UNIVERZITA J. E. PURKYNĚ V ÚSTÍ NAD LABEM
Přírodovědecká fakulta

Python and R for Data Science

RNDr. Jiří Škvára, Ph.D.

Ústí nad Labem 2021

Předmět:	Python and R for Data Science
Studijní program:	Aplikovaná informatika
Klíčová slova:	python, R, data science
Anotace:	V předmětu studenti prakticky rozvinou základní dovednosti v programovacích jazycích Python a R v oblastech stěžejních pro datové inženýry a vědce. Různé metody a techniky zpracování, analýzy a vizualizace dat si studenti osvojí ryze prakticky na vzorových řešeních, tj. na aplikační a interpretační úrovni, bez nutnosti hlubších znalostí principů těchto metod a technik, které by měli nabýt v předchozím, případně dalším studiu. Výraznou součástí výuky je práce studentů ve skupinách na řešení případových studií („inspirovaných daty“) menšího rozsahu, jejich prezentace a vzájemné kritické zhodnocení. Zdrojem dat a inspirací jsou platformy typu kaggle.com. Ve výuce se uplatní materiály výukových platforem jako je datacamp.com, které jsou jinak doporučeny zejména k samostudiu a získání certifikátů.

Jazyková korektura nebyla provedena, za jazykovou stránku odpovídá autor.

Autor: RNDr. Jiří Škvára, Ph.D.

Obsah

Úvodní slovo

Tento kurz je koncipován jako přehled základních metod používaných pro datové vědy, přičemž důraz je kladen na praktické využití těchto metod v programovacích jazycích Python a R. Pro oba tyto programovací jazyky existuje řada knihoven obsahujících snadno použitelné nástroje jak pro statistickou analýzu a vizualizaci dat, tak i pro strojové učení a jeho specifické aplikace.

První dvě kapitoly tohoto kurzu jsou zaměřeny na samotné jazyky Python a R. Čtenář je zde seznámen nejen se základy těchto jazyků, ale i pokročilými datovými strukturami pro práci s rozsáhlými daty, možnostmi načítání dat ze souborů z prostých textových souborů a základy vizualizace. Druhá část textu, kapitoly tři až pět, je pak zaměřena na pokročilejší práci s daty jako například načítání dat ze specializovaných formátů, stahování dat z webu a čištění chybných zdrojových dat. Dále je zde důraz kladen na pokročilé vizualizační metody a statistické metody využitelné pro odhalovaní vztahů mezi vstupními daty. Třetí část textu, kapitoly šest až osm, je věnována strojovému učení a jeho praktickým aplikacím. V této části se seznámíme s metodami strojového učení s učitelem a bez učitele a bude vysvětleno, pro jaké aplikace jsou tyto metody vhodné. Jejich konkrétní aplikaci si ukážeme na případech kategorizace textu a analýzy sítí. V poslední části se zaměříme na nástroje pro tvorbu reportů nutných pro zaručení reprodukovatelnosti výsledků, a tvorbou interaktivních aplikací, dashboardů, pro přehlednou prezentaci výsledných dat.

Zápočet je možné získat prezentací skupinového projektu v rámci seminářů a samostatným projektem, jehož téma musí nejprve schválit zkoušející tohoto kurzu. Projekt by měl být na téma strojového učení s využitím dat dostupných na serveru kaggle.com. Spolu s kódem by měl být dodán report a vhodně zpracovaný dashboard prezentující výstupní data. V případě studentů kombinované formy studia odpadá nutnost prezentace skupinového projektu v rámci seminářů.

Zkouška spočívá v diskusi nad dodaným samostatným projektem, přičemž hodnocena je struktura kódu a reportu, znalost použitých metod a schopnost kód upravit či rozšířit o dodatečné funkce.

Splnění kurzu

V této sekci jsou uvedeny podmínky pro splnění kurzu.

Skupinové projekty

1. vybrat data ze serveru kaggle + motivační prezentace týmu
2. prezentace výsledků

Samostatný projekt

- Data ze serveru kaggle schválená po konzultaci s vyučujícím
- Odevzdání protokolu

Ukázka zadání zápočtového programu

Na základě dat dotupných v rámci Kaggle výzvy "Titanic - Machine Learning from Disaster"dostupných na adrese https://www.kaggle.com/c/titanic/data vytvořte model predikující úmrtí/přežití pasažérů nepotopitelné lodi Titanic. Použitý model validujte na základě trénovacích dat a zhodnoťte kvalitu použitého modelu. Vytvořte dashboard, který bude obsahovat grafy ukazující vztahy mezi vstupními proměnnými se zobrazenou hodnotou korelačního koeficientu. Dashboard by měl také ukazovat překryv předpovědí vůči reálným datům na trénovací množině dat.

