Strojové učenie a neurónové siete - **zadanie 1** - analýza dát, jednoduché neurónové siete

01. október 2020
I-SUNS cv. 2

Zadanie 1

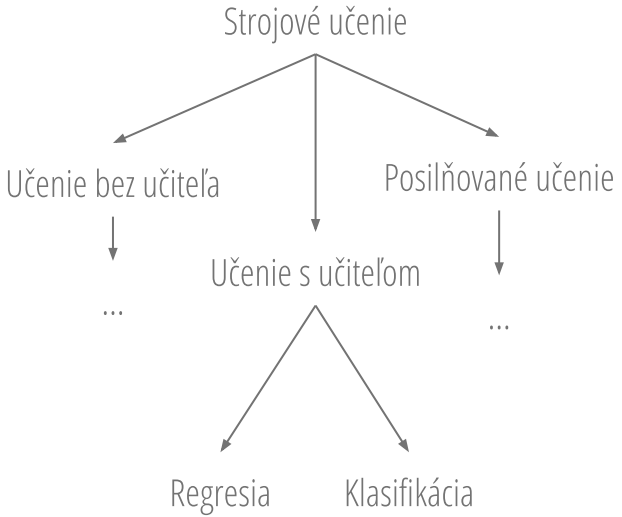
- Zadanie a bodovanie je k dispozícii na dokumentovom serveri v AIS
- Máte sa naučiť:
  1. Načítanie dát
  2. Základná analýza dát
  3. Čistenie dát, spracovanie textových stĺpcov, vyberanie množín
  4. Príprava dát na vstup do siete (škálovanie)
  5. Použitie neurónovej siete na regresiu/klasifikáciu
  6. Základné vyhodnocovanie úspešnosti sietí
- Dataset: dostupný v AIS aj s popisom stĺpcov (prescreening pacientov na srdcovú chorobu)

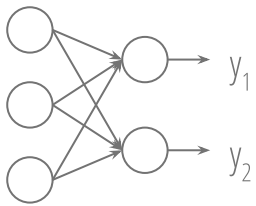
Obsah tejto prezentácie

- Regresný vs. klasifikačný problém
- Škálovanie
- Analýza vzťahov
- ANN (zatiaľ ako blackbox)
- Vyhodnocovanie modelov
- Postup riešenia demonštrovaný na datasete zo zdroja, využité knižnice numpy, pandas, plotly, sklearn

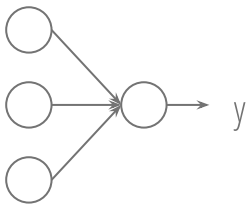
Regresia vs klasifikácia

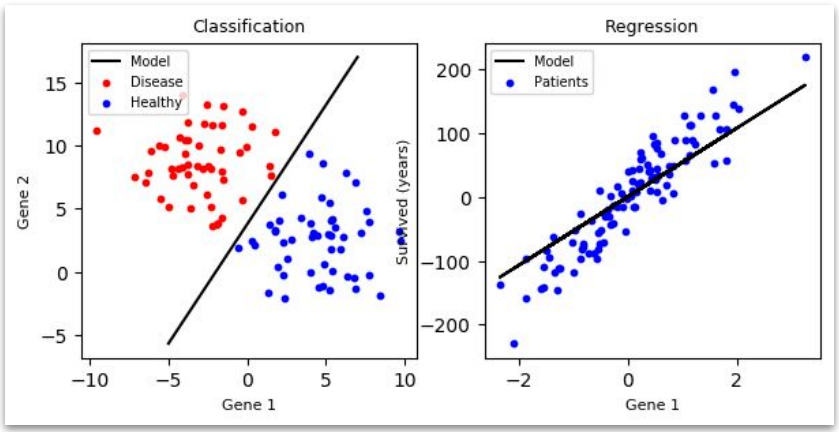
- Mapujú výstup y na vstup x: y=f(x)
- Spojitý alebo diskrétny výstup (“problému” nie modelu)





