Strojové učenie - Machine Learning

Základy internetu vecí

ÚINF PF UPJŠ

EURÓPSKA ÚNIA
Európsky sociálny fond
Európsky fond regionálneho rozvoja
OPERAČNÝ PROGRAM
ĽUDSKÉ ZDROJE
MINISTERSTVO ŠKOLSTVA, VEDY, VÝSKUMU A ŠPORTU SLOVENSKEJ REPUBLIKY

Ktoré zvieratá mám rád?

bažant, kuna, rys, **ďateľ**, vlk, jazvec, bobor, vydra, los, sob, **žralok**, **pavúk**, **dážďovka**, **had**, tiger, lev, leopard, **orol**, **sliepka**, **hus**, mačka, **sova**, **netopier**, **moriak**, kôň, mlok, **slimák**, rys, puma, slon, antilopa, nosorožec, hroch, krokodíl, gorila, **komár**, **mucha**

Ktoré zvieratá mám rád?

- pes
- delfín
- lastovička
- jašterica
- jeleň
- prasa
- sokol
- medveď
- kapor
- líška

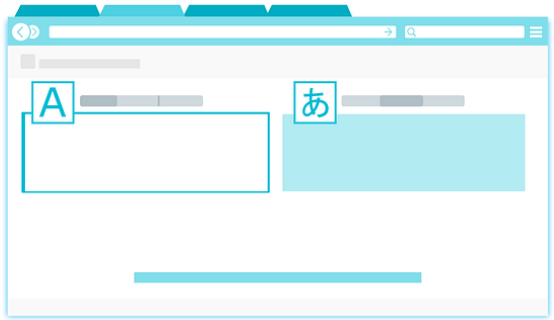
Čo je strojové učenie?

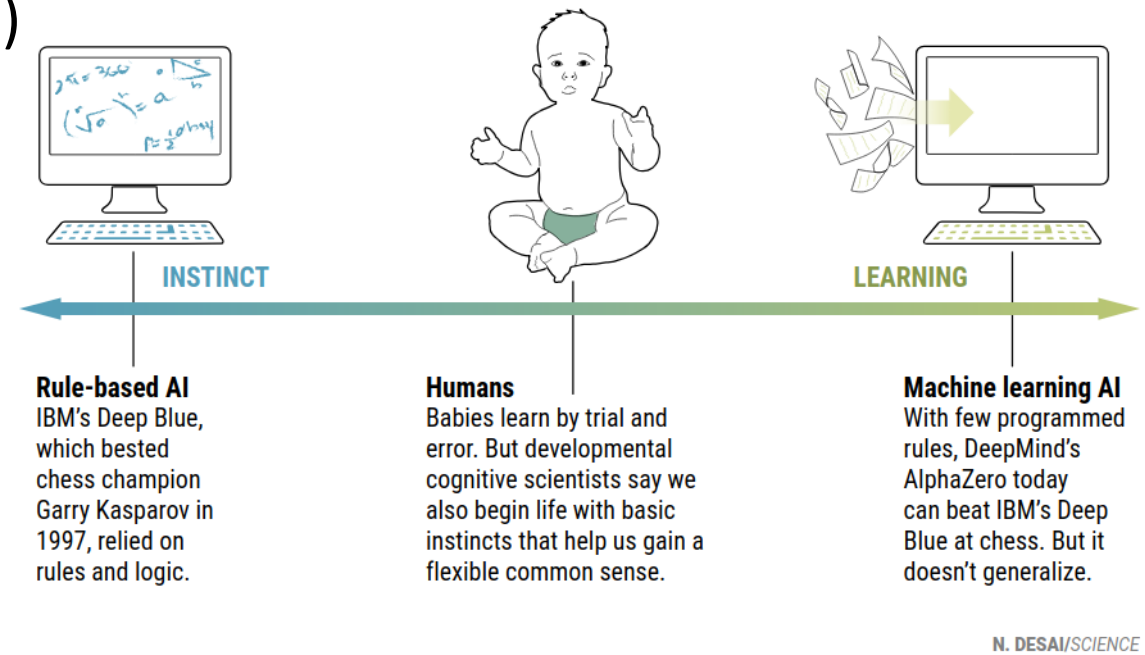
Machine learning algorithms build
a mathematical model of sample data,
known as "training data",
in order to make
predictions or decisions
without being
explicitly programmed to perform the task.

Arthur L. Samuel, 1959

Kedy použiť ML?

