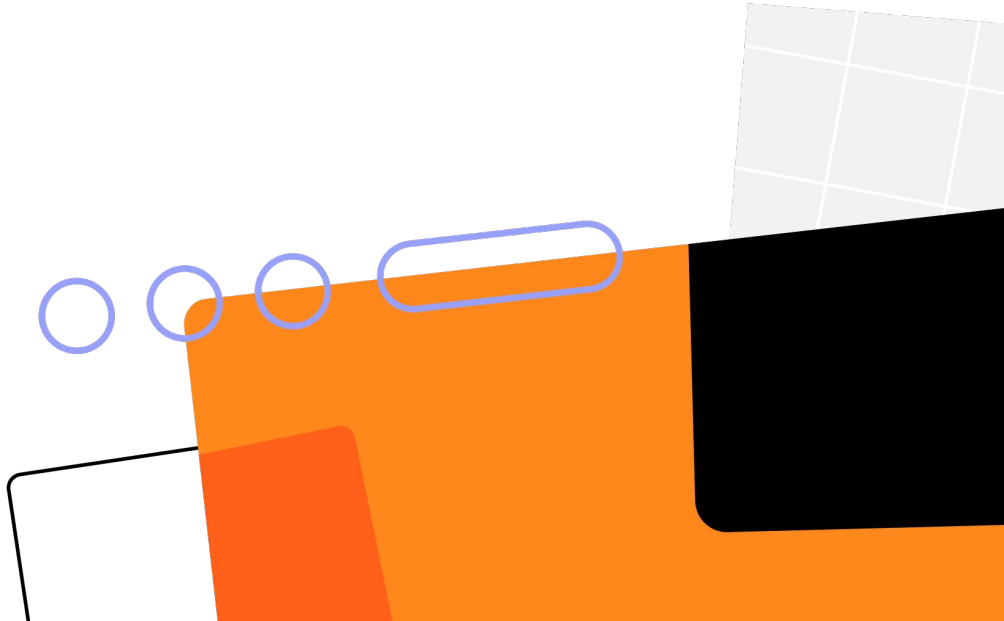
Cvičení:
Regrese (projekt)

Kurz Data Science



Kde jsme?

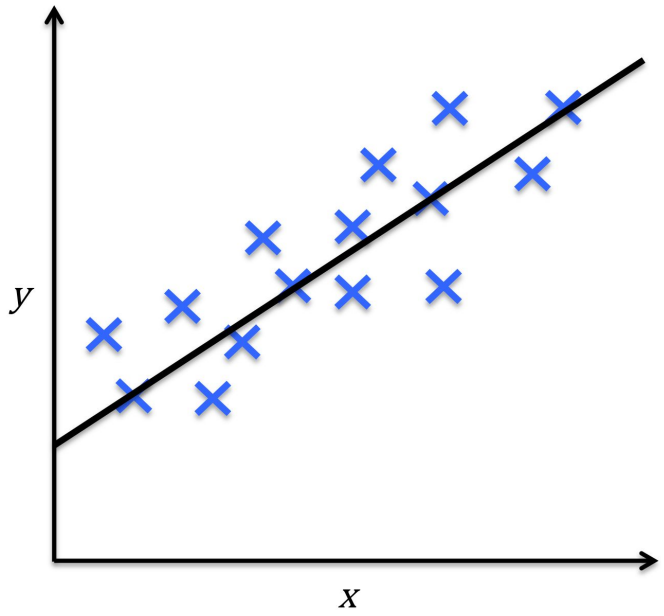
Úvod	Programování v Pythonu	Matematika & Datová analýza	**Klasické strojové učení**	Umělé neuronové sítě	Závěrečný projekt
1-3	4-6	7-10	11-17	18-21	22
★ Úvod do Data Science ★ Nástroj GIT (video) ★ Systémové nástroje, správa verzí GIT a projektový management	★ Základy Pythonu ★ Python pro mírně pokročilé ★ Python v Data Science	★ Datové pipeliny ★ Matematický základ pro Data Science ★ Zpracování dat & explorativní datová analýza ★ Vizualizace dat	★ Klasické strojové učení ★ Učení pod dohledem: Regrese ★ **Cvičení: Regrese (projekt)** ★ Učení pod dohledem: Klasifikace ★ Cvičení: Klasifikace (projekt) ★ Učení bez dohledu ★ Cvičení: Učení bez dohledu (projekt)	★ Úvod do umělých neuronových sítí ★ TensorFlow knihovna & Keras API ★ Trénování hlubokých neuronových sítí ★ Hluboké učení: Přizpůsobené trénování a aplikace	★ Závěrečný projekt

Cvičení:

Regrese (projekt)

Praktické projekty hrají klíčovou roli ve výuce ML, protože poskytují praktickou zkušenost a pomáhají upevnit vaše porozumění teoretickým konceptům s jejich aplikací v kontextu reálného světa.

