Regresja

Uczenie nadzorowane

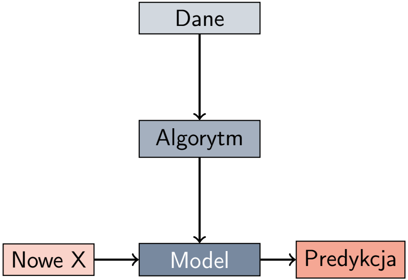
Paweł Gliwny

Wydział Fizyki i Informatyki Stosowanej, UŁ

Uczenie nadzorowane — paradygmat

Schemat:

1. Dane treningowe (wejścia + wyjścia)
2. Algorytm znajduje relacje
3. Tworzy model (równanie)
4. Model przewiduje wynik dla nowych danych



Regresja vs Klasyfikacja

Regresja

- Wyjście = **liczba ciągła**
- Przykłady:
  - Cena domu
  - Temperatura
  - Sprzedaż

Klasyfikacja

- Wyjście = **kategoria**
- Przykłady:
  - Spam / nie spam
  - Kot / pies
  - Choroba: tak / nie

Regresja liniowa — najprostsza metoda ML

- Metoda opracowana ok. 1800 roku
- Najprostszy przykład uczenia nadzorowanego
- Często pierwsze podejście do prognozowania
- Wszechobecna, skuteczna i... często nadużywana

Równanie prostej

$$y = mx + b$$

- m — współczynnik kierunkowy (slope)
- b — wyraz wolny (intercept)

Metoda najmniejszych kwadratów (OLS)

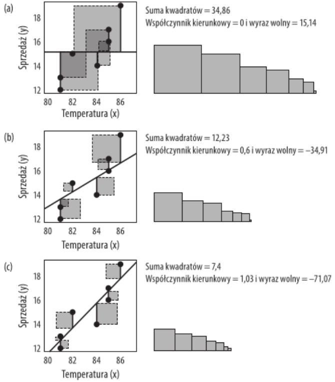
Cel: Znaleźć linię minimalizującą sumę kwadratów błędów

Funkcja kosztu

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Dlaczego kwadraty?

- Błędy dodatnie i ujemne się nie znoszą
- Duże błędy są bardziej karane
- Rozwiązanie analityczne (pochodne $= 0$)



Analityk danych, A. Gutman & J. Goldmeier

Interpretacja współczynników

Przykład: sprzedaż $= 1{,}03 \cdot$ temperatura $- 71{,}07$

Co mówią współczynniki?

- **Współczynnik kierunkowy** $m = 1{,}03$
  Wzrost temperatury o 1°F $\Rightarrow$ wzrost sprzedaży o \$1,03
- **Wyraz wolny** $b = -71{,}07$
  Teoretyczna sprzedaż przy temperaturze 0°F

Kluczowe: Model nie tylko prognozuje, ale **wyjaśnia** związek między zmiennymi.

Rozszerzenie na wiele zmiennych

Regresja wielokrotna (multiple linear regression)

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + \varepsilon$$

- b_0 — wyraz wolny (wartość Y gdy wszystkie $X_i = 0$)
- b_i — zmiana Y przy wzroście X_i o 1 jednostkę
  (przy pozostałych zmiennych stałych)
- ε — błąd losowy

Przykład (cena domu):

$$\text{cena} = b_0 + b_1 \cdot \text{powierzchnia} + b_2 \cdot \text{rok\_budowy} + b_3 \cdot \text{łazienki}$$

Interpretacja — izolowanie wpływu zmiennej

Przykład: Współczynnik dla `YrBuilt` = 818,38 $

Interpretacja

Dom zbudowany **rok później** kosztuje średnio **818$ więcej**
(przy stałej powierzchni, liczbie łazienek itd.)

Uwaga na jednostki!

- +1 sqft powierzchni $\neq$ +1 łazienka
- Współczynniki mają różne skale
- Zawsze pamiętaj o jednostkach zmiennej!

RMSE — główna metryka

Root Mean Squared Error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Właściwości:

- Wyrażone w **jednostkach Y** — łatwa interpretacja
- Reprezentuje typową wielkość błędu predykcji
- **Im niższe, tym lepiej**
- Główna metryka do porównywania modeli

RSE (Residual Standard Error) uwzględnia stopnie swobody — w praktyce różnica minimalna dla dużych zbiorów.

R^2 — współczynnik determinacji

Definicja

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

Interpretacja:

- Zakres: od 0 do 1
- $R^2 = 0{,}80$ oznacza: model wyjaśnia **80%** wariancji Y
- Pozostałe 20% to niewyjaśniona zmienność

Uwaga: Wysokie R^2 na danych treningowych $\neq$ dobry model!
Ryzyko **overfittingu**.