1 Programovací Jazyky Python a R

Cílem této sekce je připomenutí základů programování v jazycích Python a R (syntax, práce s poli a seznamy, základní funkce a metody objektů) a prohloubení již nabytých znalostí pro účely datové analýzy. Speciální důraz je kladen na knihovny pandas a tidyverse a jejich použití pro rychlou a efektivní práci s daty. Tyto knihovny umožňují s daty pracovat jako s tabulkami podobně jako v tabulkových procesorech. Tyto tabulky jsou na úrovni kódu reprezentovány objektem dataFrame.

CÍLE KAPITOLY

- Základy programovacích jazyků Python a R.
- Pokročilé metody jazyků Python a R pro datové vědy
- Práce se základními knihovnami numpy a scipy.
- Práce se slovníky a knihovnami pandas/tidyverse.

KLÍČOVÁ SLOVA

Python a R, pandas, tidyverse, numpy

ÚKOLY

1. Vytvořte slovník "detectives"popisující alespoň čtyři literární/filmové detektivy. Každý klíč ve slovníku (jméno detektiva) bude mít jako hodnotu další slovník "author", "first book", "number of books". Výsledný slovník převeďte na pandas/tidyverse dataframe.
2. Vytvořte jednoduchou simulaci výstupu na Mount Everest (8 849 m). V této simulaci udělejte 1000 opakování tohoto algoritmu:
   (a) je-li možné stoupat, výstup o 0-100 metrů s pravděpodobností 99%.
   (b) v opačném případě, je-li možné klesat, sestup o metr
      - v případě sestupu o metr šance na úmrtí $P = \frac{\text{aktuální výška}}{\text{výška hory}}$
   (c) pokud horolezec tento krok přežil, pokračuj krokem (a), jinak vyšli dalšího horolezce.
3. Vytvořte pandas/tidyverse dataFrame obsahující informace o jednotlivých horolezcích z předchozího úkolu.

OTÁZKY

1. Popište rozdíl mezi seznamy a poli.
2. Jak byste provedli základní statistické operace nad poli?
3. Jak lze provést základní průzkum hodnot v dataFrame objektu?

OTÁZKY K ZAMYŠLENÍ

1. Jaký je rozdíl mezi vestavěnými slovníky a pandas/tidyverse dataFrame objekty. V jakých situacích je vhodnější použít slovníky?

SHRNUTÍ

Po prostudování byste měli být schopni:

- Využít jazyky R a python pro a jejich vestavěné funkce pro jednoduchou práci s reálnými či náhodnými daty.
- Provést základní statistickou analýzu dat.

ODKAZY NA LITERATURU

- Vše potřebné pro využití programovacího jazyka Python spolu s knihovnou pandas pro datové vědy nalezente zde.
- Materiály popisující využití jazyka R a knihovny tidyverse pro datové vědy naleznete zde.
- Dokumentace knihovny numpy naleznete zde.
- Interaktivní materiály k procvičení vybraných témat naleznete zde

MÍSTO PRO VAŠE POZNÁMKY

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

2 Práce s daty a jejich vizualizace

V této kapitole se zaměříme na prohloubení znalostí v oblasti práce se samotnými daty a jejich vizualizace. Základním krokem pro práci s daty je jejich načtení ze souboru. Nejčastěji se setkáváme s prostými datovými soubory, tedy soubory obsahující data v textové formě bez explicitně specifikovaných vzájemných relací. V tomto případě je nutné pouze ošetřit správné rozlišení oddělovačů jednotlivých hodnot stejně jako hlaviček souborů, které mohou být odlišné pro různé datové formáty. Jakmile jsou data načtena, je na nás, abychom je transformovali do požadované podoby, tedy přeškálovali hodnoty, převedly vše na správné jednotky, vytvořili nové sloupce či naopak sloučili několik existujících sloupců načtené tabulky.

Pro vizualizaci dat je pak důraz kladen na knihovny matplotlib a ggplot2, které jsou základním vizualizačním prvkem v jazycích Python a R.

