Úkol C – Klasifikace dat – Pima Indians

LS 2023/2024

Cíl

Cílem třetího úkolu je seznámit se s praktickým použitím klasifikačních metod na příkladu klasifikace pacientů s diabetem.

Data

Indiáni z kmene Pima žijí ve střední a jižní Arizoně, USA. Zajímavé je, že se u nich vyskytuje vůbec nejvyšší naměřená prevalence obezity a diabetu. Protože byli ochotní podílet se na výzkumu, tak the National Institute of Diabetes and Digestive and Kidney byl schopen naměřit tato celkem neobvyklá biometrická data. V této studii vystupují pouze ženy.

Klasifikace subjektů s diabetem

Na základě dat vytvořte klasifikátor, který bude predikovat, zda pacient trpí diabetem. Klasifikátor zhodnoťte z různých praktických i teoretických hledisek.

Požadované kroky analýzy [20b]

- Vyberte alespoň dva klasifikátory vhodné pro tento typ dat. Podle čeho budete vybírat? [2 b]
- Na vhodné podmnožině dat vybrané klasifikátory natrénujte. Jaké úspěšnosti dosahují? [2 b]
- Podle jakých příznaků se klasifikátor rozhoduje? Dává to smysl? Lze na základě vaší analýzy omezit počet měřených příznaků při zachování stejné úspěšnosti klasifikace? [4 b]
- Jakou úspěšnost klasifikace očekáváte v hypotetickém reálném nasazení vašeho klasifikátoru, tj. v případě nově příchozího pacienta? [4 b]
- Jaká je pravděpodobnost na základě dat, že nový pacient bude nemocný? [2 b]
- Jaká je pravděpodobnost, že zdravý pacient bude klasifikován jako nemocný? Jaká bude naproti tomu pravděpodobnost, že nemocný pacient bude klasifikován jako zdravý? Výsledky diskutujte. [4 b]
- 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? [2 b]

Výsledky upravte do formy prezentace (Powerpoint nebo PDF), která bude obsahovat především stručný úvod, popis metod, které jste použili, výsledky jejich aplikace na data a závěry, které jste zjistili interpretací výsledků. Tyto prezentační slidy vám budou sloužit jako podklady pro prezentaci vašich výsledků úlohy C, která se bude konat během sloučeného bloku přednášky a cvičení v posledním týdnu semestru. Hodnocení se skládá jak z odevzdaných podkladů, tak i vašeho vystoupení. Maximální počet slidů není omezen, na prezentaci však budete mít max. 5 minut. Prezentaci ve formátu *ppt*, *pptx*, nebo *pdf* a zdrojový R skript odevzdejte pomocí *Moodle*.

Název parametru	Popis parametru
Pregnant	Kolikrát byla žena za svůj život těhotná
Glucose	Koncentrace glukózy v plazmě po 2 hodinách při orálním testu glukózové tolerance
Diastolic	Diastolický krevní tlak
Triceps	Odhad tělesného tuku (mm – tloušťka podkožního tuku). Měří se na pravé paži.
Insulin	Koncentrace sérového inzulinu po 2 hodinách testu (mu U/ml).
Bmi	Body Mass Index
Diabetes	Indikátor historie diabetu v rodině
Age	Věk
Test	Výstupní veličina. ‚negativ' – pacient je zdravý, ‚positif' – pacient trpí cukrovkou.

Tabulka 1: Výčet parametrů