Zadání semestrální práce
Téma: auta

Cíl

Cílem této práce je analyzovat data o cenách a technické specifikaci osobních aut. Vaším úkolem bude:

- pomocí technických údajů o autech odlišit auta vyrobená v USA a mimo USA,
- modelovat cenu auta v závislosti na objemu nádrže ;-),
- pokusit se nalézt vnitřní strukturu v datech.

Data

Data byla nasbírána v roce 1993 a pokrývají charakteristiky a cenu vybraných 93 osobních aut. U každého auta byla zjišťována cena, spotřeba, výkon, hmotnost, rozměry a další charakteristiky (viz Tabulka 1).

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

`Manufacturer`	výrobce
`Model`	model auta
`Type`	typ auta
`Min.Price`	minimální cena (v $1,000): cena za "základní" verzi
`Max.Price`	maximální cena (v $1,000): cena za "premiovou" verzi
`Price`	průměrná cena (průměr `Min.Price` a `Max.Price`)
`MPG.city`	údaj o spotřebě - kolik mil lze ujet na galon paliva ve městě
`MPG.highway`	údaj o spotřebě - kolik mil lze ujet na galon paliva na dálnici
`AirBags`	má auto airbagy? Faktor o úrovních: `none`, `driver only`, a `driver & passenger`.
`DriveTrain`	typ náhonu: `rear wheel`, `front wheel` nebo `4WD`
`Cylinders`	počet válců (údaj chybí u "Mazda RX-7", která má rotační motor)
`EngineSize`	objem motoru
`Horsepower`	maximální výkon
`RPM`	počet otáček za minutu (při maximálním výkonu)
`Rev.per.mile`	počet otáček na jednu míli (při maximálním rychlostním stupni)
`Man.trans.avail`	má auto ruční převodovku?
`Fuel.tank.capacity`	kapacita nádrže (v US galonech)
`Passengers`	maximální počet cestujících
`Length`	délka auta (v palcích)
`Wheelbase`	rozvor kol (v palcích)
`Width`	šířka auta (v palcích)
`Turn.circle`	prostor k otočení o 180 stupňů (ve stopách).
`Rear.seat.room`	velikost prostor u zadních sedadel (v palcích) (chybí u 2-sedadlových aut)
`Luggage.room`	velikost zavazadlového prostoru (v kubických stopách) (chybí u dodávek)
`Weight`	hmotnost (v librách)
`Origin`	indikátor, zda je výrobce auta z USA nebo mimo USA

Tabulka 1: Popis naměřených příznaků.

Požadované kroky analýzy

1. Průzkumná analýza. **[5 bodů]**
   - Kolik máte k dispozici dat (kolik aut, 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, zda bylo auto vyrobeno v USA nebo mimo USA.
2. Klasifikace aut podle místa výroby (USA, mimo USA). **[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 auta (které nemáte v datovém souboru, ale u kterého máte naměřeny příslušné příznaky)?
   - Za předpokladu, že výběr aut v datovém souboru je reprezentativní, odhadněte pravděpodobnost, že další klasifikované auto bylo vyrobeno v USA.
   - Jaká je pravděpodobnost, že další klasifikované auto, ač vyrobené v USA, bude klasifikováno jako vyrobené mimo USA? Jaká bude naproti tomu pravděpodobnost, že další klasifikované auto, vyrobené mimo USA, bude klasifikováno jako vyrobené v USA? 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 ceny auta (`Price`) na objemu nádrže (`Fuel.tank.capacity`). **[3 body]**
   - Vhodnou metodou danou závislost modelujte a formálně ji zapište. Je závislost statisticky významná? Výsledek interpretujte.
   - Existuje i souvislost mezi cenou a jinými příznaky[1]? Pokud ano, interpretujte je. Můžete se pokusit i o modelování (vysvětlení) ceny auta pomocí více příznaků. Změní se relace mezi kapacitou nádrže a cenou auta po zohlednění jiných příznaků?
4. Hledání podskupin v datech. **[4 body]**
   - Pokuste se v datech *pouze na základě hodnot příznaků* (tj. bez uvažování toho, kde bylo auto vyrobeno) 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 tím, kde bylo auto vyrobeno? Vizualizujte a/nebo popište tabulkou.

[1] nápověda: uvažte, co může cenu auta ovlivňovat, a zaměřte se jen na takové příznaky

Zpráva o řešení

Svoji práci shrňte ve formě sdělení ve formátu PDF. Kromě věcné stránky se bude hodnotit i forma textu [**8 bodů**][2]. 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 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.

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