Úkol C – Klasifikace dat – Pima indians diabetes

ZS 2016

Cíl

Cílem třetího úkolu je seznámit se s praktickým použitím klasifikačních metod na příkladě klasifikace indiánek kmene Pima podle diabetu.

Data

Vzorek populace žen, které byly přinejmenším 21 let staré, pocházely z indiánského kmene Pima a žily poblíž Phoenixu (Arizona), byl testován na diabetes podle kritérií WHO. Data byla nasbírána Národním ústavem diabetu a nemocí trávicího traktu a ledvin Spojených států. Ve srovnání s daty z populací indiánek kmene Pima z Mexika mohou data posloužit k posouzení životního stylu na rozvoj diabetu. Samotná data ze Spojených států mohou napomoci vytipovat příznaky, které ovlivňují rozvoj diabetu.

Klasifikace diabetu u indiánek

Pokuste se vytvořit takový na základě předložených dat klasifikátor, který by predikoval, zda indiánky trpí diabetem. Klasifikátor zhodnoťte z různých praktických i teoretických hledisek.

Požadované kroky analýzy

- 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 vyšetřené indiánky? [4 b]
- Jaká je pravděpodobnost na základě dat, že nově vyšetřená indiánky bude trpět diabetem? [2 b]
- Jaká je pravděpodobnost, že nově vyšetřená indiánka bude klasifikovaná jako trpící diabetem? Jaká bude naproti tomu pravděpodobnost, že nově vyšetřená indiánka bude klasifikovaná jako zdravá. 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 zprávy, která bude obsahovat stručný **úvod**, popis metod, které jste použili, v sekci **metody,** výsledky jejich aplikace na data v sekci **výsledky** a závěry, které jste zjistili interpretací výsledků v sekci **závěr**. Maximální délka zprávy je 3 stránky. Zprávu ve formátu pdf odevzdejte pomocí UploadSystemu.

no_preg	Počet těhotenství
gc_conc	Koncentrace glukózy v plazmě
dbp	Diastolický krevní tlak [mmHg]
skin_fold	Tloušťka kožní řasy na tricepsu [mm]
insulin	Koncentrace inzulinu v plazmě
bmi	Body mass index
dm_gen	Riziko diabetu z hlediska rodinné historie
age	Věk [roky]
class	Diabetes mellitus

Tabulka 1: Popis proměnných z studie indiánek kmene Pima