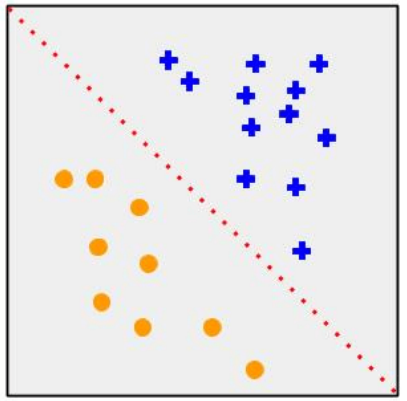
- "Tradičné" algoritmy vs. učenie z dát
- Nie je to riešenie na všetko
- Ak sa dá niečo vyriešiť bez ML, tak netreba ML
- Vhodné v situáciach:
  - ak nevieme **definovať pravidlá** (počasie)
  - ak nevieme **škálovať** (spam)



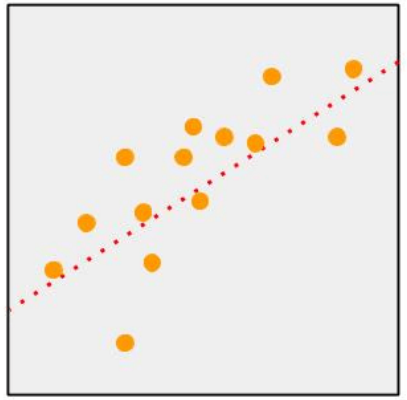


Typy problémov

- **Klasifikácia** (diskrétny výstup)
  - Binárna alebo multi-class
  - Kúpi si zákazník tento produkt?, Je tento email spam?
  - Čo robí používateľ smartfónu?
- **Regresia** (spojitý výstup)
  - Aká je výška človeka podľa jeho váhy?
  - Aká je cena domu?
- **Zhlukovanie**
  - Učenie bez učiteľa



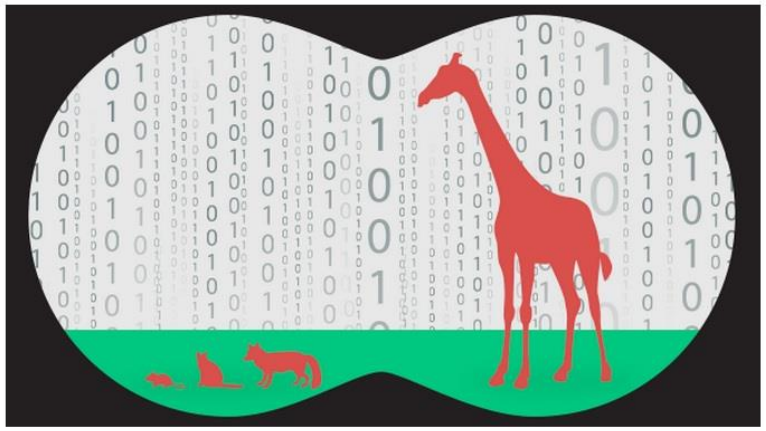
Classification

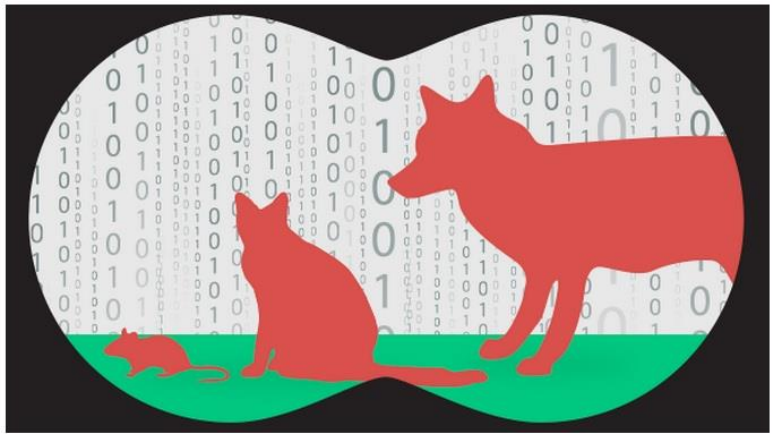


Regression

Vytvárame ML riešenie

- Definovať problém
- Zozbierať dáta (označené)
- Analyzovať dáta (prípadne upraviť)
- Vybudovať model - dodať vstupy/features
- Použiť model na predikciu





"Ak mám **hodinu** na vyriešenie problému, strávim
55 minút premýšľaním o **probléme** a
5 minút premýšľaním o **riešení**."