napr. so sigmoidou ako aktivačnou fciou - $y_n \in \langle 0,1 \rangle$

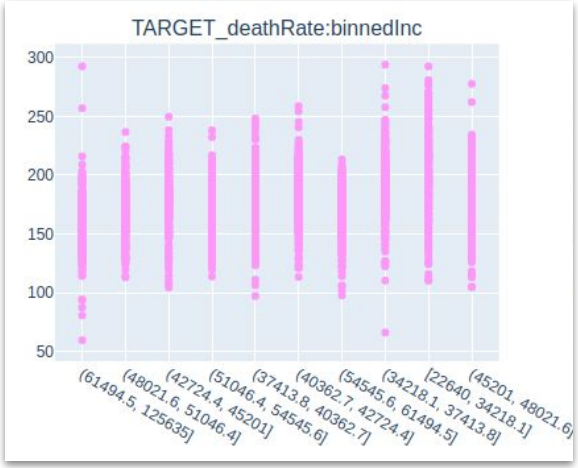


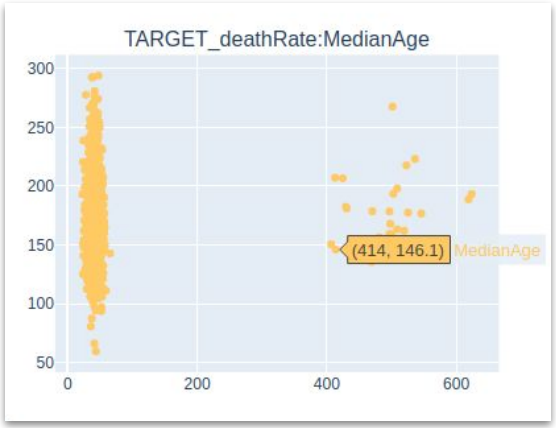


Zdroje obrázkov:
https://dev.to/petercour/machine-learning-classification-vs-regression-1gn

Príprava dát

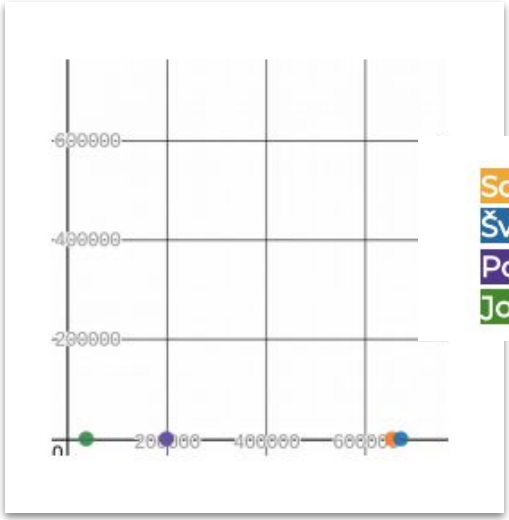
- Chýbajúce hodnoty
  - Nahrádzam nulou
  - Predpokladám hodnotu (*imputation)*
  - Ignorujem príznak/vzorku
- Nečíselné (textové) hodnoty
  - Dajú sa hodnoty mapovať na čísla?
  - yes/no - true/false
  - good/better/the best
- Kontrola správnosti dát
  - **Duplikáty**, outliers
  - Somariny sa ťažko modelujú
- Numerická stabilita
  - Normalizácia/štandardizácia/škálovanie príznakov