- průzkum a zpracování dat
  - zpracování chybějících hodnot
  - detekce odlehlých hodnot
  - feature engineering
  - normalizace/standardizace dat
- výběr modelu
  - lineární regrese (jedna nebo více proměnných)
    - polynomiální rysy
    - vztahy proměnných
    - regularizace
  - SVR - support vector regressor
  - rozhodovací strom a souborné metody
- hodnocení modelu
  - výběr metrik (MSE, R^2 nebo RMSE)
  - rozdělení trénovací/validační/testovací sada
  - křížová validace a ladění modelu



Cvičení:

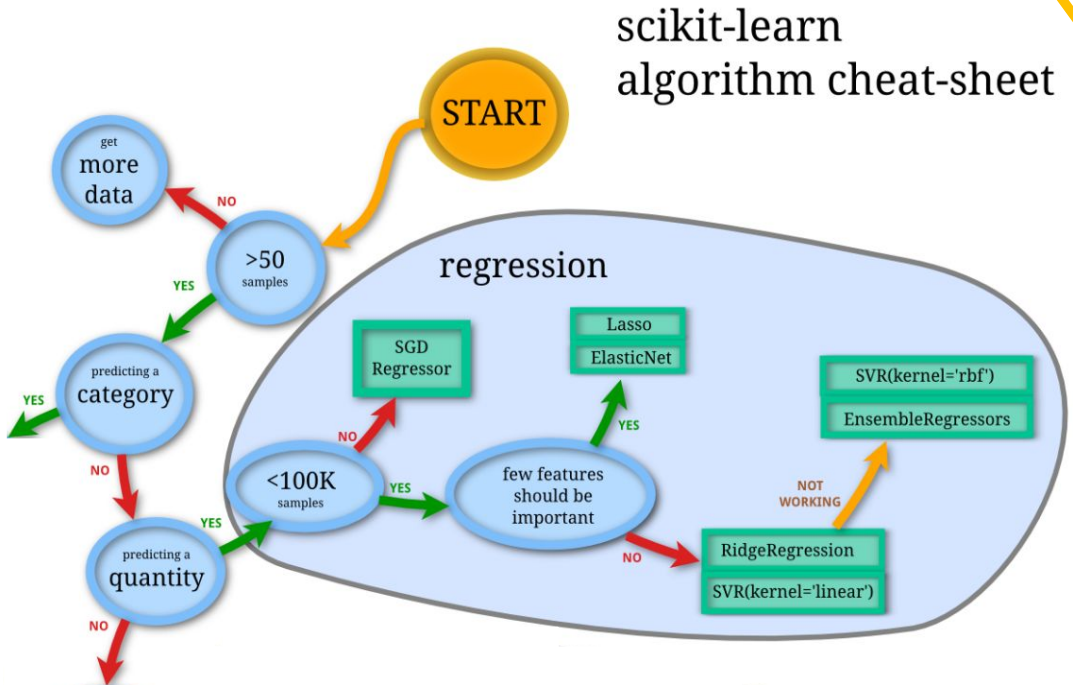
Regrese (projekt)

Praktické projekty zahrnují prezentaci vašich zjištění zainteresovaným stranám nebo členům týmu. Prostřednictvím těchto projektů rozvineme dovednosti v efektivní komunikaci výsledků a jejich vizualizaci, což je klíčové pro předávání komplexních konceptů netechnickému publiku.

- vysvětlete všechny provedené kroky
- diskutujte o potížích, se kterými jste se setkali
- zobrazte srovnání výkonnosti každého modelu
- interpretujte výsledky (včetně koeficientů a statistické významnosti proměnných)

Mít sbírku dokončených projektů prokazuje vaši schopnost používat regresní techniky a zvyšuje vaši důvěryhodnost jako datového vědce nebo analytika.

- poskytujte zdrojový kód ve skriptech .ipynb nebo .py
- nahrávejte své projekty na GitHub
- Prezentujte je ve svém portfoliu nebo při pracovních pohovorech



Cvičení:
Regrese (projekt)

Rozvrh dne

- rozdělení do projektových skupin (max. 2-3 osoby)
- KROK 1:
  - výběr projektu a stažení dat
  - shromáždění požadavků
  - porozumění datům
  - předběžná analýza dat
- SHRNUTÍ 1:
  - podrobný popis a prezentace hypotéz, které mají být testovány
- KROK 2:
  - průzkumná analýza dat (EDA)
  - vizualizace dat
  - zpracování dat
  - aplikace regresních algoritmů
  - hodnocení jejich výkonu
  - možné aplikace
- SHRNUTÍ 2

Cvičení:
Regrese (projekt)

Požadavky

- Projekt lze dodat jako:
  - PowerPoint nebo prezentace Google Slides
  - alt. lze to provést v aplikaci Jupyter Notebook nebo Google Colab
  - Kód Pythonu ve skriptech .ipynb nebo .py
- Práce se dá vyřešit
  - V prostředí vašeho počítače - nejprve nainstalujte všechny potřebné balíčky pomocí příkazu pip
  - pomocí Google Colab
- V určitém okamžiku by měl být kód refaktorován a vyčištěn, např. pomocí PyCharm nebo jiného IDE.
- Kód by měl být sdílen se všemi členy týmu prostřednictvím vzdáleného úložiště git, například na platformě GitHub.