Úprava dát

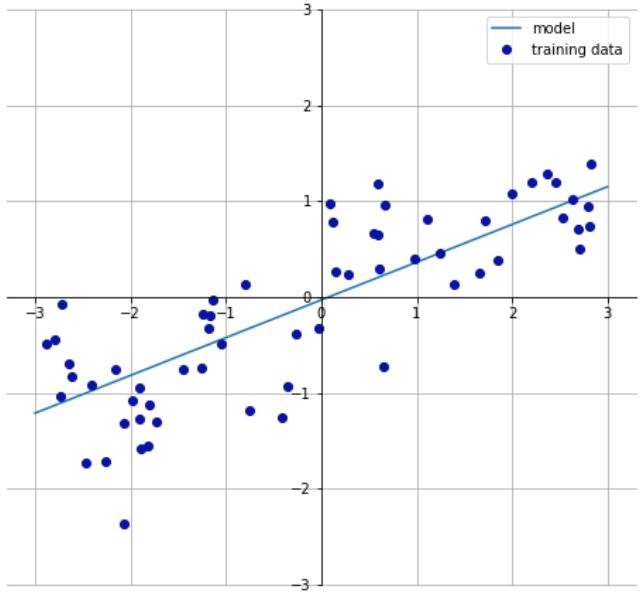
- **Nahradiť chýbajúce alebo nevyhovujúce hodnoty**
  - N/A vs. 0, alebo doplniť priemerom/mediánom
- **Kartézsky súčin**
  - produkty, farby - biele_tricko, zelene_tricko, zelene_nohavice
- **Nelineárne transformácie**
  - priradiť číselné hodnoty do kategórii (napr. vek zákazníkov)
- **Domain-specific features**
  - dĺžka, šírka, výška - pridať súčin; váha, výška - BMI
- **Variable-specific features**
  - parsovanie stringov
- **Normalizácia vstupov**
  - škálovanie a normalizácia podľa strednej hodnoty

Rozdelenie datasetu

- Cieľ - označiť neznáme dáta (**predikcia**)
- Typicky 70:30 (predvolené v AWS), niekedy až po 80:20
- Trénovacia vzorka - učenie (aké sú medzi dátami vzťahy, závislosti)
- Testovacia vzorka - evaluácia (aký dobrý je model)
- Náhodné poradie
- Pozor na prílišné účenie - nie je dostatočne všeobecné

Lineárny model

- **Model –** SVM, neurónová sieť, bayesovská sieť, rozhodovací strom
- **Lineárny model** - lineárna kombinácia vstupov (features)
- Učiaci proces nastavuje jednu váhu pre každý vstup
- `y = w[0] * x[0] + w[1] * x[1] + ... + w[p] * x[p] + b`
- **Hyperparametre**
  - Učiaci pomer
  - Veľkosť modelu
  - Počet prechodov dátami
  - Poradie vstupných dát
  - L1 a L2 regularization - vhodné ak sú korelácie vo vstupoch
    - L1 spôsobuje redší model a redukuje šum
    - L2 znižuje hodnoty váh a stabilizuje ich tam kde korelujú