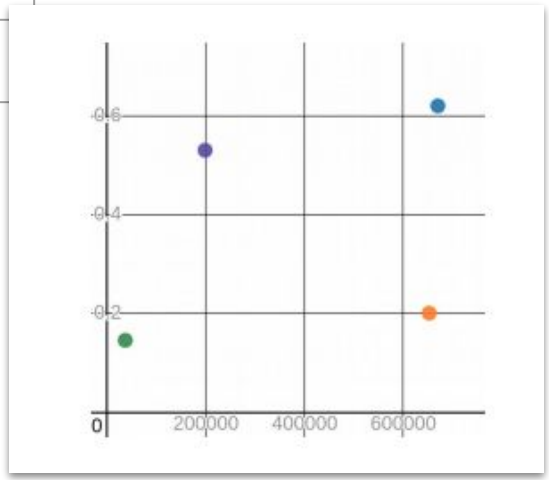


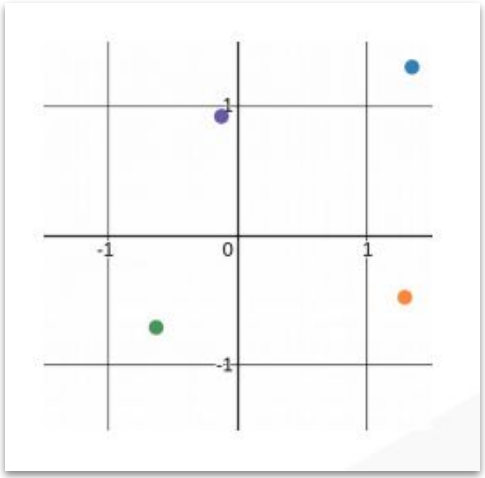
Normalizácia dát

	HDP (mil $)	Podiel žien v pracovnom pomere
Saudská Arábia	653 219	0.2
Švajčiarsko	670 790	0.62
Portugalsko	199 122	0.53
Jordánsko	37 517	0.145



Saudská Arábia
Švajčiarsko
Portugalsko
Jordánsko





Škálovanie dát

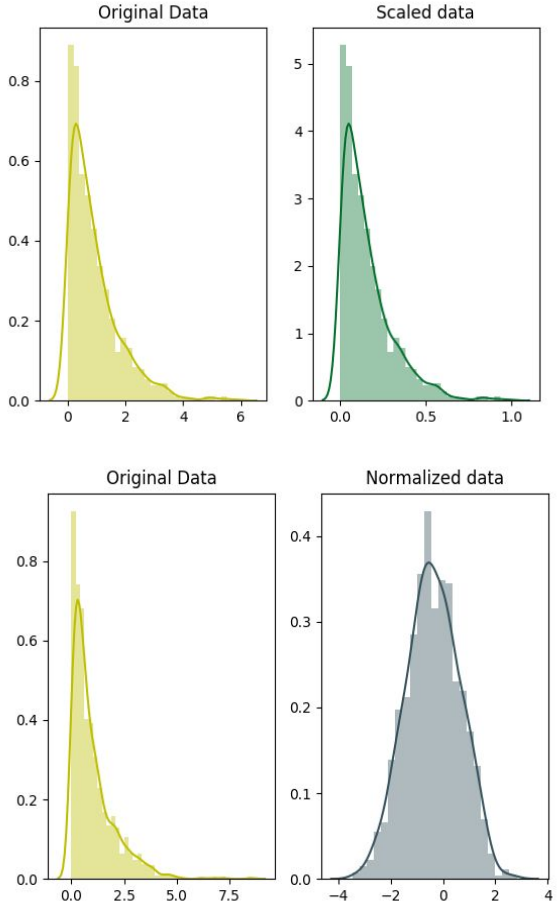
- Normalizácia
  - Mení rozsah (obv. medzi ⟨0,1⟩, ⟨-1,1⟩)
  - Ak sa premenná neriadi Gaussovým rozdelením
  - Min-max, z-score, ...

