Klasifikácia

Matej Kubinec

Čo je to klasifikácia?

- Klasifikácia je proces predikcie triedy danej množiny dát.
- Triedy sú nazývané aj cieľ/značka alebo kategória.
- Patrí do skupiny učenia s učiteľom (trenovacie dáta majú známu triedu)

Prečo vôbec klasifikovať dáta?

- detekcia spamu (binárna klasifikácia)
- detekcia podozrivého trafficu na sieti
- odporúčanie reklám
- ...

Ako môžeme klasifikovať dáta?

1. Lazy Learners
    - Trieda je určená na základe podobnosti/blízkosti k testovacím dátam.
    - Oproti Eager Learners je náročnejší výpočet predikcie.
    - napr. k-nearest neighbor, case-based reasoning
2. Eager Learners
    - Zostrojí sa klasifikátor založený na trénovacích dátach.
    - Oproti Lazy Learners je výpočet náročnejší v trénovacej fáze.
    - napr. rozhodovacie stromy, Naive Bayes klasifikátor, neurónové siete

Dáta - Golf/Weather

zdroj

Outlook	Temperature	Numeric Temperature	Nominal Humidity	Numeric Humidity	Nominal Windy	Play
overcast	83	hot	86	high	FALSE	yes
overcast	64	cool	65	normal	TRUE	yes
overcast	72	mild	90	high	TRUE	yes
overcast	81	hot	75	normal	FALSE	yes
rainy	70	mild	96	high	FALSE	yes
rainy	68	cool	80	normal	FALSE	yes

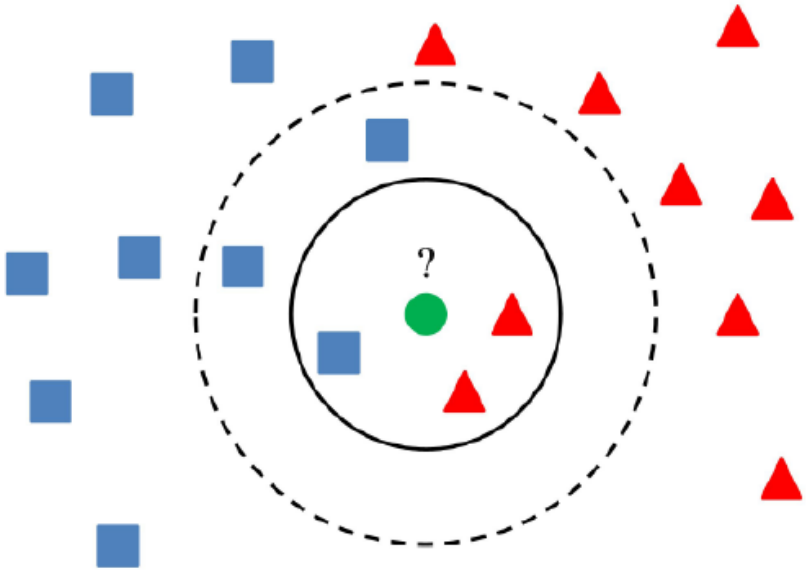
Predspracovanie

- prevod kategoriálnych atribútov

outlook	outlook_overcast	outlook_rainy	outlook_sunny
overcast	1	0	0
overcast	1	0	0
overcast	1	0	0
overcast	1	0	0
rainy	0	1	0
rainy	0	1	0
rainy	0	1	0

