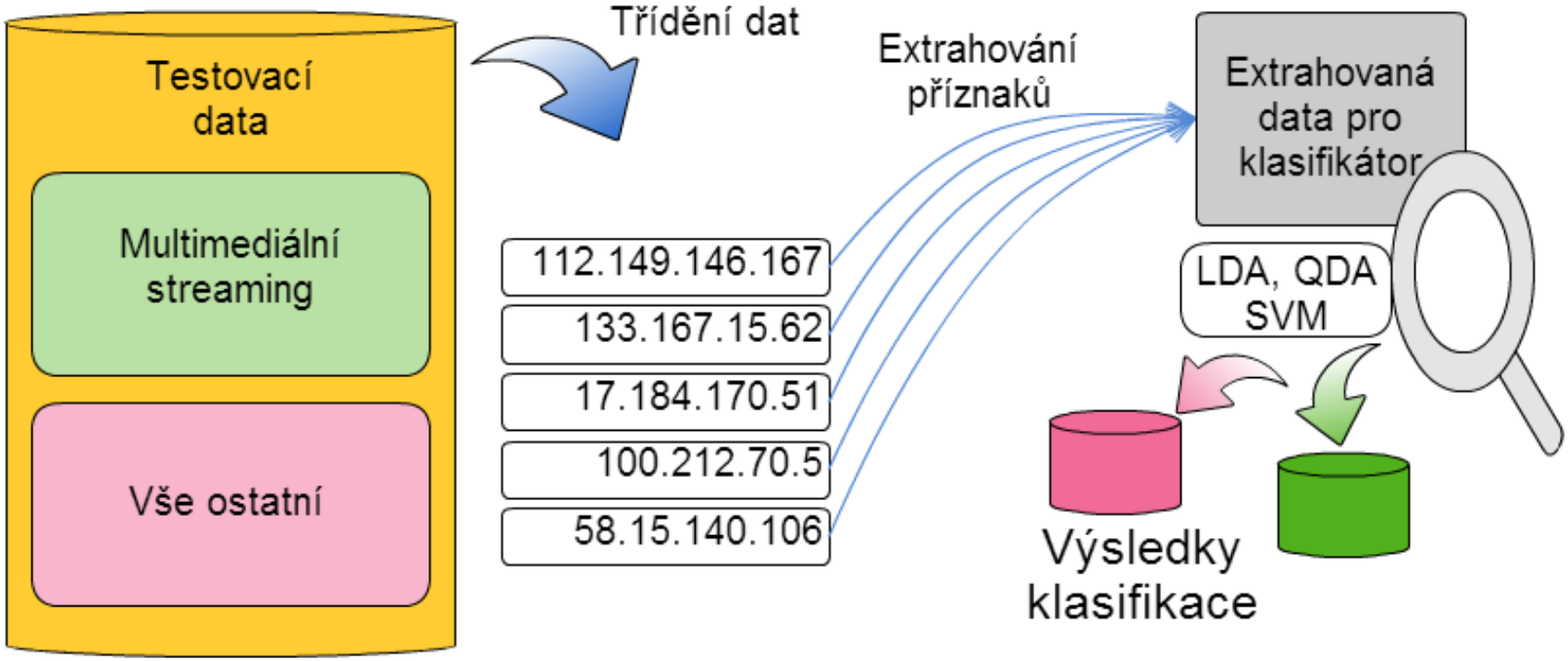
Naivní klasifikace



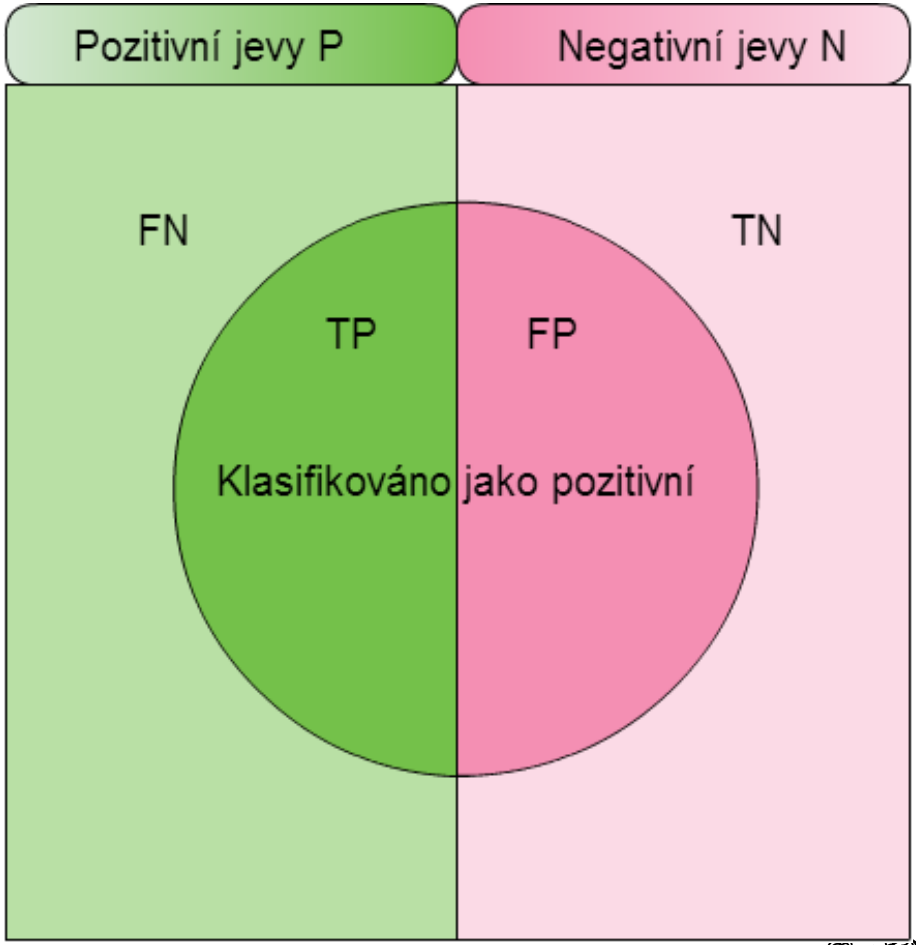
Single log klasifikace



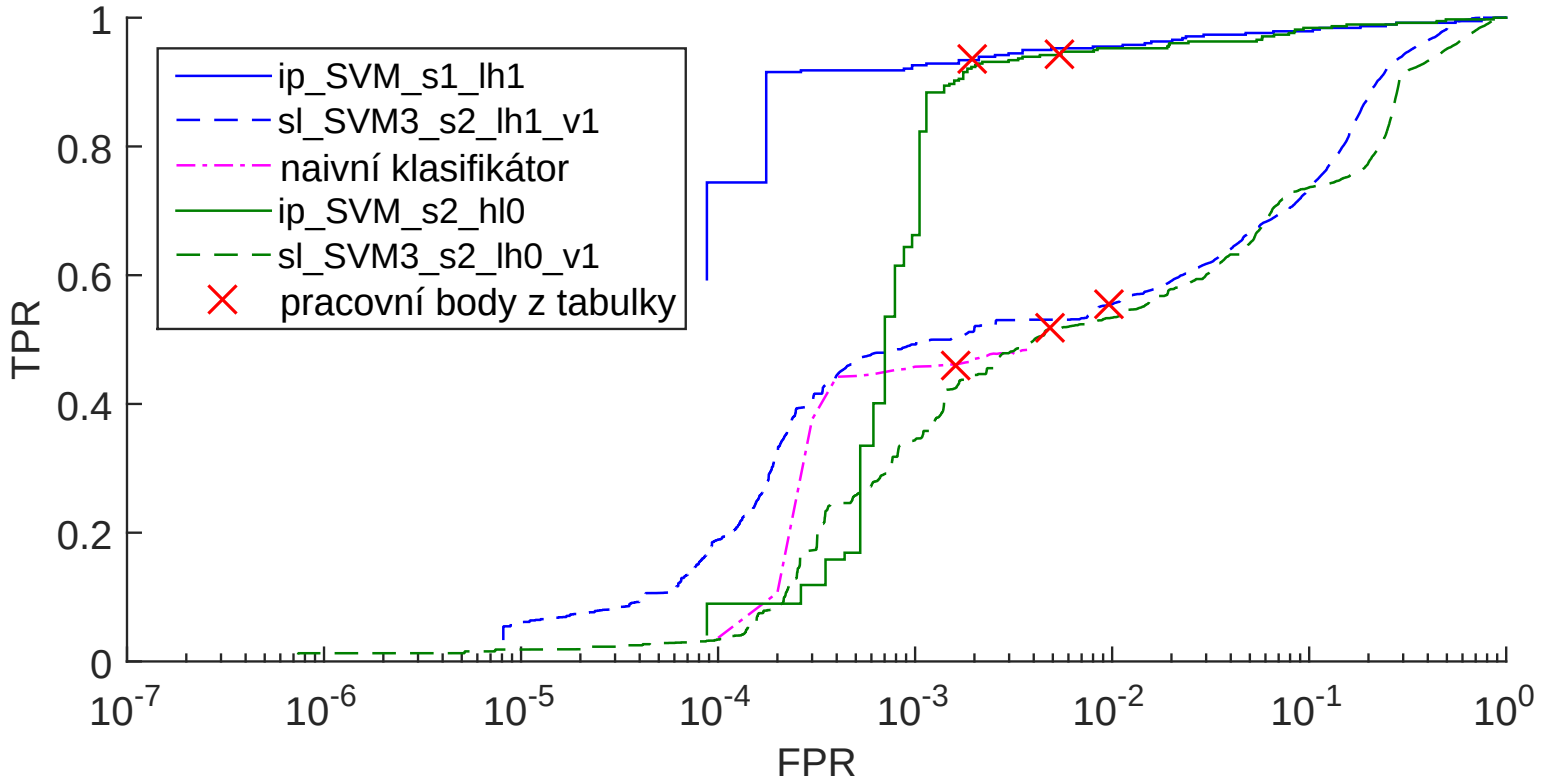
IP klasifikace



- TP, TN - správně klasifikované jevy
- FP - chyba I. druhu
- FN - chyba II. druhu
- Recall = TPR = $\frac{TP}{P}$
- FPR = $\frac{FP}{F}$
- Precision = $\frac{TP}{TP+FP}$
- ROC křivka: TPR = TPR(FPR)



Obrázek: ROC křivky 2 nejlepších klasifikátorů reprezentující každou metodu s vyznačenými pracovními body z tabulky



Tabulka: Statistické výsledky klasifikace a porovnání nejlepších klasifikátorů reprezentující každou z metod klasifikace

Classifier name	Classified as *media*	TP	FP	TPR	FPR	Precision	% of Youtube media found
ip_SVM_s1_lh1	376	354	22	93.65%	0.19%	94.15%	
sl_SVM3_s2_lh1_v1	37059	23794	13265	55.41%	0.98%	64.21%	97.18%
