Zaawansowane uczenie maszynowe:
wykład 1

Paweł Cichosz

Informacje o przedmiocie

Strona przedmiotu: `http://elektron.elka.pw.edu.pl/~pcichosz/zum`

Literatura:

1. I. H. Witten, E. Frank, M. A. Hall (2016): *Data Mining: Practical Machine Learning Tools and Techniques* (4th edition).
2. T. Hastie, R. Tibshirani, J. Friedman (2016): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd edition).
3. C. M. Bishop (2006): *Pattern Recognition and Machine Learning*.
4. P. Cichosz (2015): *Data Mining: Explained Using R*.

Zasady zaliczania

Kolokwia: 2×25 punktów

1. w połowie semestru (termin wykładu 8),
2. na końcu semestru (termin wykładu 15).

Projekt: 10 + 40 punktów

1. zespoły dwuosobowe,
2. język R lub Python,
3. ogłoszenie tematów w ciągu 2–3 tygodni,
4. dokumentacja wstępna do końca 7. tygodnia semestru,
5. zakończenie do końca przedostatniego tygodnia semestru.

Formuła wykładu

Przeplatające się tryby:

Slajdy:

- bardziej uporządkowany wywód,
- czasem mniejsza wnikliwość,
- mniej interakcji ze słuchaczami,
- większa dyscyplina czasowa.

Kreda/marker i tablica:

- czasem większa wnikliwość,
- więcej interakcji ze słuchaczami,
- zwiększone ryzyko pomyłek,
- nie zawsze wszystko zmieści się w czasie.

Niezależnie od trybu prowadzenia: dostępne slajdy.

Koncepcja przedmiotu

Cel: pogłębienie i rozszerzenie wiedzy i umiejętności w zakresie uczenia maszynowego.

Założenie: znajomość podstaw teorii i niektórych podstawowych algorytmów...

ale zaczynamy od przypomnienia!

Zakres:

- wybrane nie-neuronowe algorytmy uczenia się
- stosowane do tworzenia modeli predykcyjnych
- na podstawie danych tabelarycznych.

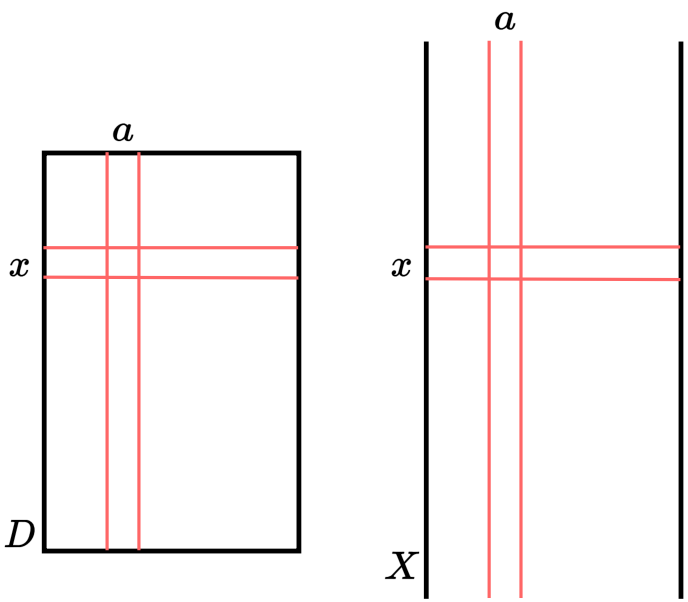
Uczenie się

- Pozyskiwanie lub doskonalenie:
  - **wiedzy**,
  - umiejętności.
- Realizowane z wykorzystaniem informacji trenującej w postaci:
  - **danych** (przykładów),
  - odpowiedzi na zapytania,
  - oceny wykonanych akcji.

Wiedza

- Odnosi się do pewnej dziedziny – zbioru obiektów, osób, przedmiotów, zdarzeń, sytuacji itp.
- Może być użyta do *wnioskowania* (dedukcyjnego): odpowiadania na pytania, podejmowania decyzji itp.
- Może być wynikiem *wnioskowania* (indukcyjnego): uogólniania obserwacji.

