

Home page

Dydaktyka – bieżące

- [AiR – Inf1](#)
- [AiR – PAOM](#)
- [AiR – ML](#)
- [Inf – AdvML](#)

- [Prace mgr 2018/2019](#)

Przydatne

- [Jak korzystać z wiki?](#)
- [WiFi AGH](#)
- [UCI AGH](#)

## Regularyzacja

### Wykład

Slajdy:  [Regularyzacja](#)

### Regularyzacja regresji liniowej

Proszę wczytać dane Boston Housing i zapoznać się z poszczególnymi cechami oraz artybutem.

Zbiór danych:  [Boston](#)

```
dataset = pd.read_csv('boston.csv.pdf')

X = dataset.drop('MEDV', axis=1)
y = dataset['MEDV']
```


**Zad 1** Proszę podzielić zbiór danych na treningowy i testowy - 70-30%.

**Zad 2** Proszę stworzyć podstawowy model regresji logistycznej i sprawdzić jego działanie. Skuteczność ok. 74 %.

```
# model - model regress liniowej + fit
print('Blad treningowy: {}'.format(model.score(X_train, y_train)))
print('Blad testowy: {}'.format(model.score(X_test, y_test)))
```

Otrzymana skuteczność nie jest zadawalająca, celem dalszej części zadania będzie uzyskanie większej dokładności zarówno dla danych treningowych i testowych.

W tym celu skorzystamy z `Polynomial features/regression`, czyli aproksymacji wielomianowej. **Zad 3**

- Korzystając z funkcji `StandardScaler()` proszę znormalizować dane
- Aproksymacja funkcji ze stopniem wielomianu 2:  [Polynomial Features](#)

```
steps = [
    ('poly', PolynomialFeatures(degree=2)),
    ('model', LinearRegression())
]
```

```
pipe = Pipeline(steps)
```

```
pipe.fit(X_train, y_train)
```

- Uczenie modelem regresji logistycznej

Oczekiwane rezultaty: skuteczność danych uczących ok. 90%, danych testowych ok. 60 %.

Otrzymane wyniki oznaczają przeuczenie modelu (ang. overfitting).

Konieczne jest zastosowanie metod regularyzacji.

**Zad 4** Proszę wykonać regularyzację metodą Ridge. Dla parametru `alpha=10` skuteczność danych uczących ok. 90%, danych testowych ok. 80 %. Proszę narysować wykres zależności skuteczności od parametry `alpha`. Parametr może przyjmować bardzo małe wartości 0.001 oraz bardzo duże.

**Zad 5** Proszę zastosować metodę regularyzacji Lasso i wyznaczyć optymalną wartość parametru `alpha` z przedziału [0;1]. Skuteczność danych uczących ok. 85%, danych testowych ok. 83 %.

### Regularyzacja regresji logistycznej

Proszę wczytać zbiór danych Breast Cancer Database oraz zapoznać się z bazą danych:  [Breast cancer](#)

Dodatkowe informacje o bazie danych:  [Breast Cancer Wisconsin](#)

```
import os
path = os.getcwd() + '/breast_cancer.txt.pdf'
dataset = pd.read_csv(path, header=None, names=['ID', 'Clump Thickness', 'Uniformity of Cell Size', 'Uniformity of Cell Shape', 'Marginal Adhesion', 'Single Epithelial C
```

Baza danych posiada dwie klasy decyzyjne: zmiana łagodna (benign - 2) oraz złośliwa (malignant - 4). W celu łatwiejszej klasyfikacji zamieniamy wartości na 0 i 1:

```
dataset['Class'].replace(2, 0, inplace=True)
dataset['Class'].replace(4, 1, inplace=True)
```

**Zad 1** Korzystając z funkcji `.isnull()`Sprawdź, czy baza danych nie posiada brakujących wartości. Uzupełnij brakujące wartości wartościami średnimi dla danych klas. Przydatne funkcje: `dataset['Clump Thickness'].fillna(median, inplace=True)`

lub `dataset.replace('?', median, inplace=True)`

**Zad 2** Podziel zbiór danych na cechy X oraz etykietę y (ostatnia kolumna). Zbiór X bez pierwszej kolumny zawierającej ID badanej osoby.

**Zad 3** Podziel zbiór danych na uczący i treningowy.

**Zad 4** Korzystając z dostępnych bibliotek w Pythonie stwórz model regresji logistycznej z regularyzacją L1 i parameterem regularyzacji C=1.

**Zad 5** Narysuj wykres zależności skuteczności algorytmu względem ścieżki regularyzacji L2 (10 wartości parametru regularyzacji C z zakresu [0.0001;1])

### Regularyzacja algorytmu kNN

Korzystając z rozwiązania z laboratorium 3 proszę narysować wykres skuteczności algorytmu w zależności od wartości k. Proszę wskazać optymalną wartość k.

### Regularyzacja drzew deycyzyjnych

Korzystając z rozwiązania z laboratorium 3 proszę narysować wykres skuteczności algorytmu w zależności od głębokości drzewa. Proszę wskazać optymalną głębokość drzewa.

teaching:air-ml:2019!labs:ab05

Table of Contents

- Regularyzacja
- Wykład
- Regularyzacja regresji liniowej
- Regularyzacja regresji logistycznej
- Regularyzacja algorytmu kNN
- Regularyzacja drzew decyzyjnych