Zadání semestrální práce předmětu DVZ ZS 2012/13
Téma: onemocnění prostaty

Ačkoli v důsledku rakoviny prostaty umírá ve světě méně než 5% mužů, bývají obtíže spojené s tímto onemocněním vážné a problémy s prostatou jsou tak častým tématem vědeckých i společenských polemik. Jasno není například v otázce, zda potenciální včasná diagnostika nemoci vyváží problémy spojené se samotným diagnostickým testem a následnou léčbou. Jedním z cílů tohoto projektu je proto modelování vztahu mezi velikostí nádoru a hladinou antigenu specifického pro prostatu, a zejména pak identifikace faktorů ovlivňujících závažnost choroby. Na vzorku dat získaných sledování několika příznaků na pacientech podstupujících resekci prostaty se budeme snažit chorobu diagnostikovat, tedy na základě příznaků co nejpřesněji rozlišit závažnost choroby, a včasnou diagnostikou tak zvýšit šanci na úspěšnou léčbu.

Soubor `prostate.csv` obsahuje data (matici pozorování), v souboru `prostate.names` pak naleznete stručný popis jednotlivých příznaků.

Vaším úkolem je strojově klasifikovat pozorování (rozlišit pacienty s pokročilou chorobou od pacientů s chorobou v zárodku), modelovat velikost nádoru v závislosti na hladině specifického antigenu, a pokusit se identifikovat vnitřní strukturu v datech.

Konkrétně:

1. Průzkumná analýza.
   - Prohlédněte si data.
   - Obsahují data nějaké chybějící hodnoty? Pokud ano, jak se s nimi vypořádáte?
   - Jsou v datech nějaká odlehlá pozorování? Pokud ano, jak se s nimi vypořádáte?
   - Je třeba data pro účely následujícího zpracování transformovat? Pokud ano, jak?
   - S ohledem na následující úkoly vizualizujte vybrané příznaky / vazby mezi příznaky / vazby mezi příznaky a typem nádoru.
2. Klasifikace pacientů podle toho, zda (ne)mají zasažené semenné váčky (seminal vesicle invasion, `svi`)
   - Vyberte alespoň dva klasifikátory vhodné pro tento typ dat. Podle čeho budete vybírat? Volbu zdůvodněte a fakticky podložte (obrázkem, číselně apod.).
   - Na vhodné podmnožině příznaků vybrané klasifikátory natrénujte. Jaké úspěšnosti dosahují? Podle jakých příznaků se rozhodují?
   - Jakou úspěšnost klasifikace očekáváte na nezávislých datech? Vyjádřete číselně. Mohli byste dát i nějakou záruku úspěšnosti takové budoucí klasifikace?
   - Jaká je pravděpodobnost, že nově příchozí pacient bude trpět zasažením semenných váčků?
   - Jaká je pravděpodobnost, že nově příchozí pacient trpící zasažením váčků bude klasifikován jako nezažený? Jaká bude naproti tomu pravděpodobnost, že nově příchozí pacient s nezasaženými váčky bude klasifikován jako zasažený? Kterou chybu považujete za závažnější a proč?

- Má ve vašem případě na přesnost klasifikace vliv to, zda je trénovací (testovací) množina vyvážená? Pokud ano, jaký? Vyvažovali jste trénovací (testovací) množinu? Pokud ano, proč a jak? Pokud ne, proč ne?

3. Modelování vztahu velikosti nádoru (cancer volume, `cv`) na hladině specifického antigenu (prostate-specific antigen, `psa`).

   - Vhodnou metodou modelujte danou závislost a formálně ji zapište. Je závislost významná? Výsledky interpretujte.

4. Hledání charakteristických vzorů v datech.

   - Pokuste se v datech na základě hodnot příznaků (tj. bez zohlednění toho, zda se jedná o pacienta se zasaženými či nezasaženými semennými váčky) nalézt vnitřní strukturu. Výsledky vizualizujte a interpretujte.
   - Je možno dát nalezenou strukturu do souvislosti s rozlišením osob (na zasažené a nezasažené)? (Tj. nelze rozdělení na osoby zasažené a nezasažené tušit již přímo v nalezené struktuře? Jinými slovy: neodpovídají nalezené podskupiny do značné míry pacientům se zasaženými/nezasaženými váčky?) Vizualizujte a/nebo popište tabulkou.

Svojí práci shrňte ve formě vědeckého sdělení ve formátu PDF. Buďte struční, avšak úplní. Pište jen to, co sami považujete za důležité, avšak tak, aby byly vaše kroky jasné i čtenáři, který neabsolvoval předmět DVZ. Zdůvodňujte učiněná rozhodnutí. Buďte konzistentní (zavedete-li nějaké termíny, používejte je v celém textu; nemíchejte desetinné čárky a tečky).

Text by měl obsahovat abstrakt, úvod, metodiku, výsledky a závěr.

V **abstraktu** velmi stručně shrňte celý text. Jednou nebo pár větami vystihněte každou následující část textu. (Abstrakt se píše na závěr.)

V **úvodu** popište, čeho se váš projekt týká, proč jej řešíte, co od něj očekáváte.

V **metodách** popište, jaká data máte k dispozici (včetně informace o tom kolik instancí a příznaků se v datech nachází, zda se v datech vyskytují chybějící a/nebo odlehlé hodnoty) a jakým způsobem budete postupovat (jaké metody použijete a proč). Diskutujte možné problémy, očekáváte-li nějaké, a navrhněte možná řešení.

Výsledky by pak měly shrnout, k čemu jste dospěli. Vyberte relevantní výsledky, není třeba čtenáře zahlcovat přílišnými detaily či marginálními výsledky. Ve výsledcích by se měly objevit přesné odpovědi na všechny položené otázky, a to buď přímo v textu, nebo jako obrázek či tabulka.

Závěr by měl diskutovat dosažené výsledky (co výsledky znamenají? jak je lze využít?) a stručně shrnout celý projekt.

Dbejte na grafické zpracování. Z obrázků i tabulek by mělo být na první pohled jasné, co vyjadřují. V obrázcích nesmí chybět popis os. Jsou-li v obrázku textové popisy, měly by být jasné a čitelné. Označujete-li si pracovně příznaky např. jako AT1, AT2 apod., nepatří tyto pracovní názvy rozhodně do publikačního výstupu. Na obrázky i tabulky je dobré se odkázat z textu (pomocí čísel).

V neposlední řadě se hodnotí i gramatická a stylistická správnost textu.