Evaluácia - príklad

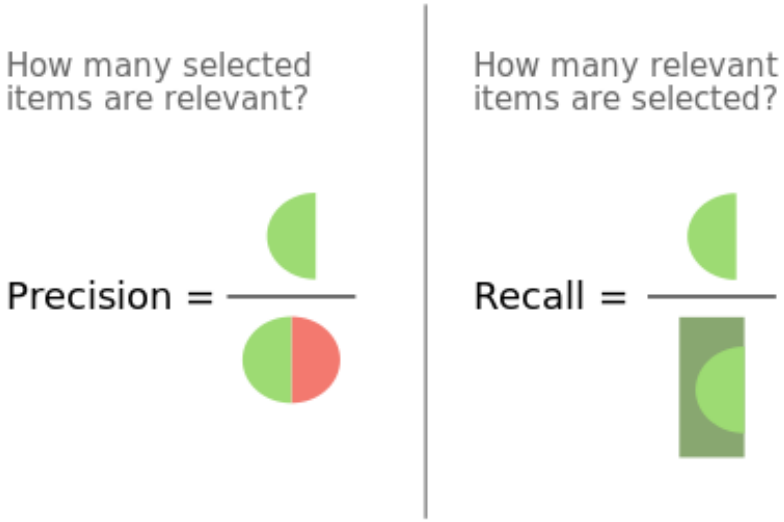
- Vzorka ľudí - chorobu má 0.5%

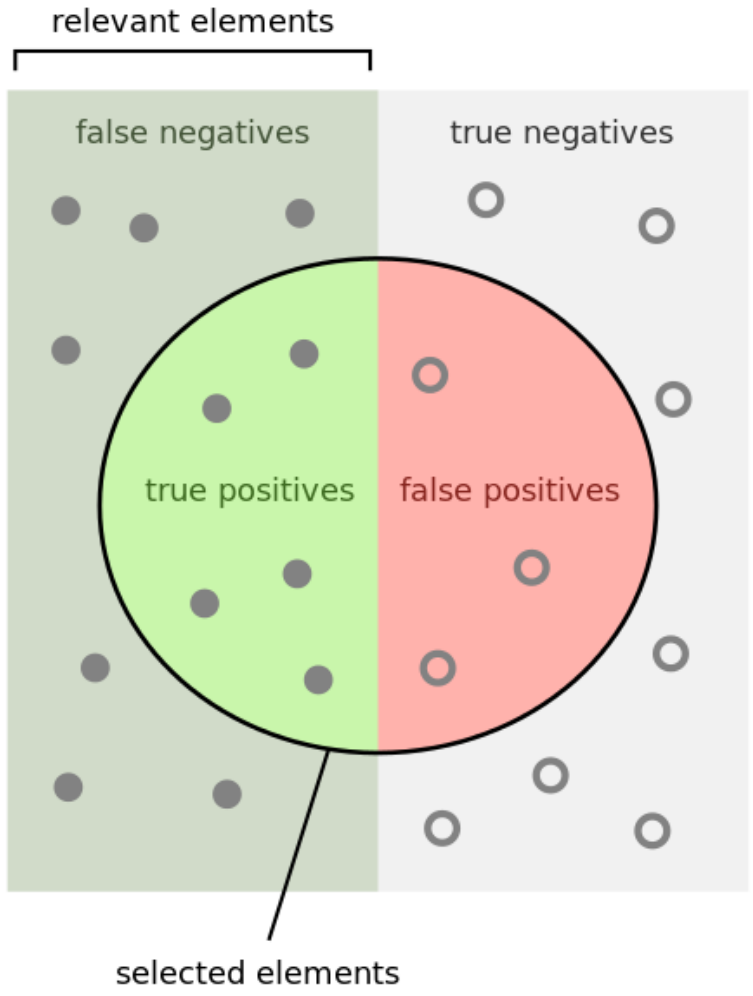
- **A1** algoritmus odhaduje či má človek chorobu
- Chyba algoritmu A1 je **1%**

- **A2** algoritmus o každom pacientovi tvrdí, že chorobu nemá
- Chyba algoritmu A2 je **0.5%**