CÍLE KAPITOLY

V této kapitole se dozvíte:

- jak načítat prosté datové soubory pomocí v jazycích Python a R.
- jak pracovat manipulovat s daty načtenými jako dataFrame objekt v pandas a tidyverse
- jak vizualizovat data pomocí matplotlib a ggplot2
- jaká grafická reprezentace dat je vhodná pro danou aplikaci

KLÍČOVÁ SLOVA

pandas, tidyverse, matplotlib, ggplot2

ÚKOLY

1. prostudujte si interaktivní materiály k tématu "základy vizualizace dat"na serveru datacamp zde
2. načtěte soubor train.csv (dostupný zde).
   - přidejte nový sloupec "Titul"obsahující titul pasažéra, který lze vyextrahovat ze sloupce "Name".
   - Vytvořte bodový graf, kde na jedné ose je cena vstuenky, na druhé věk pasažéra. Barevně oddělte pasažéry, kteří potopení lodi nepřežili.
3. využijte program s horolezci z předchozí sekce a pomocí matplotlib/ggplot2 vykreslete:
   - počet úspěšných horolezců v závislosti na počtu iterací

- počet zesnulých horolezců v závislosti na počtu iterací
- histogram nadmořských výšek, do kterých se jednotliví horolezci dostali.

OTÁZKY

1. Kdy je vhodné načítat data jako pole hodnot a kdy je lepší použít dataFrame?
2. Kdy je vhodné pro vizualizaci dat použít spojnicový graf, bodový graf a kdy histogram?
3. Jak byste zvolili vhodný počet dílků pro histogram?

SHRNUTÍ

Po prostudování byste měli být schopni:

- načíst libovolný prostý datový soubor
- transformovat a agregovat jednotlivé záznamy
- vhodným způsobem data vizualizovat.

ODKAZY NA LITERATURU

- Dokumentace knihovny matplotlib je dostupná zde.
- Úvod do práce s knihovnou ggplot2 naleznete zde.

MÍSTO PRO VAŠE POZNÁMKY

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

3 Pokročilé techniky práce s daty

V předchozích kapitolách jsme se již naučili pracovat s prostými datovými soubory načítanými z lokálního datového uložiště. V praxi se ovšem velmi často setkáváme s daty uloženými ve specializovaném formátu či relačními databázemi. Obvykle jsou také zdrojová data uložena na webu, kde sice můžeme ručně stáhnout jeden soubor po druhém, ale pro větší efektivitu je lepší tato data nějak přímo importovat v kódu samotném. Pro tento účel jsou k dispozici balíčky urllib (pro jazyk Python) a httr (pro jazyk R), které umožňují zadat požadavek pro stažení dat ze zadané webové adresy. Velmi populární je také balíček request, který celý proces zjednodušuje.

Zdrojová data velmi často obsahují různé chyby jako jsou překlepy v hodnotách, duplicitní záznamy či záznamy s prázdnými poli, nevhodný formát textu a podobně. Tyto chyby je potřeba identifikovat a odstranit ještě před tím, než na základě analýzy dat začneme formulovat závěry. Pro některé typy chyb, jako jsou například chybějící záznamy, jsou již připraveny metody pro jejich nahrazení případně odstranění. V jiných případech si však musíme sami rozhodnout, jaká data jsou chybná, a to na základě samotné povahy dat. Příkladem mohou být záznamy průměrných teplot ve vašem rodném městě za posledních dvacet let. Objeví-li se v datech chyba typu špatně zadaná hodnota nebo špatné teplotní jednotky, měli byste být schopni takovýto chybný záznam odhalit na základě histogramu.

V této kapitole si ukážeme pokročilé metody importu dat a metody jejich ošetření pro efektivní datovou analýzu.

CÍLE KAPITOLY

V této kapitole se seznámíte s:

- import lokálních dat specializovaných formátů v jazycích Python a R
- práce s relačními databázemi SQL v jazycích Python a R
- knihovny urllib /httr a requests
- objekt typu JSON v jazyce Python a R.
- čištění chybných zdrojových dat v dataFrame objektu v pandas a tidyverse

KLÍČOVÁ SLOVA

import dat, sql, JSON, čistění dat

ÚKOLY

1. Stáhněte JSON soubor obsahující hodnocení filmů podle různých kategorií.
    - vypořádejte se s chybějícími záznamy a odstraňte záznamy obsahující hodnotu "Unknown"
    - zkontrolujte, zda soubor neobsahuje duplicitní záznamy.
2. Ze serveru kaggle stáhněte SQL databázi hráčů evropské fotbalové ligy
    - v databázi odstraňte chybějící záznamy
    - porovnejte střední hodnoty a histogramy výšek a vah hráčů z různých zemí.

