

**Podstawowe pojęcia**

Wykład 2

Marcin Szczuka

<http://www.mimuw.edu.pl/~szczuka/mme/>

*Many mickles make muckle.*

## 0.1 Plan wykładu

- Wprowadzenie podstawowych pojęć.
- Proces uczenia się.
- Ocena rezultatów.
- Przygotowywanie danych wejściowych.

## 0.2 Dziedzina, przykłady

*Dziedzina* (przestrzeń, uniwersum) to pewien zbiór  $X$ , z którego pochodzą (którego elementami są) nasze *przykłady*.

Element  $x \in X$  nazywamy *przykładem* (instancją, przypadkiem, rekordem, entką, wektorem, obiektem, wierszem).

## 0.3 Atrybuty

*Atrybut* (cecha, pomiar, kolumna) to pewna funkcja

$$a : X \rightarrow A.$$

Zbiór  $A$  jest nazywany *dziedziną wartości atrybutu*, lub prościej – dziedziną atrybutu.

## 0.4 Atrybuty c.d.

Zakładamy, że każdy przykład  $x \in X$  jest całkowicie reprezentowany przez wektor

$$a_1(x), \dots, a_n(x),$$

gdzie

$$a_i : X \rightarrow A_i$$

dla  $i = 1, \dots, n$ .  $n$  nazywamy czasem rozmiarem (długością) przykładu.

W pewnych zastosowaniach wyróżniamy specjalny atrybut nazywany *decyzją* (klasą) lub *atrybutem decyzyjnym*.

## 0.5 Atrybuty nominalne

Nominalne (dyskretne, skończone, wyliczeniowe) - przestrzeń wartości atrybutu jest skończona (i niewielka) oraz nie występuje porządek na wartościach atrybutu.

Na przykład atrybut *Pe* ma przestrzeń wartości  $\{M, K, NW\}$ .

## 0.6 Atrybuty porządkowe

Porządkowe (uporządkowane) - przestrzeń wartości atrybutu jest skończona (i zwykle niewielka), ale można wprowadzić porządek na wartościach atrybutu.

Na przykład atrybut *Wysokość* może przyjmować wartości  $\{Zero, Nisko, Kosmos, Wysoko, Średnio, Bardzo Wysoko\}$  i wiemy, że:

$Zero < Nisko < Średnio < Wysoko < Bardzo Wysoko < Kosmos$

Zauważmy jednak, że znajomość porządku nie oznacza znajomości odległości pomiędzy wartościami atrybutu.

## 0.7 Atrybuty ciągłe

Ciągłe (numeryczne) - przestrzeń atrybutu jest dobrze zdefiniowanym zbiorem liczbowym np. liczby całkowite czy rzeczywiste. Znamy porządek na wartościach atrybutu, a zwykle także odległość. Możliwe są operacje algebraiczne na wartościach atrybutów.

Na przykład *Temperatura* może być wyrażona w stopniach Celsjusza (lub Kelwina) i wiemy, że różnica między  $-1^{\circ}C$  a  $35^{\circ}C$  wynosi  $36^{\circ}$ .

## 0.8 Typy atrybutów

Rozróżnienie między rodzajami atrybutów nie jest ścisłe. Istnieją także inne rodzaje atrybutów, które możemy chcieć wyróżniać:

- Binarne - przyjmują tylko dwie wartości (zwykle 0 i 1).
- Przedziałowe - uporządkowane i mierzone w ustalonych przedziałach (data, temperatura).

- Stosunek - określone przez odniesienie do punktu referencyjnego (odległość od stacji kolejowej).

## 0.9 Zbiory danych - dane tablicowe

Outlook	Temp	Humid	Wind	EnjoySpt
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
⋮	⋮	⋮	⋮	⋮
rainy	mild	high	TRUE	no

## 0.10 Zbiory danych - dane tablicowe

Outlook	Temp	Humid	Wind	EnjoySpt
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
⋮	⋮	⋮	⋮	⋮
rainy	71	91	TRUE	no

## 0.11 Pojęcia

*Pojęcie* (pojęcie docelowe) to przyporządkowanie

$$c : X \rightarrow \{0, 1\}.$$

Równoważne określenie - pojęcie jest podzbiorem  $X_c \subseteq X$ .

W praktyce rozważamy pojęcia *wielokrotne* tzn.  $c : X \rightarrow C$  dla  $|C| > 2$ .

Dla przykładu  $x \in X$  wartość  $c(x)$  nazywamy *etykietą* (decyzją, kategorią, klasą)  $x$ -a.