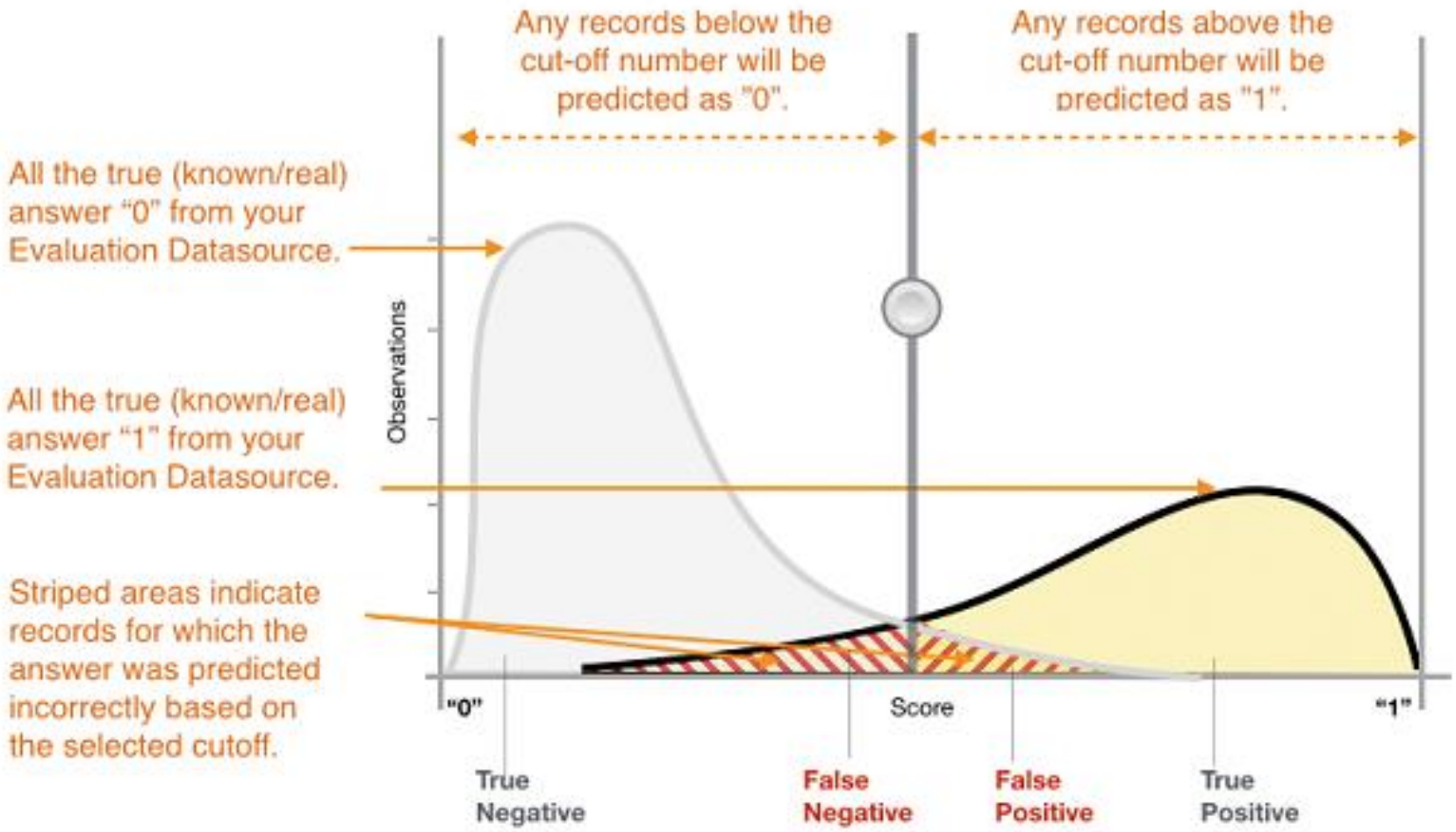
Evaluácia – F1 skóre

- $F_1\ skóre = \frac{2 * P * R}{P + R}$