Dane



- Reprezentacja tabelaryczna (wiersze, kolumny) pewnego podzbioru dziedziny – zbioru obiektów, osób, przedmiotów, zdarzeń, sytuacji itp.
- Zazwyczaj niezbędny pracochłonny proces przygotowania: pobranie danych z różnych źródeł, połączenie, transformacja, oczyszczenie itp.

Podstawowa terminologia i notacja

Dziedzina: X (także: populacja).

Przykład: $x \in X$ (także: obserwacja, przypadek, rekord, wiersz).

Zbiór danych: $D \subset X$ (także: próba) – dowolny podzbiór dziedziny.

Atrybuty: $a_1, a_2, \ldots, a_n$, $a_i : X \to A_i$ (także: cechy, zmienne) – wektorowa reprezentacja przykładów:

- ciągłe (numeryczne): reprezentacja i interpretacja liczbowa (możliwa arytmetyka),
- dyskretne (kategoryczne): brak interpretacji liczbowej:
  - nominalne: bez relacji porządku (możliwe porównania $=$, $\neq$),
  - porządkowe: z relacją porządku (możliwe porównania $=$, $\neq$, $<$, $\leq$, $>$, $\geq$),
- wektor wartości atrybutów:

$a_1(x)$	$a_2(x)$	$\ldots$	$a_n(x)$

Model

- Obliczeniowa reprezentacja wiedzy uzyskanej na podstawie danych:

  struktura: ustalona lub dopasowana do danych forma reprezentacji zależności,

  parametry: dopasowane do danych wartości konkretyzujące strukturę.
- Zastosowanie modelu: predykcja (zazwyczaj wyróżnionego atrybutu docelowego na podstawie innych atrybutów).

Zadanie klasyfikacji

Pojęcie: $c : X \to C$, C – skończony zbiór klas (kategorii).

Zbiór trenujący: $T \subseteq D \subset X$, dla $x \in D$ znane $c(x)$.

Model: $h : X \to C$, $h \approx c$.

Funkcja decyzyjna (dla niektórych reprezentacji modeli): $g : X \to \mathcal{R}$.

Predykcje probabilistyczne (dla niektórych reprezentacji modeli):
$\pi_d : X \to [0, 1]$, $\pi_d(x) = P(d|x)$, $\pi(x) = P(1|x)$.

Klasa pojęć: $\mathbb{C}$ – zbiór pojęć rozważanych dla pewnej dziedziny (podzbiór wszystkich możliwych pojęć spełniający określone warunki), determinuje złożoność *zadania* uczenia się.

Przestrzeń modeli: $\mathbb{H}$ – zbiór możliwych modeli rozważanych dla pewnej dziedziny (np. wyznaczony przez określenie reprezentacji), determinuje złożoność *algorytmu* uczenia się.

Uczenie się jako przeszukiwanie: wykorzystując T znaleźć model $h \in \mathbb{H}$ dobrze przybliżający pojęcie $c \in \mathbb{C}$.

Zadanie regresji

Funkcja docelowa: $f : X \to \mathcal{R}$.

Zbiór trenujący: $T \subseteq D \subset X$, dla $x \in D$ znane $f(x)$.

Model: $h : X \to \mathcal{R}$, $h \approx f$.

Jakość modelu

Funkcja straty:

dla klasyfikacji: $\mathcal{L} : C \times C \to \mathcal{R}$, gdzie $\mathcal{L}(c(x), h(x))$ jest wartością straty związaną z predykcją $h(x)$ dla przykładu, dla którego prawdziwą klasą jest $c(x$,

dla regresji: $\mathcal{L} : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$, gdzie $\mathcal{L}(f(x), h(x))$ jest wartością straty związaną z predykcją $h(x)$ dla przykładu, dla którego prawdziwą wartością funkcji docelowej jest $f(x$.