### 0.12 Przestrzeń pojęć

*Przestrzeń pojęć* (klasa pojęć) jest to rodzina  $C$  wszystkich pojęć dla dziedziny  $X$

Jeśli posłużymy się definicją pojęcia jako podzbioru, możemy powiedzieć, że dla skończonej dziedziny  $C \subseteq 2^X$ .

### 0.13 Hipotezy

*Hipoteza* to funkcja

$$h : X \rightarrow \{0, 1\}$$

reprezentująca przybliżenie pojęcia docelowego uzyskane w wyniku uczenia.

W przypadku pojęć wielokrotnych definicja hipotezy ulega odpowiedniej modyfikacji.

### 0.14 Przestrzeń hipotez

*Przestrzeń hipotez* jest to zbiór  $H$  wszystkich hipotez, które mogą być zbudowane w procesie uczenia.

Kształt tego zbioru zależy od sposobu reprezentacji hipotez i wybranego algorytmu uczenia.

Chcemy faworyzować podejścia w których  $C \subseteq H$  tzn.  $c \in H$ . W takim przypadku pojęcie docelowe jest *wyuczalne*. Niestety, przeważnie dostajemy  $H \subset C$

### 0.15 Przykłady etykietowane, klasy

*Przykład etykietowany* (obiekt z decyzją) dla pojęcia  $c$  jest parą  $\langle x, c(x) \rangle$ ,  $x \in X$ .

*Przykład pozytywny* dla  $c$ :  $x \in X$ ,  $c(x) = 1$ .

*Przykład negatywny* dla  $c$ :  $x \in X$ ,  $c(x) = 0$ .

*Klasa* (klasa decyzyjna, kategoria) przykładów z dziedziny:  $C_i^c = \{x \in X | c(x) = i\}$ . Przeważnie używamy zapisu  $C_i$ .

### 0.16 Zbiór treningowy

*Zbiór treningowy* (próbka treningowa/ucząca) dla nadzorowanego (z nauczycielem) wyuczania pojęcia  $c$  jest to zbiór

$$T_c = \{\langle x, c(x) \rangle | x \in T \subseteq X\}.$$

W większości przypadków, gdy  $c(\cdot)$  jest ustalone, używamy prostszego oznaczenia  $T$ .

*Zbiór treningowy* (próbka treningowa/ucząca) dla uczenia nienadzorowanego (bez nauczyciela) to podzbiór  $T \subseteq X$ .

### 0.17 Błąd próbki

*Błąd próbki* dla hipotezy  $h$  ze względu na pojęcie  $c$  i zbiór przykładów  $D$  określamy jako:

$$e_D^c(h) = \frac{\sum_{x \in D} \delta(h(x), c(x))}{|D|},$$

gdzie

$$\delta(h(x), c(x)) = \begin{cases} 1 & \text{gdy } c(x) \neq h(x) \\ 0 & \text{w.p.p.} \end{cases}$$

### 0.18 Błąd rzeczywisty

*Błąd rzeczywisty* (całkowity, globalny) dla hipotezy  $h$  ze względu na pojęcie  $c$  i rozkład prawdopodobieństwa  $\Omega$  na  $X$  określamy jako:

$$e_\Omega^c(h) = \Pr_{x \in \Omega}(h(x) \neq c(x)).$$

Wykorzystując metody rodem ze statystyki, możemy próbować wyestymować błąd rzeczywisty używając błędu próbki dla różnych próbek.

### 0.19 Macierz błędu

*Macierz błędu* jest powszechnie wykorzystywana do przedstawiania rozkładu błędów w przypadku hipotez (i pojęć) wielokrotnych. Macierz błędu to macierz kwadratowa rozmiaru  $n \times n$

$$\mathbf{E} = \begin{pmatrix} e_{11} & \dots & e_{1n} \\ \vdots & \ddots & \vdots \\ e_{n1} & \dots & e_{nn} \end{pmatrix}$$

W macierzy tej  $e_{ij}$  jest liczbą przykładów w próbce dla których  $c(x) = i$  i  $h(x) = j$ ;  $i, j = 1, \dots, n$ .  
 $n$  jest liczbą klas decyzyjnych.

## 0.20 Indukcyjne uczenie z nadzorem

Mając zbiór treningowy  $T_c$  znajdź taką hipotezę  $h \in H$ , która najlepiej aproksymuje pojęcie docelowe  $c$  ze względu na ustalone kryterium.

*Kryterium* jest zwykle (choć nie wyłącznie) oparte na błędzie (błędach) próbki. Dokładne (naiwne) kryterium formułuje się jako:

$$\forall_{x \in X} (h(x) = c(x))$$

## 0.21 Indukcyjne uczenie bez nadzoru

Mając zbiór treningowy  $T_c$  znajdź taką hipotezę  $h \in H$ , która daje najlepszą klasyfikację przykładów ze względu na ustalone kryterium.

