Interpretovateľnosť modelov neurónových sietí

Branislav Pecher

Vedúci projektu: Ing. Jakub Ševcech PhD.

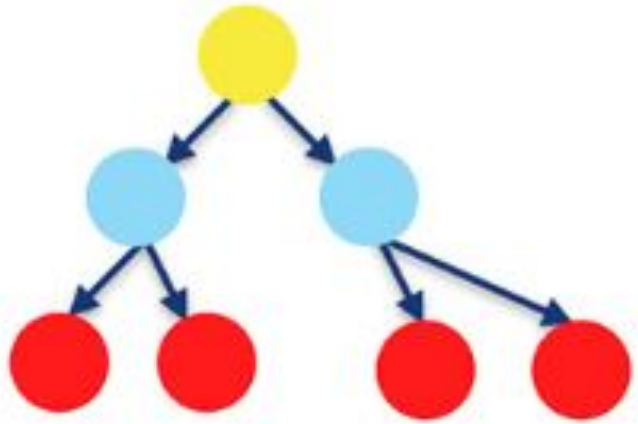
Typicky používané prístupy

- Extrakcia pravidiel
- Vizualizácia
- Analýza citlivosti (sensitivity analysis)

Extrakcia pravidiel

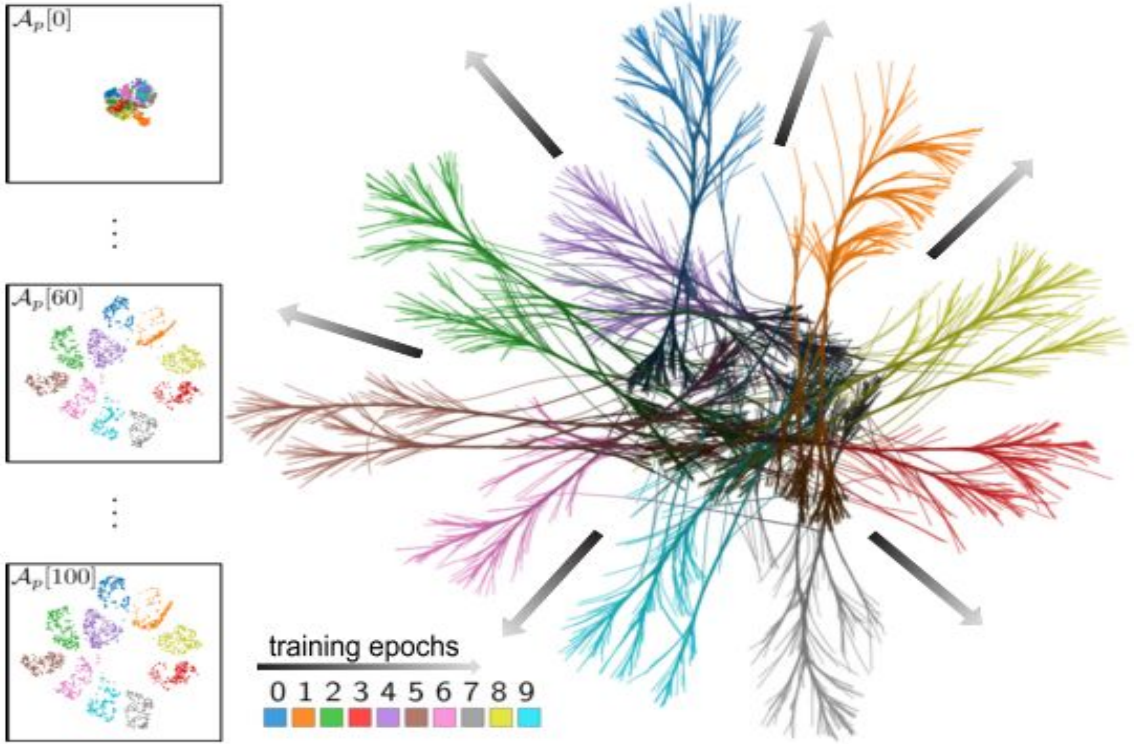
- Generovanie jednoduchých pravidiel
- Pre jednotlivé neuróny, následná agregácia
- Pre jednotlivé vrstvy, využitie predchádzajúcej vrstvy
- Pravidlá mapovania vstupu na výstup - učenie rozhodovacieho stromu