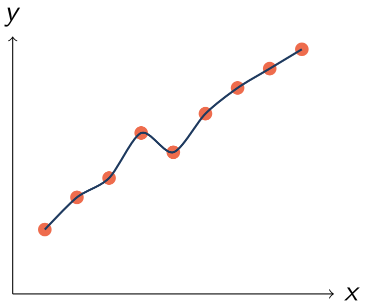
Problem overfittingu

Overfitting (przeuczenie):

- Model "zapamiętuje"dane treningowe
- Świetne wyniki na train
- Słabe wyniki na nowych danych

Przyczyny:

- Za dużo zmiennych
- Za mało danych
- Model zbyt złożony

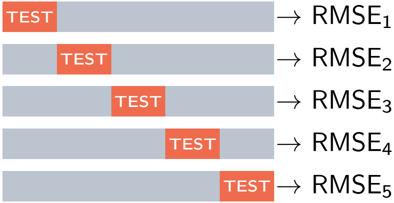


Overfitting

Rozwiązanie: Walidacja krzyżowa (Cross-Validation)

k-Fold Cross-Validation

1. Podziel dane na k części (foldów)
2. Dla każdego foldu: trenuj na $k-1$, testuj na 1
3. Uśrednij wyniki



Wynik = średnia z $\mathrm{RMSE}_1 \ldots \mathrm{RMSE}_5$

Standard: $k = 5$ lub $k = 10$

Ile zmiennych wybrać?

Problem:

- Więcej zmiennych $\Rightarrow$ niższe RMSE na train
- Ale niekoniecznie lepszy model na nowych danych!

Forward Selection

1. Zacznij z 0 zmiennych
2. Dodawaj jedną po drugiej
3. Wybieraj tę, która najbardziej zmniejsza RMSE
4. Stop gdy brak poprawy

Backward Elimination

1. Zacznij ze wszystkimi
2. Usuwaj jedną po drugiej
3. Usuwaj tę, której brak najmniej pogarsza RMSE
4. Stop gdy wszystkie istotne

Problem: regresja wymaga liczb

Zmienne kategoryczne (tekstowe) nie mogą być bezpośrednio w modelu.

Przykład:

`Typ_budynku`: "dom", "blok", "kamienica"

Rozwiązanie: Dummy Variables (zmienne fikcyjne)

Każda kategoria → osobna kolumna binarna (0 lub 1)

Typ	Typ_dom	Typ_blok	Typ_kamienica
dom	1	0	0
blok	0	1	0
kamienica	0	0	1

Reference Coding — prawidłowe podejście

Problem z One-Hot: Kolumny są liniowo zależne!
dom + blok + kamienica = 1 (zawsze)

Rozwiązanie: Usuń jedną kategorię (referencję)

Typ	Typ_blok	Typ_kamienica
dom (referencja)	0	0
blok	1	0
kamienica	0	1

Reguła: P kategorii $\Rightarrow$ $P - 1$ kolumn dummy

Interpretacja z referencją

Model: $\text{cena} = 300000 + (-50000) \cdot \text{blok} + (-20000) \cdot \text{kamienica}$

Interpretacja współczynników

- $b_0 = 300\,000$ — bazowa cena dla **referencji** (dom)
- $b_{\text{blok}} = -50\,000$ — blok jest o 50k **tańszy** niż dom
- $b_{\text{kamienica}} = -20\,000$ — kamienica jest o 20k **tańsza** niż dom

W Pythonie: `pd.get_dummies(df, drop_first=True)`

scikit-learn vs statsmodels

scikit-learn

Cel: Predykcja, ML, produkcja

```
from sklearn.linear_model
  import LinearRegression
model = LinearRegression()
model.fit(X, y)
y_pred = model.predict(X_new)
```

statsmodels

Cel: Statystyka, wnioskowanie

```
import statsmodels.api as sm
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
print(model.summary())
```

Na tym kursie: głównie `scikit-learn`

Metryki w scikit-learn

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import cross_val_score
import numpy as np

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

cv_scores = cross_val_score(model, X, y, cv=5,
      scoring='neg_mean_squared_error')
cv_rmse = np.sqrt(-cv_scores.mean())
```

Podsumowanie — kluczowe punkty

1. **Regresja liniowa** — najprostsza metoda ML, przewiduje wartości ciągłe
2. **OLS** minimalizuje sumę kwadratów błędów
3. **Współczynniki** wyjaśniają wpływ zmiennych (przy innych stałych)
4. **RMSE** i R^2 — główne metryki oceny
5. **Cross-validation** chroni przed overfittingiem
6. **Zmienne kategoryczne** wymagają kodowania (P-1 dummy)

Zadanie praktyczne

AMES Housing Price Prediction

1. Zbudować model regresji wielokrotnej
2. Znaleźć zmienne dające dobry model
3. Zacząć od danych liczbowych
4. (Dla chętnych) Dodać dane kategoryczne

Metryka: RMSE