**Klasifikace dat podle místa výroby**

**Milan Poláček**

# Zadání

## 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 (viz Tabulka příznaků z úplného zadání). Kompletní data jsou k dispozici 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]

# Úvod

S problematikou klasifikace dat jsem se setkal již dříve. A to jak na bakalářském, tak magisterském studiu. Ale nikdy jsme nepoužili nic jiného v úlohách než třízení dat pomocí rozhodovacího stromu.

Znalosti nejen ze cvičení mi ale stačili, abych byl schopen tuto úlohu s pomocí dostupné literatury na webu udělat.

# Metody

Dle zadání jsem si měl vybrat dva klasifikátory pro tento typ dat. Jelikož třízení dat bylo do dvou skupin, využil jsem proto dva základní klasifikátory. Prvním byl Support vector machines (dále jen zkráceně SVM) a druhým byl rozhodovací strom. Metodu SVM jsem zvolil kvůli svým předpokladům pro dělení do dvou skupin. Rozhodovací strom je jednoduchá metoda s učitelem vhodná téměř pro jakékoliv třízení.

Nejprve jsem data musel zbavit záznamů, kde chyběly příznaky. Po jejich odstranění nám zbylo 82 záznamů. A následně bylo možno tedy data rozdělit do dvou skupin na trénovací a testovací data.

Pro učení klasifikátorů jsem vybíral z příznaků dle řešení z přechozí úlohy. Experimentálně jsem snižoval počet a vybíral příznaky, u kterých jsem předpokládal, že by mohli kladně ovlivnit třízení dat.

Vybrané příznaky pro největší úspěšnost určování popisuji níže ve výsledcích. Trénování klasifikátorů jsem provedl 100x a výsledky jsem následně zprůměroval.

# Výsledky

Nejvyšší úspěšnost u metody SVM byla 62% (toto číslo je zaokrouhleno z průměrování z více pokusů). Tato úspěšnost určení zda je vozidlo z USA tedy byla při těchto zvolených příznacích a to výrobce, hmotnost vozidla, spotřeba průměrná jak ve městě, tak na dálnici, objem motoru, počet pasažéru a velikost zavazadlového prostoru a typ vozu. Metoda byla nastavena s lineárním jádrem a hodnotou *cost* 10.

Nejvyšší úspěšnost rozhodovacího stromu byla 59% (toto číslo je zaokrouhleno z průměrování z více pokusů)). Této úspěšnosti bylo dosaženo při stejných příznacích jako u předchozí metody. A byla využita metoda *class*.

Rozhodovací strom se rozhodoval nejčastěji dle hmotnosti vozidla, velikosti zavazadlového prostoru a počtu pasažérů. Protože se jedná o klasifikování mezi americkými a jinými vozy, je klasifikace podle těchto atributů patrně správná z obecně známého faktu, že vozy americké výroby jsou velké a také těžké.

Při dalším omezení počtu příznaků u obou klasifikačních metod došlo ke zvýšení chybovosti a tedy snížení zejména senzitivity klasifikace, ale také specificity.

Senzitivita pro rozhodovací strom byla 62% a specificita 64% (čísla byla zakouzlena z průměrování z více pokusů).

Senzitivita pro metod SVM byla 60% a specificita 69% (čísla byla zakouzlena z průměrování z více pokusů).

# Závěr

Bohužel se mi nepodařilo zjistit podle, kterých příznaků se klasifikuje metoda SVM, ale dá se předpokládat, že s takto vysokou úspěšností podle podobných příznaků jako rozhodovací strom.

V reálném nasazení při nastavení příznaků dle mého příkladu je možné předpokládat, že se bude úspěšnost klasifikace pohybovat přes 54%. Stále se však pohybujeme kolem úspěšnosti při odhadování, jako při odhadu jaká strana padne při hodu mincí.

Pokud byl výběr aut v datovém souboru reprezentativní, je možné předpokládat, že na zhruba 51,6% (číslo bylo zaokrouhleno) bude vozidlo americké výroby.

Další otázkou bylo, zda je závažnější nižší selektivita nebo senzitivita testu resp. zda je horší chyba 1. nebo 2. druhu. V našem případě nejde o nic závažného, ale závažnější je chyba 2. druhu, kdy lidově řečeno „nespustíme poplach, ale měli bychom“. A z toho vyplívaje nám vadí nízká senzitivita testu. Tato chyba velice závažná například v lékařství (např. odhalování rakoviny).

Jako na každý klasifikátor má velký vliv nevyváženost trénovacích dat, která ovlivňuje zejména senzitivitu testu. Trénovací data jsem se nevyvažoval, protože v reálném světě data nejsou nikdy zcela vyvážená a je patrné, že datový soubor byl patrně vyvážen.