OTÁZKY

1. Popište co je API. Jak nám umožňuje získávat data z webových serverů?
2. Popište vlastnosti a strukturu objektu JSON.
3. JSON není standardní objekt jazyků Python a R. Jakým datovým typem je tedy v těchto jazycích reprezentován?

SHRNUTÍ

Po prostudování byste měli být schopni:

- načíst libovolný datový soubor
- načíst data z relační databáze, použít příkazy SQL pro výběr vhodných záznamů
- stáhnout data z webu pomocí balíku requests ve formátu JSON
- vyčistit zdrojová data, odstranění redundancí a záznamů s chybějícími poli
- odstranit chybné záznamy, špatně zadané jednotky, neplatné znaky a překlepy.

ODKAZY NA LITERATURU

- Obecný popis objektů JSON naleznete zde.
- Základy práce s relačními databázemi SQL jsou k dispozici zde.

4 Pokročilé techniky vizualizace dat

V předchozích kapitolách jsme se již seznámili s možnostmi vizualizace dat pomocí knihoven matplotlib a ggplot2. Zatímco ggplot2 nabízí poměrně jednoduché rozhraní pro tvorbu pokročilých grafů, matplotlib obvykle vyžaduje mnoho řádku kódu pro správné nastavení grafu. V této kapitole se seznámíme s nástrojem Seaborn, který je dostupný pouze pro jazyk Python a funguje jako nadstavba nad knihovnou matplotlib. Tato knihovna umožňuje snadno vytvářet vizuálně atraktivní grafy na základě přednastavených schémat. Seaborn je primárně určen pro statistickou vizualizaci a umožňuje tak například vykreslovat do grafů intervaly spolehlivosti bez nutnosti dalších výpočtů.

Dále se seznámíme s knihovou Plotly, která také poskytuje vysokoúrovňové rozhraní pro tvorbu grafů publikační kvality. Plotly, navíc oproti knihovnám Seaborn a ggplot2 poskytuje možnost vytvářet interaktivní grafy či jednoduše přidávat vysvětlivky do grafů.

CÍLE KAPITOLY

V této kapitole se seznámíte s

- s pokročilou vizualizací dat pomocí knihoven Seaborn a ggplot2
- vizualizací dat pomocí knihovny Plotly v jazycích Python a R
- relačními a kategorickými grafy
- možností přidávání vysvětlivek do grafů.

KLÍČOVÁ SLOVA

Seaborn, Plotly, ggplot2, interaktivní grafy

ÚKOLY

1. použijte data o pronájmu kol ve Washingtonu (dostupná zde) a užitím Seaborn knihovny vykreslete spojnicový relační graf celkového počtu výpůjček v závislosti na měsíci. Jako datové body pro vykreslení využijte průměrnou hodnotu pro každý měsíc v roce. Vizualizaci proveďte pomocí dvou podgrafů, kde jeden bude zobrazovat také interval spolehlivosti, zatímco druhý bude zaměřen na směrodatnou odchylku. Pokuste se vytvořit obdobnou vizualizaci v jazyce R pomocí ggplot2.
2. vytvořte regresní graf nad daty z minulého úkolu. Porovnejte grafy lineární a polynomiální regrese.
3. pomocí Plotly vytvořte bodový graf celkového počtu výpůjček kol v závislosti na teplotě. Do vizualizace přidejte posuvník který bude měnit výběr vykreslovaných hodnot podle daného měsíce v roce.

OTÁZKY

1. Popište vztah mezi knihovnami matplotlib a Seaborn.
2. Popište rozdíl mezi relačními a kategorickými grafy.
3. Popište rozdíly mezi grafy vytvořenými jako plotly.express, plotly.graph_objects a plotly.figure_factory.

SHRNUTÍ

Po prostudování byste měli být schopni:

- připravit grafy publikační kvality s nástroji Seaborn, ggplot2 a Plotly
- vytvořit grafy zahrnující směrodatné odchylky, regresi či interval spolehlivosti
- vložit do grafů vysvětlivky
- vytvořit interaktivní grafy s ovládacími prvky

ODKAZY NA LITERATURU

- Interaktivní dokumentaci plotly pro jazyk Python naleznete zde.