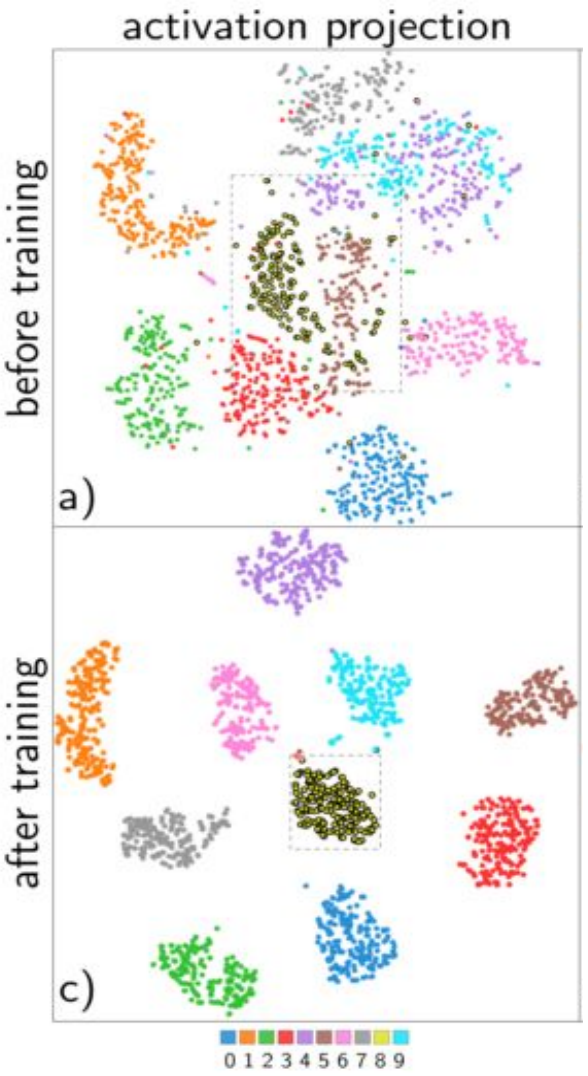
if male **and** adult **then** *survival probability* 21% (19% - 23%)
else if 3rd class **then** *survival probability* 44% (38% - 51%)
else if 1st class **then** *survival probability* 96% (92% - 99%)
else *survival probability* 88% (82% - 94%)



Vizualizácia NN

- Vizualizácia aktivácií a učenia (posledná vrstva)





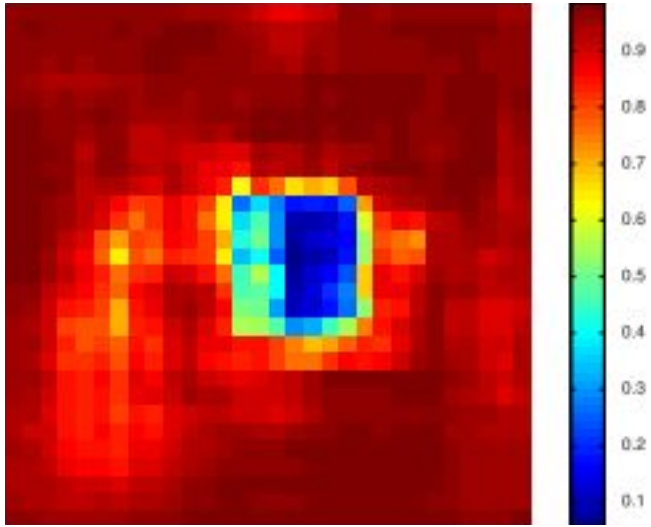
Analýza citlivosti (sensitivy analysis)

- Ako citlivý je výstup na zmenu vstupu

- Zistenie dôležitosti atribútu
  - Pre jedno pozorovanie
  - Pre model celkovo
- ‘Odstránenie’ atribútu
- Pozorovanie zmeny chyby
- Pozorovanie zmeny pravdepodobnosti pre správnu triedu

