Zadání semestrální práce

Cíl

Cílem této práce je analyzovat data kardiologických pacientů z Cleveland Clinic.
Vaším úkolem bude:

- odlišit pacienty s kardiologickým onemocněním a pacienty zdravé
- modelovat závislost maximální dosažené tepové frekvence na věku pacienta
- pokusit se nalézt vnitřní strukturu v datech

Data

Heart disease data set je soubor dat, který se stal standardním testovacím souborem pro algoritmy strojového učení. Data byla naměřena v Cleveland Clinic Foundation v roce 1988 a od té doby jsou stále objektem zájmu mnoha výzkumníků. Pacienti kardiologické kliniky byli podrobeni sérii testů, přičemž z původních 76 naměřených příznaků se používá pouze 14 pro klasifikaci pacientů na zdravé a nemocné. Popis těchto příznaků je uveden v Tabulce 1. Data jsou v souboru *heart.csv*.

age	Věk v letech
sex	Pohlaví 1 = muž, 0 = žena
cp	Typ bolesti na hrudi: 1 = typická angina, 2 = atypická angina, 3 = neanginální bolest, 4 = asymptomatická
trestbps	Klidový krevní tlak (v mmHg)
chol	Sérový cholesterol v mg/dl
fbs	Krevní cukr nalačno > 120 mg/dl: 1 = true, 0 = false
restecg	Výsledek klidového EKG: 0 = normal, 1 = abnormalita ST-T vlny, 2 = ventrikulární hypertrofie
thalach	Maximální tepová frekvence
exang	Bolest na hrudi vyvolaná cvičením: 1 = yes, 0 = no
oldpeak	ST deprese vyvolaná cvičením vztažená ke klidovému stavu
slope	Strmost píku: 1 = stoupající, 2 = neměmnící se, 3 = klesající
ca	Počet hlavních cév označených pomocí fluoroskopie
thal	3 = normální, 6 = opravený defekt, 7 = reversibilní defect
dis	Diagnóza: 0 = zdravý, 1 = nemocný

Tabulka 1: Stručný popis příznaků

Požadované kroky analýzy

1. Průzkumová analýza. [5 bodů]

- Kolik máte k dispozici dat (kolik pacientů, kolik příznaků)?
- Obsahují data nějaké chybějící hodnoty? Pokud ano, jak se s nimi vypořádáte?
- Je třeba data pro účely jejich zpracování nějak transformovat? Pokud Ano, jak?
- S ohledem na následující úkoly vizualizujte vybrané příznaky, vztahy mezi příznaky a vztahy mezi příznaky a klasifikací pacienta

2. Klasifikace pacientů na zdravé a s kardiologickým onemocněním. [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 nově příchozího pacienta?

- Jaká je pravděpodobnost na základě dat, že nově příchozí pacient bude zdravý?
- Jaká je pravděpodobnost, že nově příchozí pacient bude klasifikovaný jako nemocný? Jaká bude naproti tomu pravděpodobnost, že nově příchozí nemocný pacient bude klasifikován jako zdravý.
- 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 maximální dosažené tepové frekvence na věku. [3 body]

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

4. Hledání podskupin v datech. [4 body]

- Pokuste se v datech *pouze na základě hodnot příznaků* (tj. Bez zohlednění toho, zda se jedná o zdravého nebo nemocného pacienta) nalézt vnitřní strukturu (podskupiny). Zamyslete se, zda je vhodné data normalizovat. Výsledky vizualizujte a interpretujte. Lze popsat nalezené shluky např. Typickými/průměrnými reprezentanty?
- Je možné dát nalezenou strukturu do souvislosti s klasifikací pacientů na zdravé a nemocné? (Jinými slovy: Odpovídají nalezené podskupiny do značné míry zdravým a nemocným pacientům?) 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]. 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 zpravidla píše až na závěr.) V **úvodu** stručně popište, čeho se váš projekt týká, co je Vaším úkolem a co od své práce očekáváte. V **metodách** stručně 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 odlehlá pozorování) 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. attr1, attr2, nepatří tyto pracovní názvy rozhodně do publikačního výstupu. Na obrázky i tabulky je dobré se odkázat v 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.