MÍSTO PRO VAŠE POZNÁMKY

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

5 Analýza dat

V předchozích sekcích jste se již naučili, jak načíst zdrojová data, ověřit že nevykazují chyby a případně tyto chyby opravit. Naučili jste se také, jak tato data prozkoumat pomocí základních statistických funkcí a histogramu. V této kapitole se zaměříme na hlubší analýzu distribucí dat. Mimo to se také zaměříme na odhalování vztahů mezi různými proměnnými využitím mimo jiné korelační a regresní analýzy. Dále se zaměříme na kategorizaci dat pomocí shlukové analýzy, která soběpodobné objekty spojuje do větších shluků.

CÍLE KAPITOLY

V této kapitole se seznámíte s pojmy:

- explorační analýza a její využití v jazycích Python a R
- korelační a regresní analýza
- shluková analýza v jazycích Python a R
- faktorová analýza v jazycích Python a R
- inferenčí statistika v jazycích Python a R

KLÍČOVÁ SLOVA

exploračí analýza, korelace a regrese, shluková analýza, ověřování hypotéz

ÚKOLY

1. použijte data obecné sociálního průzkumu provedeného v USA od roku 1972. Tato data obsahují záznam o roce, kdy byl proveden průzkum a informace o respondentovi (pohlaví, věk, rok narození, rasa, délka vzdělání, celkový rodinný příjem). Vykreslete histogram, pravděpodobnostní funkci PMF a kumulativní distribuční funkci pro celkový rodinný příjem a porovnejte podle rasy a pohlaví v různých letech průzkumu.
2. vyšetřete závislost mezi daty délkou studia a příjmem v různých letech. Vypočtěte hodnotu korelačního koeficientu a porovnejte ji s hodnotou směrnice regresní přímky. Liší se tyto koeficienty v závislosti na pohlaví a rase?
3. ověřte hypotézu, že rozdělení příjmů u mužů a žen je v roce 2016 identické.
4. načtěte libovolný bitmapový obrázek a s využitím shlukové analýzy identifikujte dvě dominantní barvy.

5. využijte data z průzkumu založeného na pětibodovém osobnostním testu (dostupné zde)
    - proveďte Bartlettův nebo Kaiserův-Meyerův-Olkinův (KMO) Test pro ověření "faktorizovatelnosti"dat
    - zvolte vhodný počet faktorů pro faktorovou analýzu (například pomocí Kaiserova kritéria)
    - určete, které faktory se projevují v daných odpovědích na dotazník.

OTÁZKY

1. Vysvětlete rozdíl mezi histogramem, pravděpodobnostní funkcí PMF a kumulativní distribuční funkcí.
2. Vysvětlete rozdíl mezi kategorickými a numerickými proměnnými.
3. Korelační koeficient zkoumaných data vykazuje hodnotu blízkou nule. Znamená to, že mezi daty není žádný vztah?
4. Jak popsat vztah mezi proměnnými, víme-li že není lineární?
5. Vysvětlete princip centroidní metody hierarchického shlukování.

SHRNUTÍ

Po prostudování byste měli být schopni:

- najít vztahy mezi zadanými daty,
- porovnat různé záznamy na základě formulace a ověření nulové hypotézy,
- identifikovat faktory vystupující v datech,
- identifikovat shluky data a určit jejich charakteristické vlastnosti.

..