Naivní klasifikátor	21229	20547	682	44,22%	0,04%	96,79%	96,18%
ip_SVM_s2_lh0	419	357	62	94.18%	0.54%	85.20%	
sl_SVM3_s2_lh0_v1	28799	22189	6610	51.68%	0.49%	77.05%	91.65%

- Uspěšnější než naivní klasifikátor

- Uspěšnější než naivní klasifikátor
- Použitelná i v šifrovaném `https` provozu

- Uspěšnější než naivní klasifikátor
- Použitelná i v šifrovaném `https` provozu
- Byly nalezeny záznamy, které se v původních trénovacích množinách nevyskytovaly (po manualní analýze byly některé dokonce přidány mezi multimediální streaming)

- Uspěšnější než naivní klasifikátor
- Použitelná i v šifrovaném `https` provozu
- Byly nalezeny záznamy, které se v původních trénovacích množinách nevyskytovaly (po manualní analýze byly některé dokonce přidány mezi multimediální streaming)
- Možná vylepšení a rozšíření:
  - Pracovat s větším objemem dat

- Uspěšnější než naivní klasifikátor
- Použitelná i v šifrovaném `https` provozu
- Byly nalezeny záznamy, které se v původních trénovacích množinách nevyskytovaly (po manualní analýze byly některé dokonce přidány mezi multimediální streaming)
- Možná vylepšení a rozšíření:
  - Pracovat s větším objemem dat
  - Dále analyzovat chyby klasifikace - Jsou to opravdu chyby? Jak je blíže popsat?

- Uspěšnější než naivní klasifikátor
- Použitelná i v šifrovaném `https` provozu
- Byly nalezeny záznamy, které se v původních trénovacích množinách nevyskytovaly (po manualní analýze byly některé dokonce přidány mezi multimediální streaming)
- Možná vylepšení a rozšíření:
  - Pracovat s větším objemem dat
  - Dále analyzovat chyby klasifikace - Jsou to opravdu chyby? Jak je blíže popsat?
  - Zkusit pomocí repetitivnosti popsat jiný provoz - malware?

Děkuji za pozornost