

Klasifikace aut podle místa výroby

Cíl

Cílem tohoto úkolu je nalézt model, který bude klasifikovat auta podle místa výroby a odhadnout chování tohoto modelu na reálných datech.

Data

Data byla nasbírána v roce 1993 a pokrývají charakteristiky a cenu vybraných 93 osobních aut. U každého auta byla zjišťována cena, spotřeba, výkon, hmotnost, rozměry a další charakteristiky. Kompletní data jsou k dispozici opět v souboru *cars.csv*.

Požadované kroky analýzy:

- Vyberte alespoň dva klasifikátory vhodné pro tento typ dat. Svůj výběr zdůvodněte? [2 b]
- Na vhodné podmnožině dat vybrané klasifikátory natrénujte. Jaké úspěšnosti dosahují? [2 b]
- Lze pro učení klasifikátorů nějak využít znalosti, které jste získali o datech v předchozích úkolech (shluková analýza)? Pokud ano, jakým způsobem? [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 klasifikace dalšího auta (které nemáte v datovém souboru, ale u kterého máte naměřeny příslušné příznaky)? [2 b]
- Za předpokladu, že výběr aut v datovém souboru je reprezentativní, odhadněte pravděpodobnost, že další klasifikované auto bylo vyrobeno v USA. [2 b]
- Jaká je pravděpodobnost, že další klasifikované auto, ač vyrobené v USA, bude klasifikováno jako vyrobené mimo USA? Jaká bude naproti tomu pravděpodobnost, že další klasifikované auto, vyrobené mimo USA, bude klasifikováno jako vyrobené v USA? Kterou chybu považujete za závažnější a proč? [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.