Algorytmiczna Analiza Danych
Wykład 2
Uczenie nadzorowane

2025-10-09

1. Zbiór danych z etykietami

$$D = \left\{ \left(\bar{x}^{(1)}, y_{(1)} \right), \cdots, \left(\bar{x}^{(n)}, y^{(n)} \right) \right\} \tag{1}$$

$\overline{\boldsymbol{x}}^{(\boldsymbol{i})} = \left(\boldsymbol{x}_{\mathbf{1}}^{(\boldsymbol{i})}, \cdots, \boldsymbol{x}_{\boldsymbol{p}}^{(\boldsymbol{i})} \right)$ - cechy, zmienne wyjaśinające, predyktory
$\boldsymbol{y}^{(\boldsymbol{i})}$ - etykieta, zmienna wyjaśniająca

Problem: wyznaczyć ukrytą relację (ang. underlying realtion) f

$$y^{(i)} = f\left(\bar{x}^{(i)} \right) + \varepsilon^{(i)} \tag{2}$$

$\boldsymbol{\varepsilon}^{(\boldsymbol{i})}$ - szum

1.1. Etykieta

Regresja - zmienna ciągła
Klasyfikacja - zmienna dyskretna

2. Statistical learning vs Machine Learning

Statistical learning	Machine learning
• proste modele z mocnym wsparciem teoretycznym, np. regresja liniowa, • łatwa interpretacja modelu, modele parametryczne	• złożone modele, weryfikowane empirycznie, np. *KNN*, *DT* • zazwyczaj nieparametryczne lub z dużą ilością parametrów
• łatwe wnioskowanie, • nie wymaga mocy obliczeniowej	• duża moc predykcyjna • nie wymaha założeń o danych
• słaba moc predykcyjna, • wymaga silnych założeń o danych, • wymaga aby szum (ε) miał rozkład normalny	• trudne w interpreacji (black box), • wymaga dużej mocy obliczeniowej

$\rightsquigarrow$ wnioskowanie (ang. inference)
$\rightsquigarrow$ predykcja, przewidywanie

3. Przygotowanie danych

1. Brakujące dane

- usuwanie obserwacji,
- zastępowanie:
  - ▸ średnią
  - ▸ medianą
  - ▸ dominantą dla danych dyskretnych
  - ▸ losowanie zgodne z rozkładem
  - ▸ dodatkowa etykieta „unknown”

2. Usuwanie podejrzanych obserwacji
   - outlier (nietypowy),
   - high leverage points,
   - histogram etykiet
3. Inżynieria cech
   - dostosować dane, tak by predykcja była łatwiejsza
4. Podział danych
   - część treningowa $\approx 75\%$,
   - część walidacyjna $\approx 15\%$,
   - część testowa $\approx 10\%$,
   - typowo podział jest losowy

4. Błąd średnio kwadratowy (ang. MSE - Mean Square Error)

$$\begin{aligned} \mathrm{MSE} &= \mathbb{E}\Big(\big(\hat{f}(\bar{x}^{(0)}) - y^{(0)}\big)^2\Big) \\ &= \mathrm{Bias}\Big(\hat{f}(\bar{x}^{(0)})\Big)^2 + \mathrm{Var}\Big(\hat{f}^2(\bar{x}^{(0)})\Big) + \mathrm{Var}(\varepsilon) \end{aligned} \tag{3}$$

$$\mathrm{Bias}(\hat{f}(\bar{x}^{(0)}) = \mathbb{E}\Big(f(\bar{x}^{(0)}) - \hat{f}(\bar{x}^{(0)})\Big) \tag{4}$$