Strata zero-jedynkowa (dla klasyfikacji):

$$\mathcal{L}(c(x), h(x)) = \begin{cases} 0 & \text{jeśli } h(x) = c(x) \\ 1 & \text{jeśli } h(x) \neq c(x) \end{cases}$$

Strata kwadratowa (dla regresji):

$$\mathcal{L}(f(x), h(x)) = (f(x) - h(x))^2$$

Jakość modelu

Średnia strata na zbiorze:

$$l_{S,c/f}(h) = \frac{1}{|S|} \sum_{x \in S} \mathcal{L}(c/f(x), h(x))$$

Strata rzeczywista (oczekiwana strata na dziedzinie):

$$l_{\Omega,c/f}(h) = \mathbf{E}\big(\mathcal{L}(c/f(x), h(x)) \mid x \in X, x \sim \Omega\big)$$

gdzie

- c/f oznacza warianty: klasyfikacja (c) albo regresja (f),
- Ω oznacza rozkład prawdopodobieństwa na dziedzinie (możemy pominąć Ω w notacji jeśli rozkład traktujemy jako ustalony).

Jakość modelu

Błąd klasyfikacji na zbiorze: średnia strata zero-jedynkowa:

$$e_{S,c}(h) = \frac{\left|\{x \in S \mid h(x) \neq c(x)\}\right|}{|S|}$$

Błąd rzeczywisty klasyfikacji: oczekiwana strata zero-jedynkowa:

$$e_{\Omega,c}(h) = P(h(x) \neq c(x) \mid x \in X, x \sim \Omega)$$

Błąd średniokwadratowy na zbiorze (MSE, *mean square error*): średnia strata kwadratowa.

Błąd średniokwadratowy rzeczywisty: oczekiwana strata kwadratowa.

Obciążenie indukcyjne

- Właściwości algorytmu uczenia się, które determinują wybór h na podstawie T (sposób uogólnienia przykładów trenujących).
- Konieczne w celu wyboru jednej z wielu możliwych generalizacji.
- Realizacja:

  obciążenie reprezentacji: zawężenie przestrzeni modeli dostępnej dla algorytmu,

  obciążenie preferencji: wprowadzenie kryteriów preferencji modeli (np. preferencja dla najprostszych modeli).

Nadmierne dopasowanie

- Model $h_1 \in \mathbb{H}$ nadmiernie dopasowany do T, jeśli istnieje model $h_2 \in \mathbb{H}$ taki, że $l_{T,c/b}(h_1) < l_{T,c/b}(h_2)$, ale $l_{\Omega,c/b}(h_1) > l_{\Omega,c/b}(h_2)$.
- Zasadnicze wyzwanie przy tworzeniu modeli przez uczenie się.
- Obliczeniowa teoria uczenia się charakteryzuje poziom ryzyka nadmiernego dopasowania w zależności od przestrzeni modeli i właściwości algorytmów.
- Zapobieganie i kontrola:
  - stosowanie algorytmów o mniejszym ryzyku nadmiernego dopasowania,
  - wprowadzanie do algorytmów mechanizmów zmniejszających ryzyko nadmiernego dopasowania,
  - staranna ocena jakości modeli.

Zakres przedmiotu

- Krótki przegląd podstawowych wyników obliczeniowej teorii uczenia się: PAC-nauczalność, wymiar VC.
- Krótki przegląd podstawowych algorytmów uczenia się: drzewa decyzyjne, modele liniowe, naiwny klasyfikator bayesowski, SVM, las losowy, ocena jakości modeli.
- Drugie spojrzenie na drzewa decyzyjne: przycinanie, obsługa brakujących wartości, zastosowanie do regresji.
- Drugie spojrzenie na naiwny klasyfikator bayesowski: model wielomianowy, model dopełnieniowy.
- Drugie spojrzenie na modele liniowe: maksymalizacja logarytmu wiarygodności, regularyzacja.
- Drugie spojrzenie na algorytm SVM: postać dualna, funkcje jądrowe, zastosowanie do regresji.

Zakres przedmiotu

- Drugie spojrzenie na modelowanie zespołowe: bagging, boosting.
- Koszty pomyłek i niezrównoważone klasy.
- Drugie spojrzenie na ocenę jakości modeli: dodatkowe miary jakości, procedury oceny, kryteria selekcji modeli.
- Drugie spojrzenie na zadanie klasyfikacji: klasyfikacja wieloetykietowa, aktywne uczenie się, półnadzorowane uczenie się.
- Grupowanie: miary niepodobieństwa, podstawowe algorytmy, ocena jakości.
- Detekcja anomalii: klasyfikacja jednoklasowa, niepodobieństwo do sąsiadów, niepodobieństwo do grup.
- Selekcja i transformacja atrybutów.