Úkol A – Explorace dat a lineární regrese – bike sharing

ZS 2018

Cíl

Cílem prvního úkolu je seznámit se s daty vytěženými z bike sharingové služby a pomocí lineární regrese zjistit, jaký je vztah mezi parametry prostředí a počtem půjčených jízdních kol.

Data

Využívání služby půjčování jízdních kol má vysokou spojitost se sezónními podmínkami a parametry prostředí, např. typem počasí, srážkami, dnem v týdnu, ročním obdobím atd. Původní dataset byl pořízen během let 2011 a 2012 bikesharingovým centrem ve Washingtonu D.C., USA. Data byla agregována do hodinového měřítka a následně byly přidány údaje o počasí. *Předložený dataset obsahuje pro zjednodušení data agregovaná do celých dnů.*

Explorace dat

Data ze studie o využití této služby obsahují 16 příznaků, jejich stručný popis je uveden v *Tabulce 1* níže. V rámci explorace dat prozkoumejte příznaky: **měsíc, pracovní den, typ počasí, teplota, pocitová teplota, rychlost větru, a celkový počet vypůjčených kol**. Seznamte s typem dat v jednotlivých příznacích, rozsahem jejich hodnot, jejich rozdělením, případně počtem chybějících a odlehlých hodnot. Zvažte, jestli některé z příznaků není potřeba transformovat pro potřeby dalších analýz. Prozkoumejte závislosti mezi jednotlivými příznaky a zajímavé závislosti vizualizujte.

Modelování pomocí lineární regrese

Využití bike sharingu je velmi odvislé od počasí, dne v týdnu a dalších faktorů. Pokuste na základě dat zjistit, jaké proměnné mají na něj skutečný vliv a modelujte vztah mezi vybranými parametry a celkovým počtem půjčených kol pomocí lineárního modelu. Výsledný model zhodnoťte z hlediska přesnosti modelu a jeho statistické významnosti.

Požadované kroky analýzy

Průzkumová analýza. [4 body]

- Dimenze dat (počet příznaků, počet instancí)
- Chybějící a odlehlé hodnoty, transformace dat
- Vizualizace

Modelování vztahu mezi celkovým počtem půjčených kol a parametry popisující prostředí a čas [6 bodů]

- Formální zápis modelu
- Interpretace koeficientů modelu
- Přesnost modelu
- Statistická významnost

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* a stručně okomentovaný, přehledný zdrojový *R skript* odevzdejte pomocí *UploadSystemu*.

Název parametru	Popis parametru
dteday	Datum dne měření (yyyy-mm-dd)
Season	Roční období (1:jaro, 2:léto, 3:podzim, 4:zima)
Yr	Rok (0: 2011, 1: 2012)
Mnth	Kalendářní měsíc
Holiday	Prázdniny
Weekday	Den v týdnu (0: neděle, 1: pondělí, …, 5: pátek, 6: sobota)
Workingday	Den není ani víkend ani prázdniny
Weathersit	Typ počasí (1: jasno až polojasno, 2: mlha, zataženo, 3: lehké sněžení, přeháňky, bouřka, 4: Silný déšť, kroupy, hustá mlha, bouřky)
Temp	Normalizovaná teplota v ^{o}C. Normalizace pomocí vztahu (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39
Atemp	Normalizovaná pocitová teplota v ^{o}C. Normalizace pomocí vztahu (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50
Humidity	Normalizovaná vlhkost
Windspeed	Normalizovaná rychlost větru
Casual	Počet nahodilých uživatelů
Registered	Počet registrovaných uživatelů
Cnt	Celkový počet zapůjčených kol

Tabulka 1: Výčet parametrů

Bonusová úloha

Lineární regrese, nebo obecněji lineární model, je možné k testování řady poměrně složitých hypotéz. Lineární model může sloužit k porovnání závislostí na základě kategorických proměnných. Zajímavým příkladem v tomto může být vztah mezi dnem v týdnu a počtem vypůjčených kol. Lineární model lze použít k takovémuto porovnání.
Lze na základě dat z této studie rozhodnout, zda existuje souvislost mezi počtem půjčených kol a tím, jestli je pracovní den (nebo víkend či prázdniny), resp. zda existuje nerovnost mezi pracovním dnem a víkendem nebo prázdninami?
Bonusová úloha umožňuje získání až 4 dalších bodů, ale vyžaduje samostudium teorie a praktického použití lineárních modelů nad rámec cvičení DVZ.