MÍSTO PRO VAŠE POZNÁMKY

6 Základní aplikace metod strojového učení

Po prostudování předchozí kapitoly bychom měli být schopni analyzovat vstupní data a určit, jak na sobě jednotlivé vstupní proměnné závisí. Na základě těchto závislostí pak s využitím lidské intuice můžeme vytvářet predikce pro nová data stejného typu či v těchto datech vyhledávat vzory a anomálie. V případě, že vstupní data obsahují větší množství proměnných ovšem lidská intuice selhává a je nutné využít specializovaných algoritmů, které jsou schopné identifikovat a "naučit se" závislosti mezi obecně komplexními vstupními daty a provádět ono intuitivní vyhodnocení za nás. V této kapitole se seznámíme se základními tématy tohoto strojového učení, kterými jsou učení s učitelem a bez učitele. S metodami učení bez učitele jsme se již setkali u tématu shlukové analýzy. Toto učení slouží prozkoumání vstupní dat bez znalosti cílové proměnné. Vstupní data jsou pak spojována do shluků (někdy zvaných klastry), kterými lze množinu dat nahradit a tím i zjednodušit. Vstupní data jsou pak klasifikována na základě charakteristické vlastnosti klastru. Naproti tomu učení s učitelem předpokládá, že máme k dispozici cílovou proměnnou. Pro naučení našeho modelu pak musíme zadat takzvanou trénovací množinu parametrů, tedy kombinaci známých vstupů a výstupů. Na základě povahy dat pak nad novými vstupními daty provádíme buď jejich klasifikaci do dané skupiny, nebo se snažíme odhadnout novou výstupní hodnotu na základě regrese trénovacích dat. Alternativou k metodám strojového učení je využití vícevrstvých neuronových sítí (deep learning), které se na základě matematického modelu neuronu snaží napodobovat reálné učení biologických systémů.

CÍLE KAPITOLY

Seznámíme se s následujícími tématy:

- s obecnými metodami strojového učení (zde)
- programová realizace učení s učitelem
  - klasifikace dat v jazycích Python a R
  - regrese v jazycích Python a R
- Učení bez učitele v jazycích Python a R
- Vícevrstvé neuronové sítě (deep learning) pro Python a R

KLÍČOVÁ SLOVA

strojové učení, učení bez učitele, učení s učitelem, deep learning

ÚKOLY

1. Použije data popisující informace o zdravotním stavu pacientů (dostupná zde) a vytvořte klasifikační model, který každému z pacientů přiřadí vhodný lék na jeho obtíže. Využijte vhodnou metodu přístupu učení s učitelem a výsledky porovnejte s výstupem vícevrstvé neuronové sítě.
2. Využijte informace o klientech zdravotní pojišťovny (dostupná (zde) a vytvořte regresní model, který bude předpovídat náklady na zdravotní výdaje pojištěnce.
3. Pomocí strojového učení bez učitele vytvořte shlukujte státy podle klíčových parametrů určujících jejich rozvoj (data jsou dostupná zde). Porovnejte k-means a hierarchické klastrování a vyberte pět nejméně rozvinutých zemí, které zaslouží dotaci.

OTÁZKY

1. Pro jaké aplikace je vhodné použít učení bez učitele a kdy je naopak výhodnější použít učení s učitelem?
2. Nevhodné nastavení parametrů modelu pro strojové učení může vést k takzvanému přeučení (overfitting) nebo nedoučení (underfitting). Jak byste postupovali pro nalezení optimálních parametrů modelu?
3. Popište, jaké informace lze vyčíst z dendrogramu.
4. Jaké metody lze použít pro nalezení optimálního množství klastrů?
5. Jaký je rozdíl mezi hierarchickým klastrováním a k-means klastry?
6. Popište princip algoritmu backpropagation.

OTÁZKY K ZAMYŠLENÍ

1. Algoritmy učení s učitelem a neuronové sítě se obě používají na klasifikaci a predikci dat, přesto se ale velmi často označují za oddělené větve metod strojového učení. Jaké jsou zásadní rozdíly mezi aplikací učení s učitelem a neuronovými sítěmi?

SHRNUTÍ

Po prostudování byste měli být schopni:

- vytvořit model pro kategorizaci dat do předem specifikovaných diskrétních domén
- vytvářet prediktivní modely
- využít shlukové analýzy pro kategorizaci dat bez předem určené výstupní hodnoty
- využít vícevrstvé neuronové sítě pro klasifikaci a regrese hodnot.

ODKAZY NA LITERATURU

- Přehled metod strojového učení naleznete zde
- Dokumentace k projektu Keras k nahlédnutí zde.
- Popis API projektu TensorFlow naleznete zde

MÍSTO PRO VAŠE POZNÁMKY

7 Základy textové analýzy

Jednou z aplikací strojového učení, na kterou můžeme narazit v běžném každodenním životě, je zpracování textu. Toto zpracování spočívá v rozdělení vstupního textu na jednotlivé elementy, tokeny, a následném vyhodnocení relací mezi nimi. Typickým příkladem využití textové analýzy v praxi je doplňování slov při psaní zpráv, zvýrazňování chyb ve slovech, návrhy doporučených témat či skupin na sociálních sítích nebo spam filtry pro e-mailové schránky. V mnoha případech nás ovšem nezajímá pouze obsah zadaného textu, ale také styl jakým je text psaný. Tomuto zpracování se říká analýza sentimentu, která může být použita například pro analýzu hodnocení produktů nabízených na internetových obchodech. Analýza sentimentu může být také použita jako pomůcka při psaní e-mailů, kde může poskytovat informaci o tom, zda je námi psaná zpráva příliš formální nebo je naopak příliš emočně zabarvená. V této kapitole se seznámíme se základy textové analýzy a analýzy sentimentu.

