1. zápočtová úloha z 01RAD

Jiří Franc

2020-10-13

1. zápočtová úloha z 01RAD

Popis úlohy

V tomto úkolu je cílem provést předzpracování datového souboru, jeho vizualizaci a jednoduchou lineární regresní úlohu, kde budeme modelovat spotřebu automobilu v závislosti na jeho váze. K tomuto účelu poslouží datový souboru `car_mpg_01RAD.csv`, který obsahuje 406 pozorování o 9 proměnných. Dataset byl prvně použit americkou statistickou společností v roce 1983 a lze ho též najít na UCI Machine Learning Repository, případně na kaggle.com s několika pracovními sešity.

mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin	car_name
18	8	307	130	3504	12.0	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11.0	70	1	chrysler satellite
16	8	304	150	3433	12.0	70	1	chrysler rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10.0	70	1	ford galaxie 500

Podmínky a body

Úkol i protokol vypracujte samostatně. Pokud na řešení nějaké úlohy budete přesto s někým spolupracovat, radit se, nezapomeňte to u odpovědi na danou otázku uvést. Tato zápočtová úloha obsahuje 10 otázek po 1 bodu. Celkem za 3 zápočtové úlohy bude možné získat 30 bodů, přičemž pro získání zápočtu je potřeba 20. Další dodatečné body mohu případně individuálně udělit za řešení mini domácích úkolů z jednotlivých hodin.

Odevzdání

Protokol ve formátu pdf odevzdejte prostřednictvím MS Teams, nejpozději do 15:30, 10. 11. 2020 (tj. za 3 týdny).

Předzpracování dat:

Otázka 01

Zjistěte, zdali data neobsahují chybějící hodnoty (`NA`). Pokud ano, tak rozhodněte zdali můžete příslušná pozorování z dat odstranit a proč. Které proměnné jsou kvantitativní a které kvalitativní? Jeli možno některé zařadit do obou skupin, pro kterou se rozhodnete? Které proměnné budete brát jako faktorové a proč? Spočtěte základní statistiky pro jednotlivé proměnné.

Otázka 02

Proměnnou `mpg` nahraďte proměnnou `spotreba` kde bude místo počtu ujetých mil na galon paliva uvedena hodnota počet litrů na 100 Km. Proměnnou `cylinders` přejmenujte na `pocet_valcu`. Proměnnou `displacement` přejmenujte na `zdvihovy_objem` a převeďte z kubických palců na litry. Proměnnou `horsepower` přejmenujte na `výkon` a převeďte na Watty. Proměnnou `weight` přejmenujte na `hmotnost` a převeďte z liber na kilogramy. Odstraňte proměnnou `acceleration`. Proměnnou `model.year` přejmenujte na `rok_vyroby` a upravte ji tak, aby její hodnoty popisovaly celý rok 19XX. Proměnnou `origin` přejmenujte na `puvod` a upravte ji tak, že místo 1 bude USA, místo 2 Evropa a místo 3 Japonsko. Z proměnné `car.name` vytvořte proměnnou `vyrobce` podle prvního slova obsaženého v řetězci proměnné `car.name`.

Vizualizace dat

Otázka 03

Vykreslete scatterploty pro všechny numerické proměnné. Pro proměnné `spotreba` a `hmotnost` vykreslete histogramy spolu s jádrovými odhady hustot. Pro proměnné `pocet_valcu` a `rok_vyroby` vykreslete krabicové diagramy, kde odezvou bude `spotreba`. Je z těchto grafů vidět, že některá auta mají jinou, než očekávanou spotřebu? Navrhněte úpravu těchto dvou proměných (případně úpravu datasetu) tak, aby obě proměnné `pocet_valcu` a `rok_vyroby` byly faktorové a obsahovaly právě 3 úrovně. Pro takto upravená data vykreslete místo výše zmííněných boxplotů violin ploty.

Otázka 04

Pro kombinace faktorizovaných proměnných `pocet_valcu`, `rok_vyroby` a `puvod` vykreslete spotřebu aut, aby bylo na obrázku vidět, jestli se liší spotřeba u aut pocházejících z různých kontinentů v závislosti na počtu válců, roku v7roby a naopak. Obsahují všechny kombinace relavantní množství dat?

Otázka 05

Pro auta výrobce Chrysler vykreslete závislost spotřeby na váze automobilu, kde jednotlivé události označíte barvou podle počtu válců a velikost bodů v grafu bude odpovídat objemu motoru.

Jednoduchý lineární model

Otázka 06

Sestavte jednoduchý regresní model, kde vysvětlovaná proměnná bude spotřeba automobilu. Vybrali jste model s interceptem, nebo bez a proč? Svůj výběr řádně zdůvodněte a spočtěte pro oba modely R^2 a F statistiky, co nám o modelech říkají. Na základě zvoleného modelu zjistěte, zdali spotřeba automobilu závisí na hmotnosti automobilu. Pokud ano, o kolik se změní očekávaná spotřeba automobilu pokud se jeho hmotnost zvýší o 1000kg?

Otázka 07

Sestavte obdobný model jako v předchozí otázce, ale pouze na základě dat výrobce Chrysler. Liší se tento model od předchozího? Jaký model vykazuje silnější linearní vztah mezi hmotností a spotřebou a proč? O kolik roste spotřeba s rostoucí hmotností pro vozy Chrysler rychleji než pro libovolný automobil? Spočtěte 95% konfidenční intervaly pro regresní koeficienty popisující sklon regresnní přímky v obou modelech a zjistěte, zdali se protínají? Co z toho můžeme vyvozovat? Na základě těchto modelů zjistěte o kolik procent bude mít automobil značky Chrysler a hmotnosti 1,6 tuny vyšší očekávanou spotřebu než průměrný automobil o stejné hmotnosti.

Otázka 08

Vykreslete scatterplot hmotností automobilů a jejich spotřeby. Do tohoto grafu vykreslete regresní přímku modelu s interceptem i bez. Sestrojte navíc lineární model, kde budete uvažovat, že spotřeba závisí na kvadrátu hmotnosti. Příslušnou křivku popisující odhady středních hodnot z tohoto modelu přidejte do obrázku k oboum předchozím modelům. Pro účely predikce spotřeby automobilů, na základě jakých statistik byste mezi těmito modely vybíraly, nebo byste se rozhodovali na základě něčeho jiného a proč?

Otázka 09

Pro vámi vybraný finální lineární model popisující vztah mezi hmotností a spotřebou automobilu ověřte předpoklady pro použití metody nejmenších čtverců. Každý předpoklad zmiňte a uveďte jak byste ho validovali pomocí reziduí.

Otázka 10

Přidejte k vysvětlující proměné `hmotnost`, i proměnnou `puvod`. Navrhněte aditivní lineární model (případně 3 modely pro každý region zvlášť), ve scatterplotu vykreslete 3 skupiny různými barvami a data proložte třemi odpovídajícími regresními přímkami. Uvažujeme 3 auta o hmotnosti 2 tuny zastupující jednotlivé regiony původu. Sestrojte 90% konfidenční intervaly okolo očekávaných spotřeb a na jejich základě rozhodněte, zdali a jak se očekávané spotřeby budou lišit. Je to porovnávání správné? Zdůvoněte.