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

- Štandardizácia
  - Mení rozsah aj distribúciu
  - Ak sa premenná riadi Gaussovým rozdelením (povedzme)

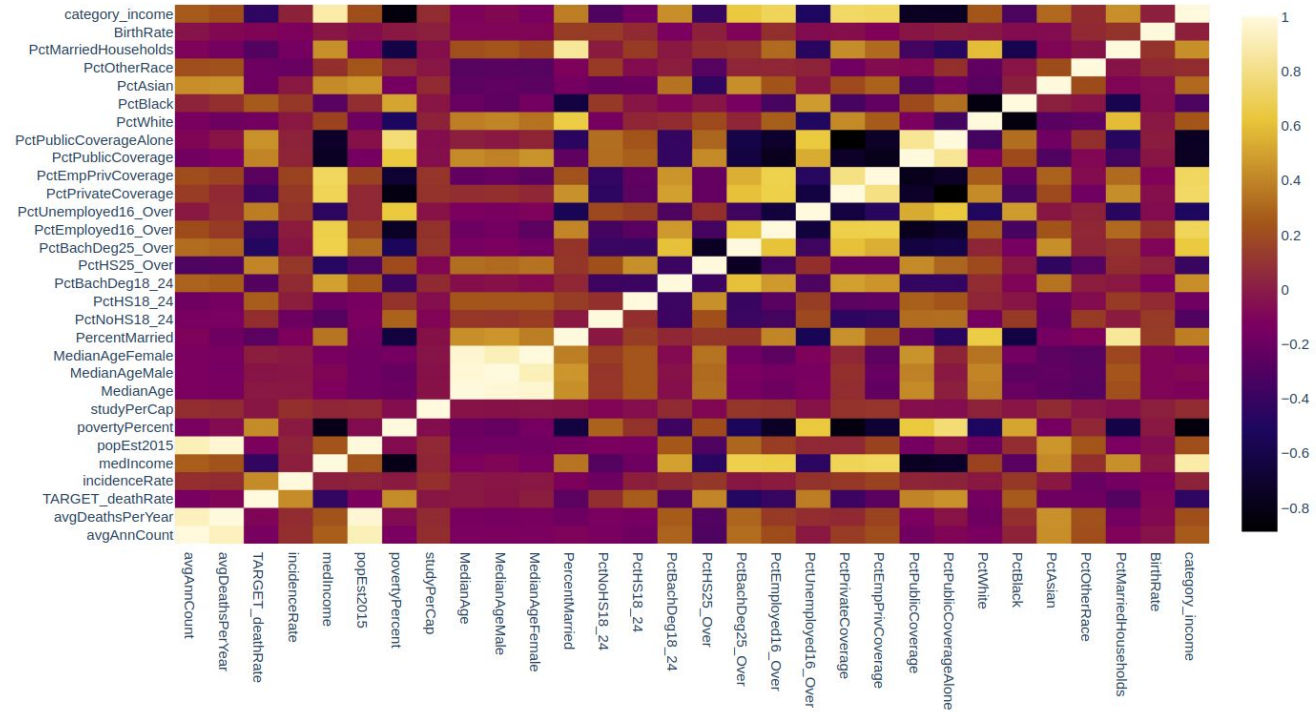
$$x_{norm} = \frac{x - \mu}{\sigma}$$

- Pre modely využívajúce gradientný zostup aj založené na vzdialenostiach
- Diskusia - vstupné a výstupné veličiny



Analýza dát

- Je dobré si položiť otázku, ktoré príznaky od seba závisia (a ako veľmi)
- Objavovanie vzťahov, vyberanie príznakov do modelov
- Napr. korelačná matica:



Neurónové siete

- Delenie trénovacia/validačná/testovacia množina (napr. 8:1:1)
  - Trénovacia množina - vstup do trénovania
  - Validačná množina - počas trénovania overuje schopnosť generalizovať na nevidených dátach
  - Testovacia - vyhodnocovanie výsledkov
- Architektúra:
  - Počet vstupných a výstupných neurónov
  - Počet skrytých neurónov
  - Aktivačné funkcie
- Trénovanie
  - Kriteriálna funkcia
  - Parameter rýchlosti učenia
  - Zastavovacia podmienka
  - Solver

```
classifier = MLPClassifier(hidden_layer_sizes=(20,), alpha=0.001, tol=0.00001, random_state=1, verbose=True,
                           max_iter=1000)
classifier.fit(train_X, train_y)
y_pred_nn = classifier.predict(test_X)
```

```
classifier = MLPRegressor(hidden_layer_sizes=(20,), alpha=0.001, tol=0.00001, random_state=1, verbose=False,
                          max_iter=5000)
classifier.fit(train_X, train_y.values.ravel())
y_pred_nn = classifier.predict(test_X)
```