CÍLE KAPITOLY

Seznamte se s následujícími tématy:

- Regulární výrazy a tokenizace v Python a R
- Identifikace prostých témat v textu a bag-of-words model v Python a R
- Rozpoznávání pojmenovaných entit v Python a R

KLÍČOVÁ SLOVA

analýza sentimentu, zpracování textu, bag-of-word model, vektory slov, n-gramy

ÚKOLY

- Popište a vysvětlete mode bag-of-words.
- Vytvořte fake news klasifikátor využitím Naive Bayes modeulu. Jako vstup využijte data dostupná zde.
- Proveďte analýzu sentimentu u příspěvků ze sociální sítě twitter (dostupné zde). Jednotlivé tweety jsou ohognoceny hodnotami 0 = negativní, 2 = neutrální a 4 = pozitivní.

SHRNUTÍ

Po prostudování byste měli být schopni:

- Efektivně prohledávat textové soubory pomocí regulárních výrazů.
- Vytvořit program, který dokáže kategorizovat text dle jeho obsahu.
- Efektivně nastavit model pro analýzu sentimentu u dodaného textu.

OTÁZKY

1. Vektory slov pro textovou analýzu mohou být pro každý text vytvořeny nově pomocí strojového učení a nebo načteny pro daný jazyk z dostupných knihoven. Jaké jsou výhody a nevýhody těchto přístupů. V jakých případech je vytvoření nových vektorů výhodnější.
2. Co je a k čemu je dobrý mrak slov (word cloud).
3. Co jsou to stopslova a proč je vhodné je nezačleňovat do analýzy textu.

OTÁZKY K ZAMYŠLENÍ

1. Pro analýzu sentimentu v mnoha případech nestačí využívat jednotlivé tokeny (unigramy), jelikož ztrácíme informaci o vztahu mezi slovy. Příliš mnoho tokenů (n-gramů) ale může vést k přeučení a chybným výsledkům. Jakým způsobem byste zvolili správný rozsah n-gramů?

MÍSTO PRO VAŠE POZNÁMKY

8 Analýza sítí

V této kapitole zúročíme znalosti z teorie grafů a zaměříme se na jejich využití pro identifikaci struktur v různých typech sítích. Těmito sítěmi můžeme v praxi rozumět například sociální sítě, kde uzly jsou tvořeny jednotlivými uživateli sítě, případně skupinami uživatelů či nabízené produkty a hrany sítě jsou pak vazby mezi nimi. Na základě analýzy sítě pak můžeme nalézt důležité uzly sítě, identifikovat komunity uzlů a vytvořit systém doporučující nová propojení. V této kapitole se seznámíte s důležitými pojmy a algoritmy nutnými pro analýzu sítí.

CÍLE KAPITOLY

Seznamte se s následujícím:

- knihovny NetworkX pro Python a igraph pro jazyk R
- grafické reprezentace sítí
- vyhledání důležitých uzlů v jazycích Python a R.
- analýza struktury sítě Python a R

KLÍČOVÁ SLOVA

teorie grafů, kliky grafu, sociální sítě, detekce komunit

ÚKOLY

1. Napište program, který načte síť uživatelů sociální sítě Youtube (dostupné zde).
2. Vytvořte maticový, obloukový a "circos" graf načtené sítě.
3. Vyhledejte v síti všechny maximální kliky.
4. Vytvořte doporučení jací uživatelé by se měli propojit na základě společných uzlů sítě.

SHRNUTÍ

Po prostudování byste měli být schopni:

- Vizualizovat studovanou síť pomocí vhodné grafické reprezentace.
- Vyhledat nejkratší cesty v síti.

- Analyzovat síť a ohodnotit jednotlivé uzly.
- Vyhledat v síti struktury se stejným či podobným zaměřením (komunity).
- Vytvořit doporučení pro zapojení nezávislých uzlů do komunity.

