Ukládání a příprava dat

2022/2023

Příprava dat a jejich popisná charakteristika

Kulíšek Vojtěch(xkulis03)
Plevač Lukáš(xpleva07)
Šesták Pavel(xsesta07)

Brno, 18. listopadu 2022

Obsah

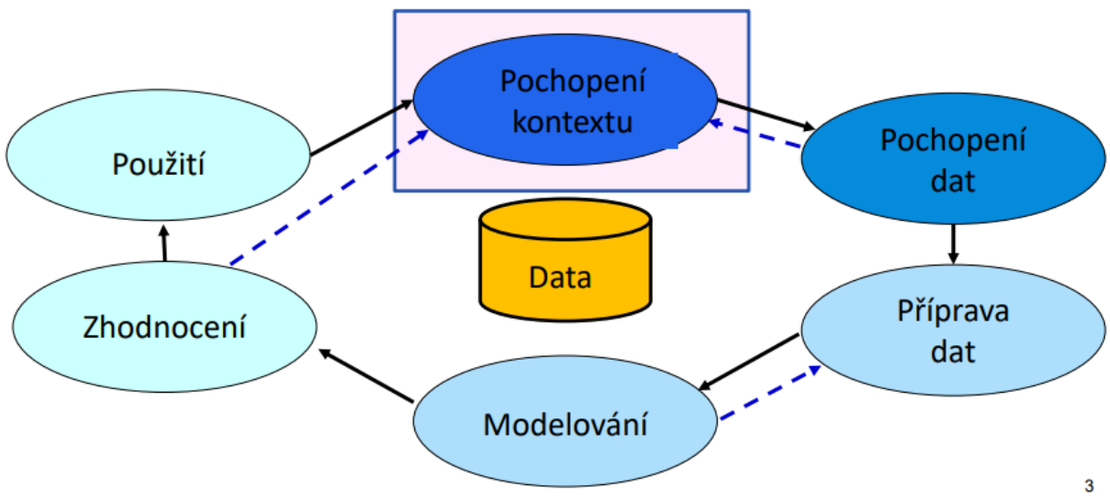
Seznam obrázků

1 Zadání

Z dostupných datových sad si zvolte jednu datovou sadu, kterou se budete dále zabývat. Stáhněte si zvolenou datovou sadu z uvedeného zdroje a prostudujte si dostupné informace k této datové sadě. Proveďte explorativní analýzu zvolené datové sady. Pro každý následující bod implementujte odpovídající sekci ve zdrojovém kódu a zjištěné výsledky popište v dokumentaci: prozkoumejte jednotlivé atributy datové sady, jejich typ a hodnoty, kterých nabývají (počet hodnot, nejčastější hodnoty, rozsah hodnot atd.) prozkoumejte rozložení hodnot jednotlivých atributů pomocí vhodných grafů, zaměřte se i na to, jak hodnota jednoho či dvou atributů ovlivní rozložení hodnot jiného atributu. Do dokumentace vložte alespoň 5 různých grafů, zobrazujících zjištěná rozložení hodnot. Použijte různé typy grafů (např. bodový graf, histogram, krabicový nebo houslový graf, graf složený z více podgrafů apod.). zjistěte, zda zvolená datová sada obsahuje nějaké odlehlé hodnoty. proveďte podrobnou analýzu chybějící hodnot (celkový počet chybějících hodnot, počet objektů s více chybějícími hodnotami atd.). proveďte korelační analýzu numerických atributů (k analýze využijte i grafy a korelační koeficienty). Připravte 2 varianty datové sady vhodné pro dolovací algoritmy. Můžete uvažovat dolovací úlohu uvedenou u datové sady nebo navrhnout vlastní dolovací úlohy. V případě vlastní dolovací úlohy ji specifikujte v dokumentaci. V rámci přípravy datové sady proveďte následující kroky: Odstraňte z datové sady atributy, které jsou pro danou dolovací úlohu irelevantní. Vypořádejte se s chybějícími hodnotami. Pro odstranění těchto hodnot využijte alespoň dvě různé metody pro odstranění chybějících hodnot. Vypořádejte se s odlehlými hodnotami, jsou-li v datové sadě přítomny. Pro jednu variantu datové sady proveďte diskretizaci numerických atributů tak, aby výsledná datová sada byla vhodná pro algoritmy, které vyžadují na vstupu kategorické atributy. Pro druhou variantu datové sady proveďte vhodnou transformaci kategorických atributů na numerické atributy. Dále pak proveďte normalizaci numerických atributů, které má smysl normalizovat. Výsledná datová sada by měla být vhodná pro metody vyžadující numerické vstupy.

