Zadání semestrální práce

Cíl

Cílem této práce je analyzovat data o vysokoenergetických nárazech v polském dole. Vaším úkolem bude:

- predikovat nebezpečné situace v dole
- modelovat závislost počtu nárazů naměřených pomocí nejaktivnějšího geofonu na seismické energii naměřené pomocí nejaktivnějšího geofonu
- pokusit se nalézt vnitřní strukturu v datech

Data

Dolování uhlí je spojeno s velkým rizikem. Jedním z takových rizik je seismické riziko, které je v tomto ohledu srovnatelné se zemětřesením. Seismické riziko je velmi obtížně predikovatelné. V datech jsou uvedené celkové počty seismických nárazů během jedné směny a příznaky popisující hodnocení rizika. Stručný popis příznaků je v Tabulce 1. Data jsou v souboru *bumps.csv*.

seismic	zhodnocení rizika seismického rizika pomocí seismické metody (a - bez rizika, b - nízké riziko, c - vysoké riziko, d - nebezpečný stav)
seismoacoustic	zhodnocení seismického rizika pomocí seismoakustické metody
shift	typ směny (W - dolování uhlí, N -přípravná směna);
genergy	seismická energie naměřená předchozí směnou z nejaktivnějšího geofonu
gpuls	počet pulsů naměřený předchozí směnou z nejaktivnějšího geofonu
gdenergy	odchylka energie od průměru energie z 8 předchozích směn
gdpuls	odchylka počtu pulsů od průměru energie z 8 předchozích směn
ghazard	výsledek zhodnocení rizika seismického rizika pomocí seismoakustické metody (pouze GMax)
nbumps	počet seismických nárazů naměřených během předchozí směny
nbumps2	počet seismických nárazů (v rozsahu energie [10^2,10^3)) zaznamenených během předchozí směny
nbumps3	počet seismických nárazů (v rozsahu energie [10^3,10^4))
nbumps4	počet seismických nárazů (v rozsahu energie [10^4,10^5))
nbumps5	počet seismických nárazů (v rozsahu energie [10^5,10^6))
nbumps6	počet seismických nárazů (v rozsahu energie [10^6,10^7))
nbumps7	počet seismických nárazů (v rozsahu energie [10^7,10^8))
nbumps89	počet seismických nárazů (v rozsahu energie [10^8,10^10))
energy	celková energie seismických nárazů během předchozí směny
maxenergy	maximum energie seismických nárazů během předchozí směny
class	rozhodovací atribut - '1' nebezpečný stav, '0' bezpečný stav

Tabulka 1: Stručný popis příznaků

Požadované kroky analýzy

1. Průzkumová analýza. [5 bodů]

- Kolik máte k dispozici dat (kolik směn, 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 rizikem při směně

2. Klasifikace směn na rizikové a bezpečné. [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 nové směny?
- Jaká je pravděpodobnost na základě dat, že nová směna bude bezpečná?
- Jaká je pravděpodobnost, že nová směna bude klasifikovaná jako nebezpečná? Jaká bude naproti tomu pravděpodobnost, že nová skutečně nebezpečná směna bude klasifikovaná jako bezpečná?
- 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 počtu nárazů naměřených pomocí nejaktivnějšího geofonu na seismické energii naměřené pomocí nejaktivnějšího geofonu. [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 bezpečnou a nebo nebezpečnou směnu) 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 bezpečností směny? (Jinými slovy: Odpovídají nalezené podskupiny do značné míry bezpečným a nebezpečným směnám?) 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.