Úkol A – Explorace dat a lineární regrese – kvalita bílého vína

ZS 2018

Cíl

Cílem prvního úkolu je seznámit se s daty ze studie kvality bílého vína a pomocí lineární regrese zjistit, jaký je vztah mezi fyzikálně-chemickými parametry vína a jeho subjektivním hodnocením.

Data

Tato data pocházejí ze severozápadní oblasti Portugalska zvaného Minho. Každé z vín bylo slepě hodnoceno nejméně 3 nezávislými degustátory podle stupnice 1 (velmi špatné) až 10 (excelentní). Z podstaty věci jsou však taková ohodnocení velmi subjektivní. Z toho důvodu byl každý vzorek také podroben fyzikálně-chemické analýze a zjištěny jeho základní parametry. Je tak možné zkoumat jednak vztahy mezi jednotlivými parametry, tak i mezi nimi a subjektivním hodnocením kvality.

Explorace dat

Data ze studie o kvalitě bílého vína obsahují 12 příznaků, jejich stručný popis je uveden v *Tabulce 1* níže. 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

Senzorické hodnocení kvality vína je velmi subjektivní. Pokuste na základě dat zjistit, jaké proměnné mají skutečný vliv na jeho hodnocení a modelujte vztah mezi vybranými parametry a subjektivním hodnocením kvality vína 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 subjektivním hodnocením kvality vína a fyzikálně-chemickými parametry [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
Fixed acid (vázaná kyselost)	Obsah kyseliny vinné (g/l)
Volatile acid (těkavá kyselost)	Obsah kyseliny octové (g/l)
Citric acid (kyselina citronová)	Obsah kyseliny citronové (g/l)
Chlorides (chloridy)	Obsah chloridu sodného (g/l)
Free sulfur dioxide (volný SO_2)	Obsah volného oxidu siřičitého (mg/l)
Total sulfur dioxide (celkem SO_2)	Celkový obsah oxidu siřičitého (mg/l)
Density (hustota)	Hustota (g/l)
pH	Kyselost – pH
Sulphates (sulfáty)	Obsah síranu draselného (g/l)
Alcohol (alkohol)	Obsah alkoholu (% vol.)
Sweet	Víno sladké? 1 – sladké, 0 - suché
Quality (kvalita)	Výstupní veličina – subjektivní kvalita vína, stupnice 0: velmi špatné víno – 10: excelentní víno

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 sladkostí vína a jeho hodnocením. Lineární model lze použít k takovémuto porovnání.
Lze na základě dat ze studie kvality vín rozhodnout, zda existuje spojitost mezi sladkostí vína (sladké/suché víno) a jeho subjektivním hodnocením, resp. poukazují data na to, že existuje kvalitativní rozdíl mezi sladkými a suchými víny?
Bonusová úloha umožňuje získání až 4 dalších bodů, ale vyžaduje samostudium teorie a praktického použití lineárních modelů nad rámec cvičení DVZ.