**Žilinská univerzita v Žiline**

**Fakulta riadenia a informatiky**

**Data imputation using PCA**

*Semestrálna práca 2*

**Úvod do strojového učenia**

**David Kučera**

Akademický rok 2024/2025

Obsah

[Popis práce 3](#_Toc186646075)

[Stručný popis programu 3](#_Toc186646076)

[Funkcionality 3](#_Toc186646077)

[Metóda nahradzovania chýbajúcich dát 3](#_Toc186646078)

[Metóda PCA 3](#_Toc186646079)

[Prezentácia a zhodnotenie výsledkov výpočtových experimentov 4](#_Toc186646080)

[Experiment porovnania implementácií PCA 4](#_Toc186646081)

[Experiment zmeny počtu komponentov 5](#_Toc186646082)

# Popis práce

Druhá semestrálna práca sa zameriava na implementáciu a výpočtové experimenty s vybranou metódou strojového učenia – **Data imputation using PCA**.

# Stručný popis programu

Program bol implementovaný v programovacom jazyku **Python** vo vývojovom prostredí JetBrains PyCharm. Obsahuje spúšťací **main.py** súbor, v ktorom sa nachádzajú zdrojové kódy programu. Ďalšie potrebné zdrojové kódy sú v súboroch *metrics.py*, *pca.py* a *pca\_sklearn.py*.

## Funkcionality

* načítanie vstupnej dátovej množiny zo súboru,
* deskriptívna štatistika dát,
* doplnenie chýbajúcich dát strednou hodnotou,
* výpočet MAE, RMSE,
* doplnenie chýbajúcich dát pomocou PCA z knižnice sklearn,
* doplnenie chýbajúcich dát pomocou vlastne implementovanej PCA,
* vlastná implementácia metódy PCA,
* štandardizácia dát pre metódu PCA,
* experiment s počtom komponentov,
* experiment porovnania vlastnej implementácie PCA s implementáciou knižnice scikit-learn,
* grafické zobrazenie výsledkov experimentov.

## Metóda nahradzovania chýbajúcich dát

Nastav

Cyklicky sa vykonáva pre :

1. Pomocou PCA sa vypočítajú chýbajúce hodnoty.
2. Tieto hodnoty sa vložia na do matice na miesta, kde chýbali.
3. Vypočíta sa chybovosť po tejto úprave.
   1. Ak platí skonči.
   2. Inak polož a opakuj.

Ako default hodnoty sa použilo ako 10 a  1e-4.

## Metóda PCA

1. Vycentrovanie dát matice.
2. Vypočítanie kovariančnej matice.
3. Vypočítanie vlastných vektorov a hodnôt matice.
4. Zoberie sa n posledných vlastných vektorov, pričom n je počet komponentov pre PCA.
5. Po výbere týchto vektorov sa spraví rekonštrukcia matice.

# Prezentácia a zhodnotenie výsledkov výpočtových experimentov

## Experiment porovnania implementácií PCA

Pri tomto experimente sme rovnakú funkcionalitu implementovali pomocou vlastne naprogramovanej PCA a pomocou implementácie PCA z knižnice **scikit-learn**. Ich výsledky sme porovnali. Týmto experimentom sme tiež overili správnosť našej funkcie. Výsledky experimentu sú uvedené v prehľadnej tabuľke.

Obrázok, na ktorom je text, rad, diagram, vývoj

Automaticky generovaný popisObrázok, na ktorom je rad, diagram, text, vývoj

Automaticky generovaný popis

Grafy metrík implementácie vlastnej PCA

Obrázok, na ktorom je text, rad, diagram, vývoj

Automaticky generovaný popisObrázok, na ktorom je rad, diagram, text, vývoj

Automaticky generovaný popis

Grafy metrík PCA z knižnice sklearn

Na základe výsledok je zjavné, že naša implementácie je rovnako dobrá, ako implementácia z knižnice scikit-learn.

## Experiment zmeny počtu komponentov

V tomto experimente sme sledovali, ako zmena počtu komponentov ovplyvní výsledné metriky RMSE a MAE. Používali sme pri tom vlastnú implementáciu PCA. Výsledky experimentu sú uvedené v prehľadnom grafe. Rozdiely sú uvedené hodnoty imputovaných dát pomocou priemeru a pomocou PCA, teda .

Obrázok, na ktorom je rad, diagram, vývoj, text

Automaticky generovaný popis

Graf experimentu metriky MAE

Obrázok, na ktorom je text, diagram, rad, vývoj

Automaticky generovaný popis

Graf experimentu metriky RMSE

Je zrejmé, že najlepšie výsledky sú pri počte **komponentov = 6**.