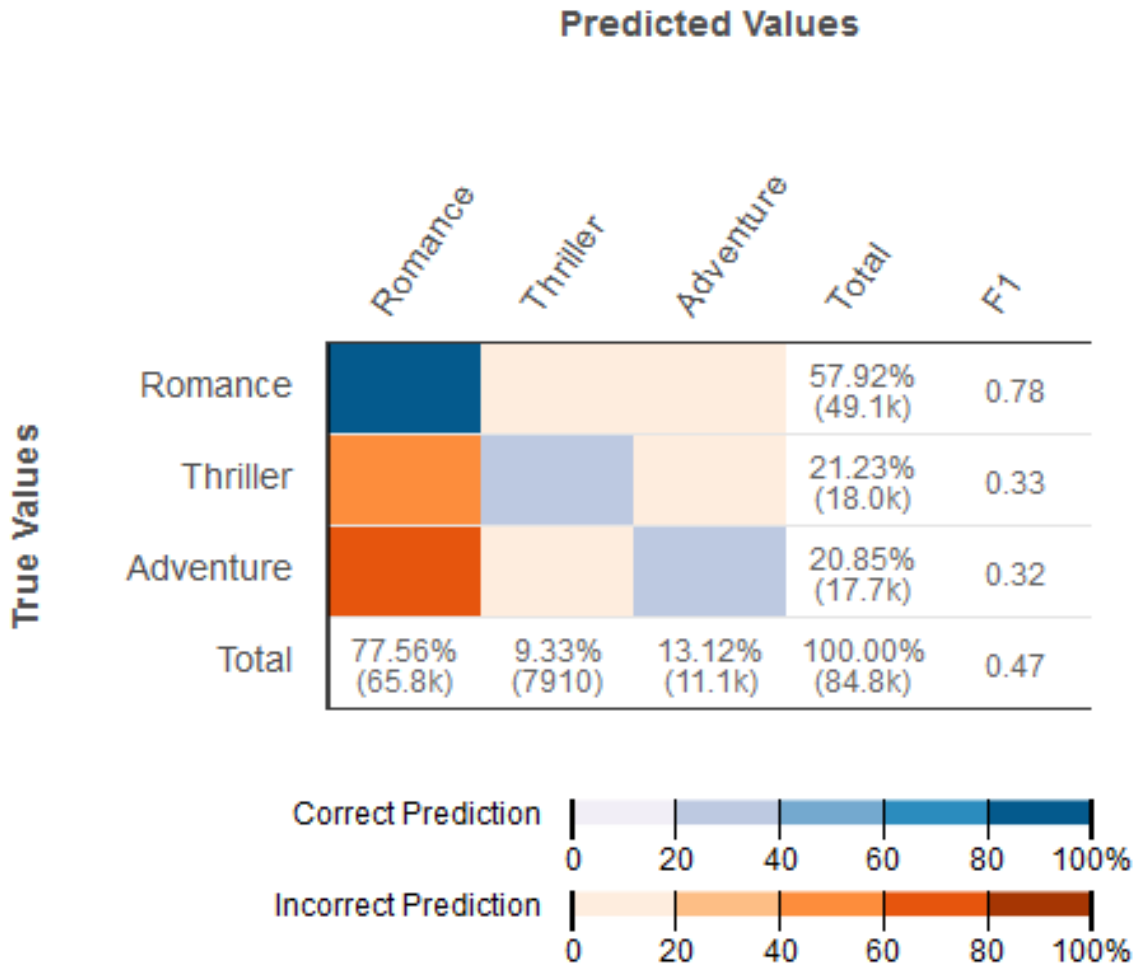




Binárna klasifikácia (cut-off)



