Cvičení z předmětu 17BIEHT - Automatická klasifikace

Zadání

Cílem cvičení je vyzkoušet si základní použití automatické klasifikace, včetně předzpracování a testování. Každý student má k dispozici dva datové soubory datax_tr a datax_ts, které obsahují trénovací a testovací data set a v jejichž názvu x reprezentuje číslo datového souboru (viz. přiřazení na konci tohoto zadání). Soubory obsahují numerické hodnoty oddělené čárkou, kde první dva sloupce odpovídají vstupním příznakům a poslední sloupec označuje třídu, do které každý řádek patří.

Během cvičení proveďte následující, výsledek předveďte a odpověďte na dotazy:

1. Naimplementujte klasifikátor 1-nearest neighbor (1NN). Tento jednoduchý klasifikátor můžete implementovat sami nebo můžete použít i libovolné pattern recognition knihovny ve vámi preferovaném programovacím prostředí.
2. Klasifikátor naučte na trénovacích datech a stanovte klasifikační přesnosti na trénovacích i testovacích datech.
3. Zjistěte průměrné hodnoty příznaků a data si vykreslete ve 2D prostoru příznaků (tzv. scatter plot). Co zajímavého pozorujete?
4. Z trénovacích data odstraňte tzv. outliers. Mohli byste použít nějaké vámi nalezené metody. V tomto modelovém příkladu ale víme, že outliers jsou první čtyři instance v trénovací množině.
5. Zjistěte průměrné hodnoty příznaků a data si vykreslete ve 2D prostoru příznaků. Co se změnilo?
6. Klasifikátor naučte na trénovacích datech a stanovte klasifikační přesnosti na trénovacích i testovacích datech. Jaký mělo odstranění outlierů vliv?
7. Normalizujte data tak, aby každý příznak měl nulovou střední hodnotu a jednotkovou standardní odchylku.
8. Zjistěte průměrné hodnoty příznaků a data si vykreslete ve 2D prostoru příznaků. Co se změnilo?
9. Klasifikátor naučte na trénovacích datech a stanovte klasifikační přesnosti na trénovacích i testovacích datech. Jaký měla normalizace dat vliv?

V případě, že nestihnete úlohu během cvičení, je nutné odevzdat odpovědi na všechny výše uvedené otázky. Navíc ještě musíte udělat totéž, ale s jiným (vámi vybraným) klasifikátorem, který není typu nejbližší soused. Hotové řešení obsahující odpovědi na otázky a kód řešení v libovolném programovacím prostředí v takovém případě odevzdejte přes upload systém do 19. 5. 2017.

Přiřazení datových souborů studentům:

X	Jméno
1	Čumrdová
2	Dudař
3	Dušek
4	Jirsa
5	Mařík
6	Štítová
7	Zoul