Redukcia dimenzie dát – motivácia a obsah

Na predmete sa venujeme dvom hlavným oblastiam, ktorých typické použitie je ako "pomocné" nástroje pre viacrozmerné štatistické metódy / metódy strojového učenia: umožňujú efektívnu reprezentáciu výstupov týchto metód a efektívny výber premenných do modelov.

Metódy extrakcie premenných

Ak máme dáta, ktoré majú veľa premenných (napr. pre ľudí: vek, príjem, počet detí, počet rokov vzdelania, ...), tak častokrát ich môžeme chcieť dobre reprezentovať menším množstvom premenných. Typické dôvody sú:

- Vykreslenie: ak chceme na obrázku vidieť, ako sa z pohľadu nameraných dát líšia muži a ženy, alebo ak chceme vykresliť dáta spolu s výstupmi nejakej metódy (klasifikácia firiem na zbankrotované/nezbankrotované, rozdelenie ľudí do niekoľkých zhlukov, ...). To je celkom bezproblémové, keď majú dáta iba 2 premenné (vykreslíme "ofarbené bodky" do 2-rozmeru), ale ak je premenných napr. 10, tak by sme potrebovali 10-rozmerný obrázok.
- Zabránenie pretrénovaniu: ak je v modeli priveľa premenných, môže dôjsť k pretrénovaniu (overfitting), čo znamená, že sa model príliš prispôsobí existujúcim dátam a potom nevie veľmi dobre predikovať.

Preto boli vyvinuté metódy, ktoré skonštruujú menší počet nových premenných na základe už existujúcich (napr. nájdu najinformatívnejšie kombinácie veku, príjmu, počtu detí, počtu rokov vzdelania, ...) tak, aby stále celkom dobre reprezentovali pôvodné dáta.

Metódy: metóda hlavných komponentov (PCA - principal component analysis), faktorová analýza, projekčné sledovanie, mnohorozmerné škálovanie, Isomap, t-SNE, autoencoders.

Metódy selekcie premenných (výber premenných do modelu)

Príbuzný problém je, že máme napr. 10 vysvetľujúcich premenných v lineárnej regresii a aby sme zabránili pretrénovaniu, do modelu nezahrnieme nutne všetky. Tentokrát teda nekonštruujeme nové premenné, ale metódy určujú, akú sadu z existujúcich premenných je ideálne zobrať. Metódy budú pravdepodobne demonštrované na lineárnej regresii, ale väčšina z nich je priamo aplikovateľná a aplikovávaná aj v iných predikčných a klasifikačných metódach.

Metódy a témy: dopredná a spätná selekcia (s pripomenutím informačných kritérií); trénovacia, validačná a testovacia vzorka; krosvalidácia; penalizácie v modeloch: hrebeňová regresia a Lasso.

Ak vyjde čas, tak sa stručne zoznámime aj s treťou oblasťou: *metódy výberu podvzorky* (subsampling), ktoré riešia problém, keď máme priveľa pozorovaní (ľudí/firiem/objektov ...) – typicky priveľa v zmysle, že s následnými modelmi sa počítač nezvláda vysporiadať v rozumnom čase. V takých prípadoch je možným riešením práve výber iba časti pozorovaní (podvzorka), ktoré by ale mali byť informatívne alebo reprezentatívne.