Zadání semestrální práce

Cíl

Cílem této práce je analyzovat data odvozená ze snímků listů stromů, keřů a bylin.
Vaším úkolem bude:

- odlišit stromy, keře a byliny
- modelovat závislost prodloužení listu na základě jeho výstřednosti
- pokusit se nalézt vnitřní strukturu v datech

Data

Klasifikace rostlin byla tradičně úkolem specializovaných biologů – taxonomů. V současné době je nedostatek taxonomů, zatímco výpočetní technika dospěla do stavu, kdy je možné automatizovat i takovýto komplexní úkol. Pro účely vývoje automatického systému pro klasifikaci listů rostlin na základě tvaru jejich listů byla vytvořena databáze fotografií listů 40 rostlin. Na základě fotografií listů bylo odvozeno několik parametrů popisujících tvar a texturu listu viz Tabulka 1. Data jsou k dispozici v souboru *leaves.csv.*

Class	Indikátor druhu rostliny
Specimen_Number	Indikátor vzorku
Eccentricity	Výstřednost listu
Aspect_Ratio	Poměr stran listu
Elongation	Prodloužení listu
Solidity	Míra konvexity listu
Stochastic_Convexity	Stochastická míra konvexity listu
Isoperimetric_Factor	Míra členitosti obvodu tvaru listu
Maximal_Indentation_Depth	hloubka největšího zářezu v listu
Lobedness	Laločnatost listu
Average_Intensity	Průměrná intenzita obrazu
Average_Contrast	Průměrný kontrast
Smoothness	Míra „hladkosti“ intenzit obrazu
Third_Moment	Míra šikmosti histogramu intenzit obrazu
Uniformity	Míra jednolitosti intenzit obrazu
Entropy	Míra náhodnosti intenzit obrazu
Species	Latinské jméno rostliny
Type	Indikátor strom – keř – bylina

Tabulka 1: Stručný popis příznaků

Požadované kroky analýzy

1. Průzkumová analýza. [5 bodů]

- Kolik máte k dispozici dat (kolik druhů rostlin, 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 typem rostliny

2. Klasifikace rostlin na stromy, keře a byliny. [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 vyfoceného listu nějaké rostliny?
- Jaká je pravděpodobnost na základě dat, že nově vyfocený list bude utržený ze stromu?
- Jaká je pravděpodobnost, že nově vyfocený list bude klasifikovaný jako list keře? Jaká bude naproti tomu pravděpodobnost, že nově vyfocený list keře bude klasifikován jako list stromu nebo byliny.
- 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 prodloužení listu na základě jeho výstřednosti. [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 strom, keř nebo bylinu) 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 rostlin na stromy, keře a byliny? (Jinými slovy: Odpovídají nalezené podskupiny do značné míry stromům, keřům a bylinám?) 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.