Úkol A – Explorace dat a lineární regrese – Galtonova studie

ZS 2016

Cíl

Cílem prvního úkolu je seznámit se s daty z Galtonovy studie a pomocí lineární regrese zjistit, jaký je vztah mezi výškou rodičů a jejich potomků.

Data

Galtonova data pochází ze studie provedeného v roce 1885 Francisem Galtonem a následně shrnuté v článku „*Regression Towards Mediocrity in Hereditary Stature*“. Data Francisi Galtonovi posloužila ke studiu vztahu mezi výškou rodičů a jejich potomků. Kromě praktických závěrů ohledně zmiňovaného vztahu výšky rodičů a jejich potomků, se Galtonova studie a data z ní pocházející stala jedním z datasetů používaných pro výuku lineární regrese a demonstraci jejích vlastností, jako je např. regrese k průměru.

Explorace dat

Data z Galtonovy studie obsahují 6 příznaků, jejich stručný popis je uveden v tabulce 1 níže. V rámci explorace dat prozkoumejte všechny příznaky. 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

Ve společnosti panuje (pravděpodobně) názor, že děti vysokých rodičů budou také vysoké. Pokuste se na základě Galtonových dat zhodnotit, jakým způsobem je výška potomka ovlivněna výškou jeho matky, otce a počtem jeho sourozenců. Na základě exploračních vizualizací a modelů, rozhodněte které parametry jsou pro výšku potomků vhodné a modelujte vztah mezi vybranými parametry a výškou potomku 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 výškou rodičů a jejich potomků [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 odevzdejte pomocí UploadSystemu.

Family	Identifikátor rodiny
Father	Výška otce v palcích
Mother	Výška matky v palcích
Gender	Pohlaví potomka
Height	Výška potomka v palcích
Kids	Počet sourozenců

Tabulka 1: Popis proměnných z Galtonovy studie

Bonusová úloha – rozdíly v dědičnosti výšky na základě pohlaví potomků

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 pohlavím potomka a vlivu jeho rodičů na jeho vzrůst. Lineární model lze použít k takovémuto porovnání. Lze na základě dat z Galtonovy studie rozhodnout, zda existují vztahy ve smyslu ‚výška syna je významně závislá na výšce otce', ‚výška syna je významně závislá na výšce matky', ‚výška dcery je významně závislá na výšce otce' a nebo ‚výška dcery je významně závislá na výšce matky'?

Bonusová úloha umožňuje získání až 4 dalších bodů, ale vyžaduje samostudium teorie a praktického použití lineárních modelů nad rámec cvičení DVZ.