Zadání semestrální práce
Téma: kriminalita

Cíl

Cílem této práce je analyzovat data popisující kriminalitu v jednotlivých státech USA. Vaším úkolem bude:

- odlišit jižní státy od severních,
- modelovat kriminalitu v závislosti na nerovnosti v příjmech,
- pokusit se nalézt vnitřní strukturu v datech.

Data

Data byla agregována v roce 1960 ve 47 státech USA za účelem studia faktorů ovlivňujících kriminalitu. U občanů každého státu byla zjištěna pravděpodobnost, že budou uvězneni, jejich vzdělání, příjmy, etnické složení a další příznaky (viz Tabulka 1). *Pozor: hodnoty některých příznaků byly při sběru transformovány (přeškálovány), a tak nelze u všech příznaků spoléhat na jejich absolutní interpretovatelnost. Předpokládejme však, že přeškálování sestávalo z přičtení konstanty a vynásobení jinou konstantou, takže relativní interpretovatelnost by měla být zachována.*

Data jsou k dispozici v souboru `crime.csv`.

`M`	procento mužů ve věku 14-24 let
`So`	indikátor jižního státu
`Ed`	průměrný počet let školní docházky
`Po1`	výdaje na policejní systém v roce 1960
`Po2`	výdaje na policejní systém v roce 1959
`LF`	index zapojení do pracovního procesu
`M.F`	počet mužů na 1000 žen
`Pop`	populace
`NW`	počet obyvatel jiných než bělochů (na 1000 obyvatel)
`U1`	index venkovské nezaměstnanosti u mužů ve věku 14-24 let
`U2`	počet venkovské nezaměstnanosti u mužů ve věku 35-39 let
`GDP`	hrubý národní produkt (na hlavu)
`Ineq`	příjmová nerovnost
`Prob`	pravděpodobnost uvěznění
`Time`	průměrný čas strávený ve vězení
`y`	index výskytu zločinu určité kategorie (na hlavu)

Tabulka 1: Popis naměřených příznaků.

Požadované kroky analýzy

1. Průzkumná analýza. **[5 bodů]**
   - Kolik máte k dispozici dat (kolik států, kolik příznaků)?
   - 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 resp. vztahy mezi příznaky resp. vztahy mezi příznaky a indikátorem jižního státu.

2. Klasifikace států na jižní a severní. **[10 bodů]**

   - Vyberte alespoň dva klasifikátory vhodné pro tento typ dat. Podle čeho budete vybírat?
   - Na vhodné podmnožině dat vybrané klasifikátory natrénujte. Jaké úspěšnosti dosahují? Podle jakých příznaků se rozhodují? Dává to smysl?
   - Jakou úspěšnost klasifikace očekáváte v hypotetickém reálném nasazení Vašeho klasifikátoru, tj. v případě klasifikace dalšího státu (který nemáte v datovém souboru)?
   - Na základě dat, která máte k dispozici, odhadněte pravděpodobnost, že takový další klasifikovaný stát bude jižní?
   - Jaká je pravděpodobnost, že další klasifikovaný stát, který bude ve skutečnosti jižní, bude klasifikován jako severní? Jaká bude naproti tomu pravděpodobnost, že další klasifikovaný stát, který bude severní, bude klasifikován jako jižní? 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 závislosti pravděpodobnosti uvěznění (`Prob`) na nerovnosti v příjmech (`Ineq`). **[3 body]**

   - Vhodnou metodou danou závislost modelujte a formálně ji zapište. Je závislost statisticky významná? Výsledek interpretujte.
   - Existuje i souvislost mezi pravděpodobností uvěznění a jinými příznaky? Pokud ano, interpretujte je. Můžete se pokusit i o modelování (vysvětlení) pravděpodobnosti uvěznění pomocí více příznaků.

4. Hledání podskupin v datech. **[4 body]**

   - Pokuste se v datech *pouze na základě hodnot příznaků* (tj. bez zohlednění toho, zda se jedná o severní či jižní stát) nalézt vnitřní strukturu (podskupiny). Zamyslete se, zda je třeba data normalizovat. Výsledky vizualizujte a interpretujte. Lze popsat nalezené shluky např. typickými/průměrnými reprezentanty?
   - Je možno dát nalezenou strukturu do souvislosti s rozlišením států na severní a jižní? (Jinými slovy: neodpovídají nalezené podskupiny do značné míry právě severním a jižním státům?) Vizualizujte a/nebo popište tabulkou.

Zpráva o řešení

Svoji práci shrňte ve formě sdělení ve formátu PDF. Kromě věcné stránky se bude hodnotit i forma textu **[8 bodů]**[1]. Buďte struční, avšak úplní. Váš postup popište tak, aby byl srozumitelný i čtenáři, který neabsolvoval předmět DVZ. Snažte se v textu odpovědět na všechny položené otázky, zdůvodňujte učiněná rozhodnutí. Text by měl obsahovat abstrakt, úvod, metodiku, výsledky a závěr. V *abstraktu* velmi stručně shrňte celý text. Jednou nebo pár větami vystihněte každou následující

[1] Celkově tak můžete za vypracování semestrální práci získat 30 bodů (za její prezentaci pak dalších 10 bodů).

část textu. (Abstrakt se píše zpravidla až na závěr.) V *úvodu* stručně popište, čeho se váš projekt týká, co je vaším úkolem a co od své práce 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 aplikací zvolených metod na data. 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 i na grafické zpracování. Z obrázků i tabulek by mělo být na první pohled jasné, co vyjadřují. V obrázcích nesmí chybět rozumný popis os (např. nikoli `attr1`, ale *hmotnost [kg]*). Jsou-li v obrázku textové popisy, měly by být jasné a čitelné. Označujete-li si pracovně příznaky např. jako `attr1`, `attr2`, nepatří tyto pracovní názvy rozhodně do publikačního výstupu. Na obrázky i tabulky je dobré se odkázat z textu (pomocí čísel). Buďte konzistentní (zavedete-li nějaké termíny či zkratky, používejte je v celém textu; nemíchejte desetinné čárky a tečky).

V neposlední řadě se hodnotí i pravopisná správnost textu.