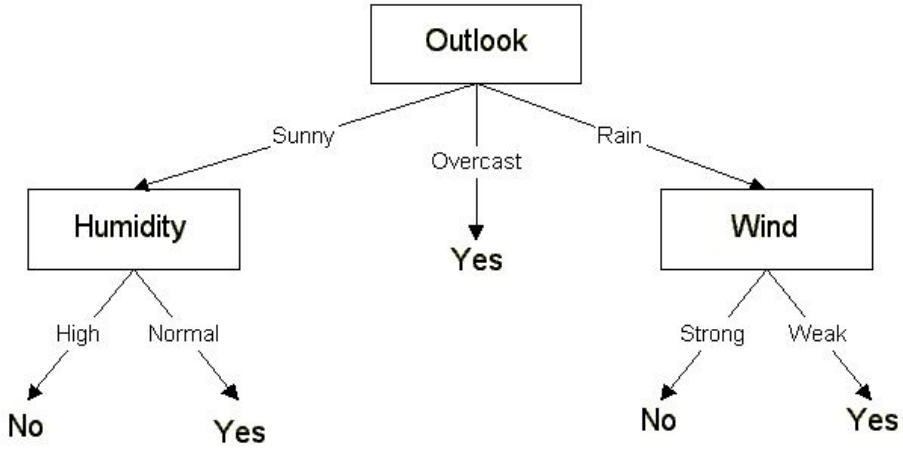
K-Nearest Neighbour

- jeden z najzákladnejších algoritmov klasifikácie
- učenie s učiteľom
- jeden parameter *k* - počet najbližších prvkov
- rozhoduje sa len na základe trénovacích dát
- knižnice: sklearn



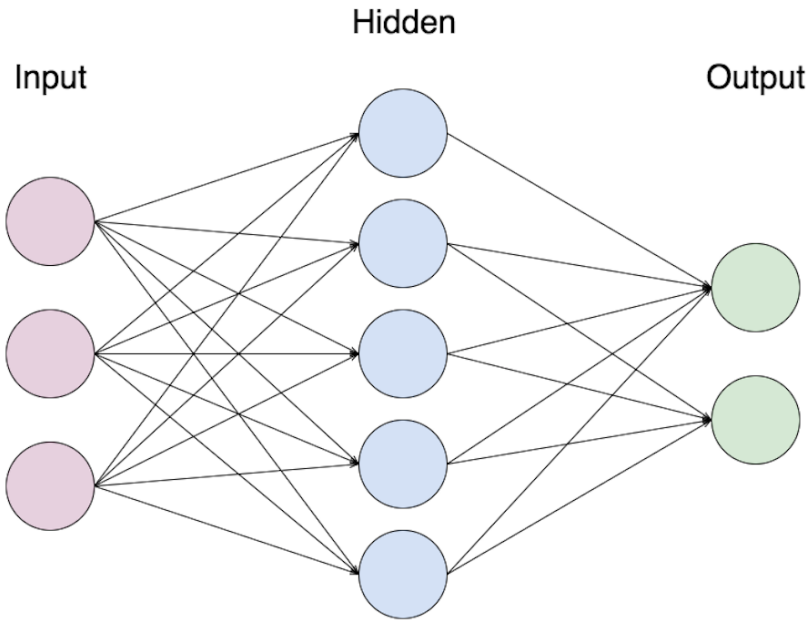
Rozhodovacie stromy

- vetvia sa na základe atribútov
  - kategoriálne - na zaklade hodnoty (outlook = sunny)
  - numerické - väčšie/menšie ako nejaký prah (temperature > 60)
- je jasné prečo sa zaradil záznam do danej triedy
- top-down pristup
- delenie na základe zisku informácie
- vhodné pre kategoriálne dáta
- knižnice: sklearn



Neuronové siete

- množina prepojených neurónov
- vstupná vrsta + výstupná vrstva + skryté vrstvy
- každý vstup na neuróne má priradenú váhu
- funguje ako BlackBox, po naučení nie je možné určite na základe čoho sa rozhoduje
- vhodnejšie na číselné hodnoty
- knižnice: tensorflow, keras, sklearn



Vyhodnocovanie

- rozdelenie dát na trénovaciu a testovaciu množinu
  - naučenie na trénovacej množine
  - otestovanie presnosti na testovacej
- k-fold validácia
  - dataset je rozdelený na k disjunktných podmnožín, približne rovnakej veľkosti

Otázky

Ďakujem za pozornosť!