Vytěžování dat: první semestrální práce

Hledání nejlepšího modelu

Zadání:

Vyberte si data, která jsou použitelná pro klasifikaci nebo regresi. Data předzpracujte, a naimportujte do DM aplikace. Zvolte si klasifikátor (kNN, lineární a polynomiální separace, rozhodovací strom) nebo regresní algoritmus (kNN, lineární a polynomiální regrese, regresní rozhodovací strom). Najděte nastavení parametrů zvoleného algoritmu tak, aby produkoval co nejlepší modely.

Podrobnější specifikace:

1. Výběr dat
Data pro semestrální práci můžete najít na internetu. Doporučená knihovna volně dostupných dat pro strojové učení (UCI = http://archive.ics.uci.edu/ml/) obsahuje stovky různých datových souborů.
Pokud máte zájem zpracovávat vlastní data, musí být k dispozici nejpozději následující cvičení.
Volbu dat schvaluje váš cvičící!

2. Předzpracování
Dbejte, aby vámi zvolená data nebyla náročná na předzpracování - měla by obsahovat hlavně číselné atributy (nominální atributy je třeba zakódovat 1zN), minimum chybějících dat (vymazat), žádné odlehlé hodnoty (smazat nebo softmax scaling), atributy stejných rozsahů (minmax normalizace), atd.

3. Import dat
DM aplikace, kterou probíráme na cvičeních je Matlab. Pokud vám nevyhovuje, je možno (na vlastní nebezpečí) použít jiný software - např. WEKA, RapidMiner, Mathematica, jazyk R.

4. Volba algoritmu
Záleží na charakteru dat (problému). Pro klasifikační problémy (kategorický výstup) máte na výběr z následujících algoritmů:

- klasifikace pomocí nejbližších sousedů (kNN),
- lineární a polynomiální separace,
- rozhodovací stromy.

Pro regresní problémy (výstup je reálné nebo celé číslo) si můžete vybrat algoritmus

- nejbližších sousedů (kNN),
- lineární a polynomiální regrese,
- regresní rozhodovací strom.

Všechny algoritmy jsou již naimplementované ve formě Matlabovských skriptů (funkcí) a najdete je v materiálech na cvičení 5, 6.

5. Experimenty s parametrizací algoritmů
Proveďte experimenty, které jasně demonstrují, jak nastavit zvolený algoritmus pro vaše data. Kritériem pro nastavení parametrů je co nejlepší generalizace produkovaných klasifikátorů (modelů). U algoritmů budete nastavovat následující parametry:

- kNN - počet nejbližších sousedů (k)
- polynomiální regrese a separace - stupeň polynomu (deg)
- rozhodovací strom - minimální počet dat pro další dělení uzlu (splitmin)

Pokud graf závislosti generalizace modelu na jeho plasticitě obsahuje hodně šumu (i přes použití křížové validace), je třeba experimenty opakovat a data pro finální graf zprůměrovat např. ze 100 měření.

6. Report
Výstupem vaší úlohy bude report (max 2 stránky), kde popíšete zvolená data (problém), jejich předzpracování, volbu algoritmu, výsledky parametrizace algoritmu (formou grafů) **s vaší interpretací**, diskutujete výsledky a vyslovíte závěr. Použijte předpřipravenou šablonu.

Hodnocení:

Úloha má dotaci 20b, je však k dispozici poměrně dost bonusových bodů, které můžete získat za nadstandardní práci (definováno dále).

20 bodů dostanete za správně vyřešenou a kvalitně zdokumentovanou úlohu.

Hlavní kritéria pro hodnocení:

- **vlastní interpretace výsledků** (každý graf a tabulka náležitě okomentována a konsekvence diskutovány)
- **reprodukovatelnost** (uvést nastavení experimentů, aby se podle popisu daly zopakovat)
- **správnost** (v experimentech nejsou zjevné metodologické chyby)

Formální stránka se také hodnotí, dávejte si pozor na správné uvedení referencí.

Nadstandardní práce:

- porovnání různých algoritmů
- různé velikosti trénovací množiny
- křížová validace pro různé počty foldů
- nadstandardní předzpracování - kódování, transformace, čištění dat, ...
- nadstandardní vyhodnocení - ROC, ...
- a další zajímavé experimenty
- bonusové body lze udělit i za použití latexu (2b) a čitelné angličtiny (3b)