Analýza citlivosti (sensitivity analysis) - CNN

- Postupné zakrývanie obrázku
- Sledovanie pravdepodobnosti pre správnu triedu



Čo plánujeme robiť

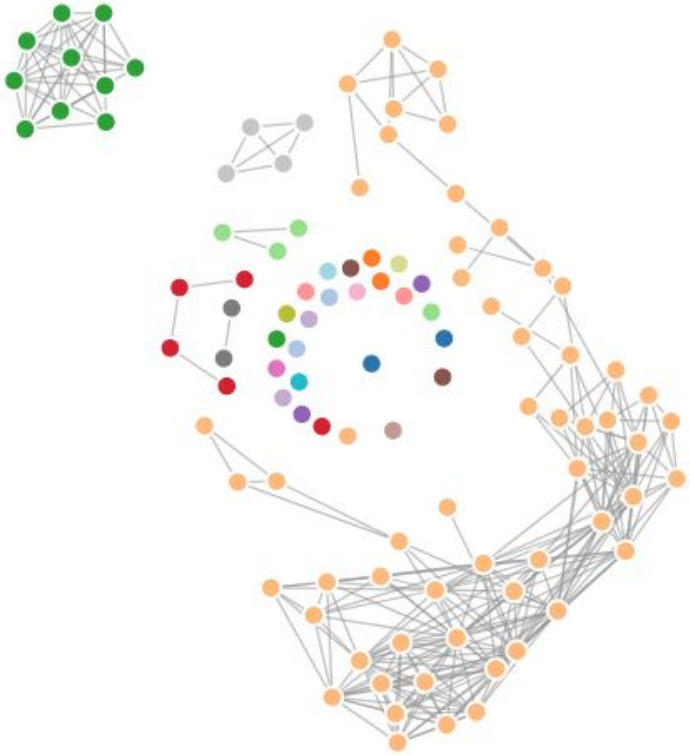
- Zovšeobecnenie metódy analýzy citlivosti použitej pri CNN na bežné siete
  - Použitie aj na iné dáta ako obrázky
- Bežná analýza citlivosti - problém so závislosťami medzi atribútmi

- Analýza citlivosti pre dáta so závislými atribútmi
- Lepšie určenie dôležitosti atribútov
  - Pre jedno pozorovanie
  - Pre model celkovo

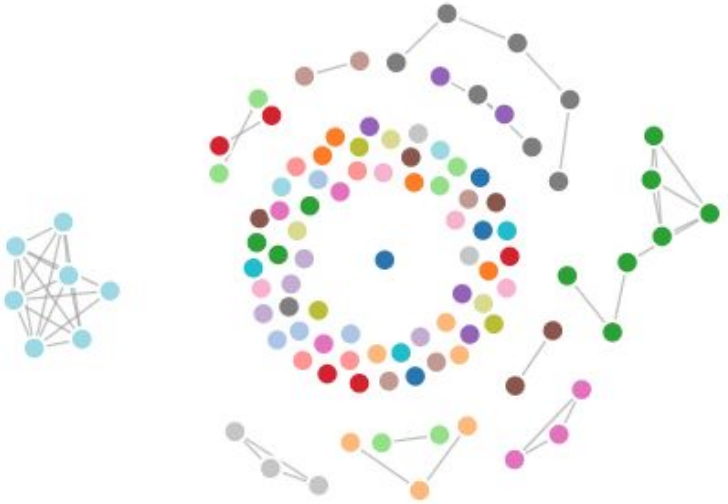
Postup

1. Nájdenie vzájomne závislých atribútov
    a. Korelačný koeficient
    b. Iný prístup ?

Závislé atribúty



Korelácia cez 0,7



Korelácia cez 0,9

Postup

1. Nájdenie vzájomne závislých atribútov
   a. Korelačný koeficient vyšší ako hranica
   b. ? iný prístup ?
2. Pre každý atribút - ‘odstránenie’ spolu s jeho korelovanými atribútmi
   a. Hodnota 0
   b. Stredná hodnota
   c. Hodnota na základe korelácie ?
3. Sledovanie zmien výstupu po ‘odstránení’
4. Určenie dôležitosti atribútu pre dané pozorovanie
5. Agregovanie cez všetky pozorovania - dôležitosť pre celý model

Otázky

- Je to relevantné?
- Problémy pri tomto postupe?
- Vyhodnotenie metódy?
- Dataset so závislými atribútmi?