Cvičení:
Regrese (projekt)

Detailní popis

1. Rozdělení do projektových skupin
   - Můžete si vybrat sami nebo vás trenér náhodně rozdělí.
2. Výsledkem tohoto cvičení by měl být krátký 1-2 stránkový **úvod analyzující problém** z obchodního/obsahového hlediska:
   - seznamte se s vybraným souborem dat a popisem úkolu, který má být proveden
   - stáhněte požadovaná data do počítače nebo na Disk Google
   - definujte, co o daném tématu víme
   - načtěte data, seznamte se s jejich strukturou a základními informacemi

Dále uveďte popis toho, co vaše skupina hodlá s datovým souborem dělat a jaká jsou omezení, tzn. co nelze udělat kvůli nedostatku informací nebo příliš velkému počtu odlehlých nebo chybějících hodnot.

Cvičení:
Regrese (projekt)

Detailní popis

3. Příprava dat včetně:
    - extrahujte číselné a kategorické rysy
    - připravte data
        - zbavte se nebo imputujte chybějící/neúplné hodnoty
    - agregujte informace (groupby)
    - vyčistěte data
    - zpracujte data
        - transformace atributů
        - diskretizace
        - škálování
        - shlukování
        - …
    - základní statistiky pro každý atribut
4. Seznam algoritmů strojového učení, které plánujete použít k řešení vašeho problému

Cvičení:
Regrese (projekt)

Detailní popis

5. Vizualizujte data, EDA
    - spojnicové grafy
    - histogramy
    - rozptylové grafy
    - tepelné mapy
    - sloupcové grafy

    Deskriptivní statistika:

    - korelační koeficienty
    - rozptyl, kovariance
    - standardní odchylka
    - variační koeficient
    - statistické rozdělení
    - korelační matice

Cvičení:
Regrese (projekt)

Detailní popis

6. S využitím všech shromážděných informací o problému, zejména popisu dat z předchozí části, navrhněte 3 algoritmy k implementaci.
    - Implementujte algoritmy lineární regrese (pro jednu a více proměnných), polynomiální regrese a rozhodovacího stromu.
    - Zkuste minimalizovat hodnotu nákladové funkce, v případě potřeby pomocí metody gradient nebo jiné funkce zabudované do balíčku sklearn.
    - Zobrazte hodnotu nákladové funkce v grafu pro počátečních 10 a 20 iterací a váhy algoritmu.
    - Proveďte předpovědi pomocí trénovaného modelu.

Cvičení:
Regrese (projekt)

Detailní popis

7. Rozdělte data do trénovací a validační sady
    - Použijte křížovou validaci a leave-one-out (nezapomeňte zamíchat data)
    - Vyhodnoťte výkonnost modelů (koeficient determinace a dvě metody měření chyby predikce)
    - Stanovte řešení problému nadměrného nebo nedostatečného přizpůsobení algoritmu (overfitting a underfitting)
    - Nalezněte kompromis mezi odchylkou a rozptylem (bias a variance) a zobrazte jej v grafu

Cvičení:
Regrese (projekt)

Detailní popis

8. Vyvoďte závěry z předchozích kroků a na 1–2 snímcích vysvětlete, proč jste použili tento konkrétní model, přičemž mějte na paměti:
    - rozlišení lineárních a nelineárních problémů
    - vliv statistik, jako je korelace
    - odlehlé hodnoty
    - váhy algoritmu
    - hodnocení jeho výkonu
    - prediktivní schopnost

Cvičení:
Regrese (projekt)

Detailní popis

9. Odpovězte na otázky na 1-2 snímcích:
    - Jaké problémy jste schopni vyřešit pomocí svého algoritmu?
    - Jak se dá použít? Jak předáte získané informace někomu dalšímu?
    - Jak interpretujete výsledky?
    - Jak interpretujete výkon algoritmu?
    - Jak můžeme model v budoucnu zlepšit?

Cvičení:
Regrese (projekt)

Souhrn

1. Identifikujte obtíže při práci v týmu.
2. Identifikujte, která část způsobila největší problémy.
3. Identifikujte témata výuky, která potřebujete zopakovat..
4. Pomohl vám tento projekt lépe porozumět regresním problémům, které se dříve naučili?
5. Může vám způsob realizace projektu pomoci s nějakými budoucími problémy?

ML v praxi:
Regrese (projekt)

- dokumentace sklearn https://scikit-learn.org/stable/
- Data: https://github.com/matzim95/ML-datasets
- Regrese – kompletní příklad https://towardsdatascience.com/machine-learning-with-python-regression-complete-tutorial-d2c99dc524ec
- Rozhodovací stromy - vysvětlení https://www.youtube.com/watch?v=7VeUPuFGJHk
- Přednáška SVM - MIT (50 minut) https://www.youtube.com/watch?v=_PwhiWxHK8o