DVZ: poznámky k semestrálním pracem v ZS 2012/13

Obecně

Nepoužívat **metody** jako černé skříňky, rozumět jim - vědět, co, proč a jak dělají. Metody mohou mít určité nároky na data, která zpracovávají (např. v ohledu na chybějící data, rozložení dat, počet instancí a příznaků, normalizace apod.). Znalost metod rovněž umožní kontrolovat, zda dostávám to, co čekám. Příklady: *Split Data* použité pro vyvažování (!), nesmyslná regrese modelující něco jiného, než je žádoucí.

Struktura práce: v abstraktu vše důležité včetně výsledků, vlastní práci rozdělit na metody/výsledky, v závěru diskutovat dosažené výsledky.

Předzpracování dat

Transformace: normalizace příznaků do srovnatelných škál (kvůli k-NN, shlukování), logaritmická či jiná transformace (pokud si to data očividně žádají).
Transformace se provádí zpravidla na začátku celého procesu zpracování (před identifikací odhlehlých hodnot) a nikoli pouze pro účely vizualizace, ale i pro účely aplikace strojových metod (př. regrese na logaritmovaných veličinách).

Odlehlá pozorování: nemusí být nutné zahazovat celé instance, může stačit prohlásit odlehlá pozorování za chybějící a řešit je jako chybějící.

Chybějící data: zahazovat instance a/nebo zahazovat příznaky (někdy lze, jindy ne / kdy?) vs. doplňovat. Pozor: každý zásah do dat může znamenat zanášení "umělé" informace, která data poškodí. Řešením může být mnohonásobné nahrazování (*multiple imputation*) nebo citivostní analýza (*sensitivity analysis*).
Informativní "*missingness*" - pokud budou instance s chybějícími příznaky pouze v trénovací množině, nikoli v testovací, nebudou trénovací a testovací množiny výběry ze "stejné populace".

Nahrazené průměrem vs. mediánem.

Klasifikace

Selekce příznaků: ne vždy nutná. Nutná např. tehdy, pokud klasifikátor vyžaduje nekorelované příznaky. Vhodná v případě, kdy máme více příznaků, než pozorování (ale existují metody, které se vyrovnají i s takovou situací, pomocí tzv. regularizace, např. *Lasso*, hřebenová regrese (*Ridge regression*), *Elastic net)*.

Vyvažování není stratifikace!

Přesnost klasifikace na trénovací sadě: schopnost klasifikátoru modelovat daný typ dat (viz křivka učení).
Přeučení vs. generalizace.
Dobrý výkon na trénovacích datech nemusí znamenat, že klasifikátor bude úspěšný i na nezávislých (testovacích) datech.
Špatný výkon na trénovacích datech však znamená, že klasifikátor nedosáhne výrazně větší přesnosti

na nezávislých testovacích datech. Může to být problém klasifikátoru ale i toho, že data nejsou dobře klasifikovatelná.
Přesnost klasifikace na testovací sadě: odhad toho, jak se bude klasifikátor chovat v reálném nasazení. Přesnost na testovací sadě nebude větší než přesnost na trénovací sadě. Výrazně horší přesnost na testovací sadě ve srovnání s trénovací sadou je znakem přeučení.

Výpočet pravděpodobností chyb I./II. druhu. Závažnost chyb I./II. druhu. ROC křivka.

Při trénování je dovoleno "vše" (např. aplikace naivního Bayese při nesplnění předpokladů) - při testování se ukáže, zda to k něčemu bylo.

Metody

K-means není klasifikátor.
Perceptron není "lék první volby".
Je dobré metody intepretovat.

Výpočet pravděpodobnosti, že nový pacient bude mít chorobu

Zpravidla z původních dat, nikoli po předzpracování.
Nikoli z výstupů modelu, ale z dat (ptáme se na realitu, nikoli na predikci nějakého modelu)!

Modelování závislosti mezi veličinami (regrese)

Pozor na porozumění, co operátor *Linear regression* dělá (závislá proměnná musí být označena jako "*label*", za nezávislé jsou považovány všechny příznaky. Odhadnutý koeficient nenese sám o sobě informaci o statistické významnosti vztahu! Ani koeficient deteminace ne!
Koeficient korelace není obecně to samé jako koeficient nalezený metodou nejmenších čtverců v regresi.
Intercept může být vhodný, i když není významný, ale má v modelu své opodstatnění.

Hledání skryté struktury v datech (shluková analýza)

Chceme-li v datech hledat vnitřní strukturu, neznamená to, že se nutně objeví právě struktura, kterou zrovna čekáme (např. ta, podle které klasifikujeme).

Korelační matice odhalí pouze některé (!) závislosti mezi dvojicemi příznaků, nikoli vnitřní strukturu! K tomu je zpravidla třeba zohlednit všechny veličiny současně, což právě umí např. shluková analýza. Korelace navíc nemusí odhalit ani jednoduchou nelineární strukturu!

Samotný graf "paralelních souřadnic" průměrů shluků (bez zohlednění rozptylu) nemůže, pokud se průměry liší, být důkazem že shluky reprezentují odlišné podskupiny.

(Výběrový) korelační koeficient (***r***) vs. koeficient determinace ($\boldsymbol{R^2}$) vs. RMSE vs. odhad koeficientu $\boldsymbol{\beta}$ v lineární regresi. Tyto veličiny nenesou samy o sobě informaci o statistické významnosti!

Korelace vs. kovariance. Není pravda, že kovariance odhalí nelineární strukturu!

Tato věta není pravdivá: "*Tento reálný předpoklad vyplývá z centrální limitní věty: provedeme-li velké množství nezávislých pozorování, pak budou mít tato pozorování normálné rozdělení.*" Proč?