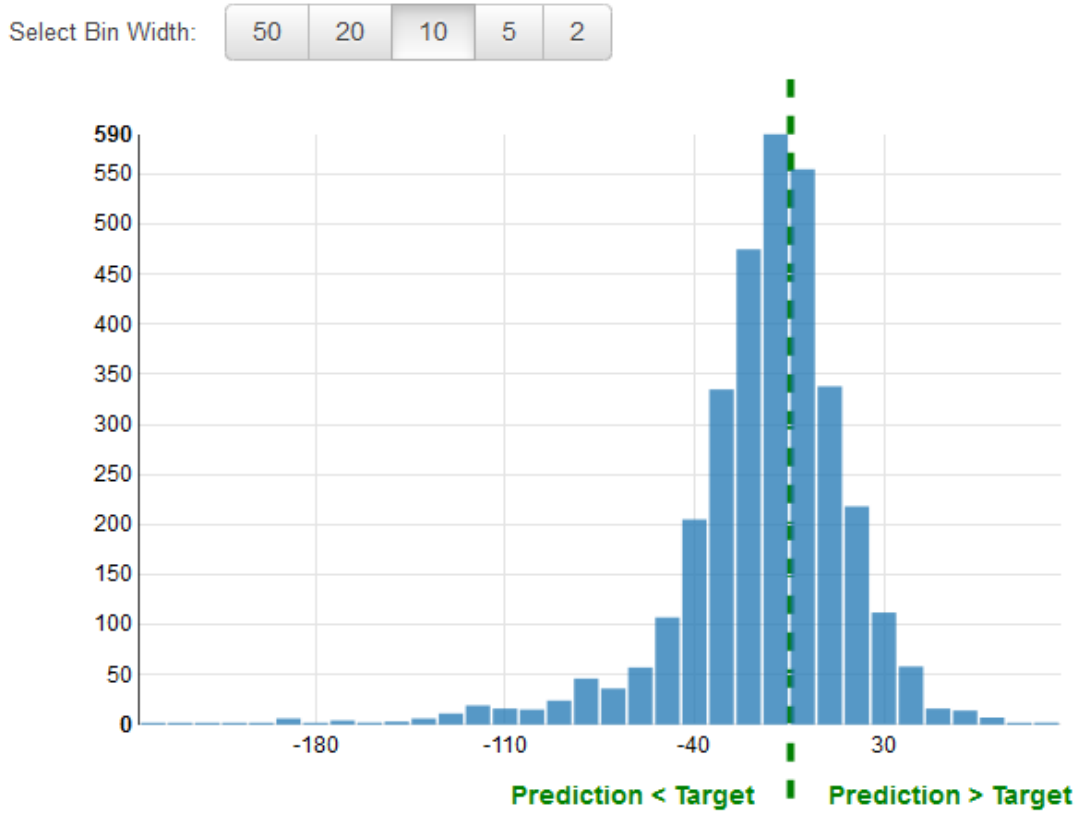
Evaluácia multiclass klasifikácie



Evaluácia regresie

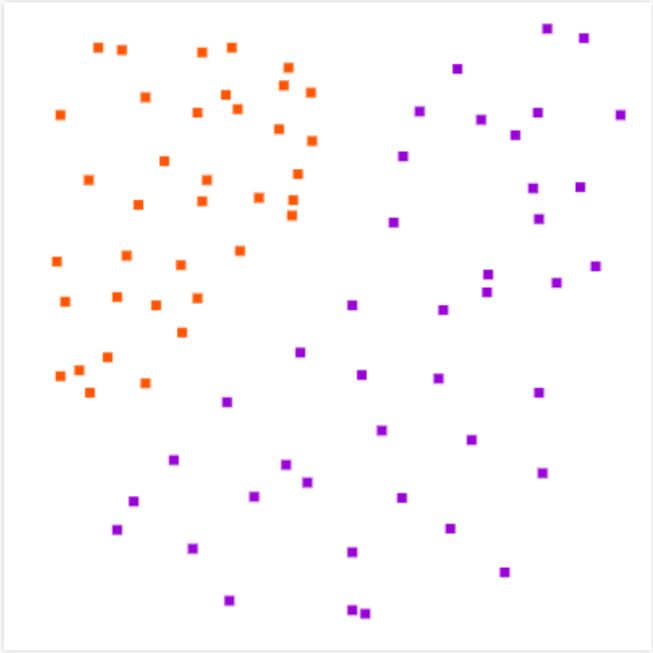
- **Residual** - rozdiel medzi cieľovou a predikovanou hodnotou
- Positive residual
  - model podhodnocuje
- Negative residual
  - model nadhodnocuje

- *Error* – stredná hodnota populácie
- *Residual* - stredná hodnota vzorky
- Rôzna metrika



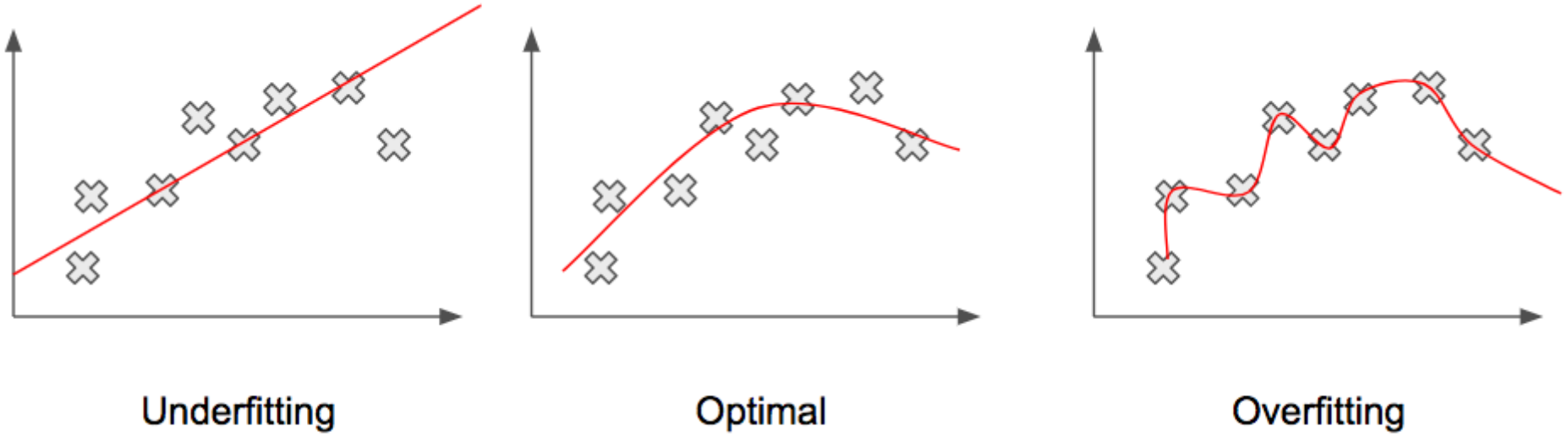
Zlepšenie presnosti modelu

- Pridať nové dáta do trénovacej vzorky
- Pridať viac premenných a zlepšiť spracovanie vstupov
- Vyladiť parametre modelu (napr. učiaci pomer, počet prechodov dátami)