Vyhodnocovanie úspešnosti regresorov

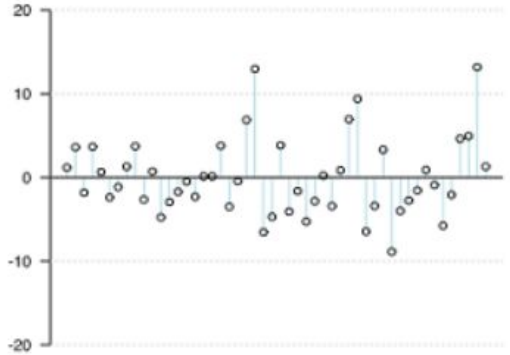
- Štatisticky
  - Mean squared error:

$$MSE = \frac{1}{n}\Sigma\left(y - \widehat{y}\right)^2$$

  - R^2 :

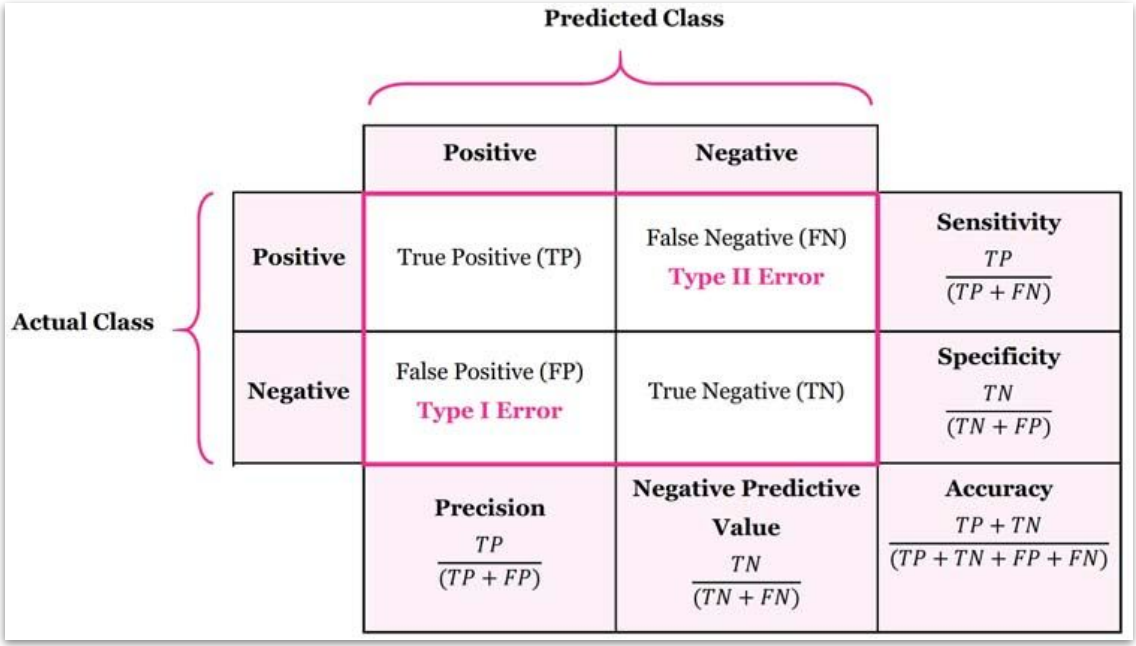
$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

- Graficky
  - Napr. residual plot



Vyhodnocovanie úspešnosti klasifikátorov

- Štatisticky
  - Celková úspešnosť
  - Top-x chyba
- Graficky
  - Confusion matrix



Zdroje obrázkov:
https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

Priestor na otázky