2 Úvod

Cílem tohoto projektu je příprava datové sady k dolování znalostí z dat. Součástí této transformace je pochopení dat, detekce anomálii, doplnění chybějících hodnot a další. Projekt je implementován v Python Jupyter notebooku, kde samotné výpočty jsou proloženy popisným textem a obrázky se přímo vkládají do kontextu dané úlohy což umožňuje rychlejší orientaci v datech a jejich snazší porozumění.



Obrázek 1: Model životního cyklu analytického projektu CRISP [1]

3 Explorační analýza

V této sekci se snažíme blíže seznámit s daty. V této části je nutné si zobrazit a pochopit hodnoty jednotlivých atributů. Z grafů detekujeme například odlehlé hodnoty, zastoupení nevyplněných hodnot pro jednotlivé atributy. Dále zde můžeme detekovat například kvantitativní atributy, které z důvodu pár textových hodnot jsou chybně klasifikovány jako kvalitativní. Snažíme se odhalit korelace mezi atributy, a to ze dvou důvodů. Velmi korelované atributy nám v dolovací úloze

nepřinesou informaci navíc a pouze budou výpočty trvat déle. Jako další možnost pro využití korelovaných atributů je dopočítání chybějících hodnot, přičemž tato hodnota bude přesnější než použití například mediánu ze souboru dat.

4 Příprava dat

V této sekci se pokusíme popsat metody použité pro očištění a přípravu dat pro dolovací úlohy. Napřed odstraníme atributy, které nám zjevně nepomohou v dalším rozhodování a záznamy, které obsahují příliš málo vyplněných atributů, takový záznam by nám novou informaci do modelu nepřinesl, jelikož má většinu atributů dopočítaných. V rámci explorační analýzy jsme identifikovali korelace mezi atributy, proto nyní vytvoříme trénovací datovou sadu, která bude obsahovat všechny záznamy bez chybějících hodnot. Na této trénovací sadě následně natrénujeme regresní modely pro odhad korelovaných atributů. Zbylé nekorelované atributy doplníme podle povahy atributu očekávanou hodnotou, kterou uživatel nemusel vyplnit, pře se domníval, že se jedná o výchozí hodnotu například. U textových kvalitativních atributů je nutné se pokusit textové hodnoty agregovat pomocí různých textových transformací jako je například převod na malé písmena. Po doplnění chybějících hodnot je nutné odstranit anomálie, aby grafy byly čitelné a nedošlo k zbytečné změně měřítka os kvůli jedné odlehlé hodnotě. Pro detekci anomálií použijeme z skóre, jelikož předpokládáme, že se jedná o data pocházející z normálního rozložení.

Data převedeme do dvou datových sad, jedna bude obsahovat veškeré atributy převedené do kvalitativní formy a druhá do kvantitativní formy. Pro převod na kvantitativní atributy použijeme kódování 1 z n. Pro převod na kvalitativní atribut použijeme funkci qsub, která diskretizuje atribut do košů se stejným počtem prvků pomocí kvantilů.

5 Závěr

Po provedení prvních dvou částí z životního modelu CRISP jsme si chtěli ověřit, jak jsou naše výsledky použitelné. Pomocí lineární regrese jsme naučili model na naše očištěná data a přesnost predikce se pohybovala nad 90%.

Reference

[1] Ing. Ivana Burgetová, P.: Porozumění datům. 2022.