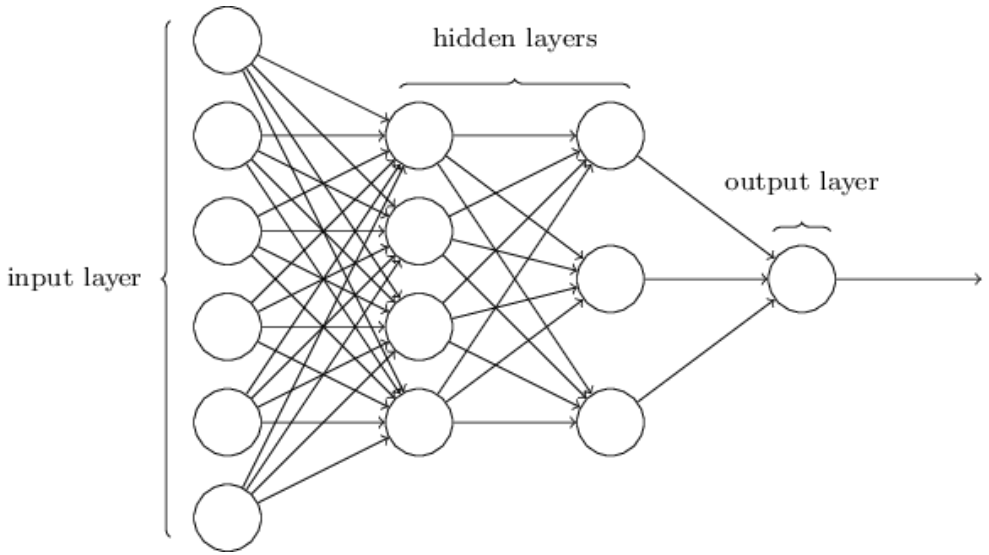
Underfitting & Overfitting

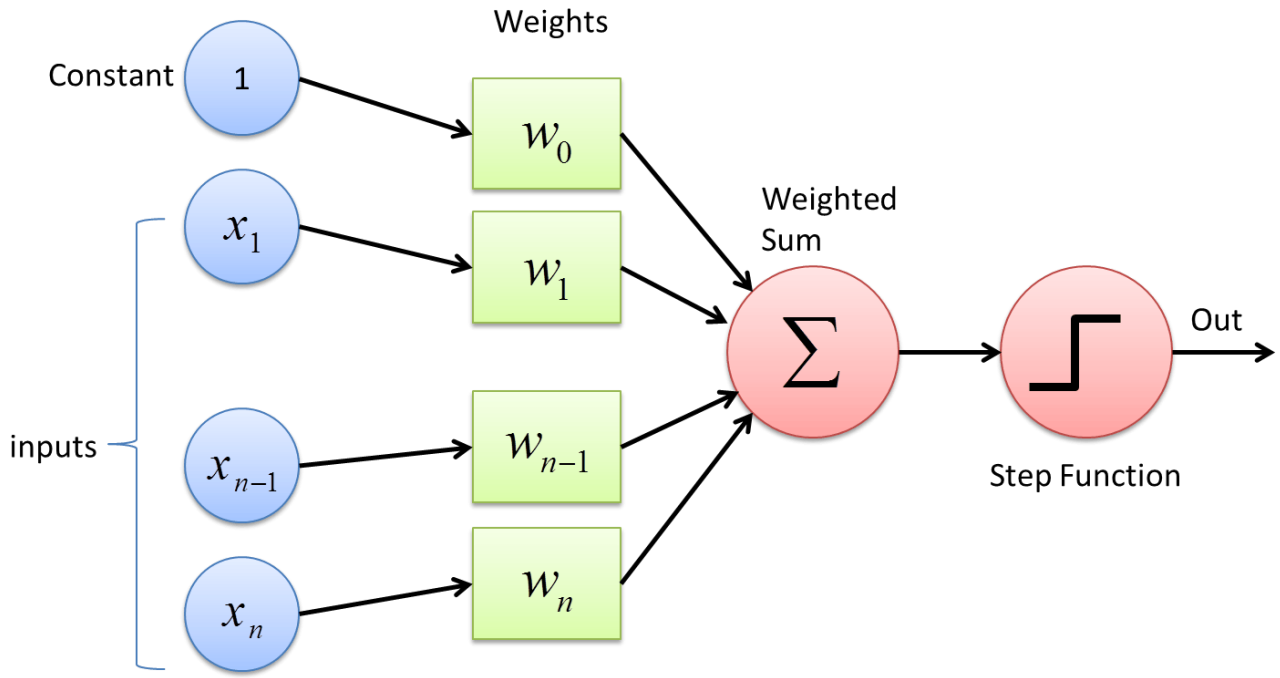
- **Underfitting**
  - pridať domain-specific vstupy, kartézske súčiny ...
  - použiť menej regularizácie
- **Overfitting**
  - menej kombinácii vstupov
  - viac regularizácie



Neurónové siete

- Playground - https://playground.tensorflow.org
- Perceptron, Backpropagation, FeedForward NN, aktivačná funkcia, online vs. batch





TensorFlow

- https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/eager/custom_training_walkthrough.ipynb



Ďakujem za pozornosť