OTÁZKY

1. Popište algoritmus hledání nejkratší cesty v grafu.
2. Jaké grafické reprezentace jsou vhodné pro vizualizaci sítí?

MÍSTO PRO VAŠE POZNÁMKY

9 Reporty, dashboardy a interaktivní vizualizace dat

V předchozích kapitolách jsem se soustředili primárně na zpracování a vyhodnocení dat. Důležitou součástí praxe datového vědce je však nejen získat výsledné hodnoty, ale také je vhodně prezentovat a zajistit reprodukovatelnost dosažených výsledků. Vhodná prezentace dat nemusí nutně znamenat pouze vytvoření vizuálně atraktivních grafů. Důležitější roli může hrát interaktivní složka prezentace. V mnoha aplikacích je vhodné mít možnost přepínat mezi různými vstupními daty a modely použitými pro práci s daty, měnit typy grafů či rozsahy os a podobně. Pro zaručení reprodukovatelnosti získaných výsledků je pak důležité sepsat srozumitelný report obsahující nejen samotné výsledky a popis použitých metod, ale i úryvky důležitých částí vlastního kódu. V této kapitole se naučíte pracovat s nástroji pro vytváření efektivní reportů k sumarizaci analýz a diskusi výsledků stejně jako možnostmi tvorby interaktivních aplikací.

CÍLE KAPITOLY

Seznamte se s následujícím:

- tvorba reportů v R Markdown a jupyter-notebook
- tvorba dashboardů pomocí plotly Dash, shiny a jupyter-dashboard
- tvorba interaktivních aplikací anvil

KLÍČOVÁ SLOVA

R Markdonw, jupyter-notebook, Plotly, shiny, anvil

ÚKOLY

1. Podívejte se na tutorial pro nástroj Anvil pro tvorbu interaktivních aplikací zde. Zkuste si dle návodu vytvořit jednoduchou aplikaci.
2. Vypracujte report pro úlohu z kapitoly analýza sítí pomocí R Markdown nebo Jupyter-notebook. Report by měl obsahovat hlavičku s názvem projektu, jménem univerzity, jménem autora a aktuálním datem. Následovat by měla stručná anotace projektu. Vkládejte segmenty kódu, které poskytují nějaký výstup.

3. Připravte interaktivní vizualizaci libovolného z předchozích úkolů. Aplikace by měla obsahovat text popisující vybranou úlohu, tlačítko pro načtení vstupních dat, tlačítko pro provedení výpočtu, výsledná vizualizace výpočtu (graf či vhodně formátovaný textový výstup)

SHRNUTÍ

Po prostudování byste měli být schopni:

- vytvořit report vašich výsledků zaručující reprodukovatelnost analýzy
- vytvářet pokročilejší interaktivní vizualizace pomocí Plotly
- vytvářet interaktivní aplikace či dashboardy pro prezentaci dat pomocí nástrojů Anvil, Plotly Dash, nebo Jupyter-notebook

ODKAZY NA LITERATURU

- Průvodce Plotly Dash pro grafickém návrhář zde a pro Python API zde.

MÍSTO PRO VAŠE POZNÁMKY

Literatura

[1] VANDERPLAS, Jake. Python Data Science Handbook: Essential Tools for Working with Data. Dostupné z: https://jakevdp.github.io/PythonDataScienceHandbook/

[2] GROLEMUND, Garrett a Hadley WICKHAM. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Dostupné z: https://r4ds.had.co.nz/

[3] DataCamp: Learn Data Science Online. Dostupné z: https://www.datacamp.com/

[4] Kaggle: Your Machine Learning and Data Science Community. Dostupné z: https://www.kaggle.com/

[5] W3Schools. Dostupné z https://www.w3schools.com/default.asp

[6] EDUCBA: Become an Awesome Data Analyst! Dostupné z https://www.educba.com/data-science/

[7] Keras: Developer guides. Dostupné z https://keras.io/guides/

[8] TensorFlow: API Documentation. Dostupné z https://www.tensorflow.org/api_docs

[9] Anvil: Learning Centre. Dostupné z https://anvil.works/learn

[10] Jupyter Dashboards Layout Extension. Dostupné z https://jupyter-dashboards-layout.readthedocs.io/en/latest/index.html#

[11] Plotly: Dash User Guide. Dostupné z https://dash.plotly.com/