*Kryterium* jest zwykle bardzo zależne od konkretnego zadania.

## 0.22 Obciążenie indukcyjne, założenie indukcyjne

Obciążenie indukcyjne – preferencje w wyborze pewnych hipotez przez system uczący się. Kombinacja wszystkich czynników, które w połączeniu ze zbiorem treningowym determinują wybór ostatecznej hipotezy.

Założenie indukcyjne – hipoteza, która jest akceptowalna dla dostatecznie dużej próbki treningowej, jest także dopuszczalna dla całej dziedziny.

## 0.23 Założenie zamkniętości świata

*Closed world assumption* (CWA) - jeśli nie potrafimy zweryfikować, czy jakiś przykład jest pozytywny, czy negatywny to przyjmujemy, że jest negatywny.

## 0.24 Uczenie w praktyce

- Zapoznaj się z danymi
- Przygotuj zbiór treningowy (i testowy).
- Zastosuj algorytm uczący się.
- Oceń uzyskane hipotezy.
- Powtarzaj powyższe tyle razy, ile trzeba.

### 0.25 Zbiór testowy, próbka testowa

Hipotezy wyuczone na podstawie samego zbioru treningowego mogą odzwierciedlać zależności, które są charakterystyczne tylko dla tej właśnie próbki. Aby zapewnić sobie właściwy poziom ogólności, powinniśmy stosować procedury testowe. Dlatego często fragment etykietowanych danych nie jest wykorzystywany w uczeniu, lecz używany jako próbka testowa.

Efekt wyuczania się zależności lokalnych, zbyt szczegółowych i charakterystycznych jedynie dla próbki nazywamy *przeuczeniem* (ang. overfitting).

### 0.26 Ocena błędu

W większości zastosowań uczenia z nadzorem wykorzystujemy poziom błędu jako kryterium oceny. Jest to po prostu błąd próbki, ale rozpatrywany zarówno dla zbioru treningowego, jak i walidacyjnego oraz testowego. W przypadku wielokrotnych wartości decyzji posługujemy się odpowiednimi macierzami błędów.

### 0.27 Przygotowanie próbek

Popularne metody wyboru próbki treningowej i testowej:

- Train-and-test.
- Cross-validation – walidacja krzyżowa, kroswalidacja.
- Leave-one-out.
- Bootstrap.

### 0.28 Train-and-test

To najbardziej komfortowa sytuacja. Podział na część treningową i testową jest zadany.

**Uwaga!** Próbka treningowa i/lub testowa mogą być kiepsko dobrane.

### 0.29 Cross-validation

k-krotna walidacja krzyżowa (k-fold cross-validation):

1. Podziel dane na k równych części (np. losowo).
2. Wybierz jedną z części jako zbiór testowy, a sumę pozostałych k-1 jako zbiór treningowy.

3. Wykonaj uczenie i testowanie.
4. Powtarzaj 2-3 dla każdej z  $k$  części.

### 0.30 Cross-validation

Najczęściej wykorzystywane: 10-krotna, 3-krotna i 5-krotna walidacja krzyżowa.

Dla zapewnienia lepszej stabilności warto jest powtórzyć wielokrotnie walidację krzyżową (np. 10 razy 10-krotną) i przyjąć średnią z tych powtórzeń jako wynik. Niestety, w wielu praktycznych sytuacjach nie możemy sobie pozwolić na wykonanie tylu eksperymentów.

### 0.31 Leave-one-out

Technika leave-one-out to w zasadzie  $n$ -krotna walidacja krzyżowa, gdzie  $n$  jest liczbą przykładów którymi dysponujemy. Jest to kosztowna technika, gdyż wymaga  $n$  wykonań algorytmu uczącego. Jest polecana tylko dla zadań, w których liczba przykładów jest relatywnie mała.

### 0.32 Bootstrap

Wybieramy losowo **ze zwracaniem**  $n$  przykładów ze zbioru  $T$ , dla  $n = |T|$ . Te przykłady stają się zbiorem treningowym.

Wszystkie przykłady, które nie zostały wybrane do zbioru treningowego tworzą zbiór testowy.

Wyliczamy błąd próbki dla zbioru treningowego i testowego, a następnie składamy te błędy.

### 0.33 0.632 bootstrap

Szansa nie wybrania jakiegoś przykładu w  $n$  próbach ze zwracaniem wynosi:

$$\left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e} = 0.36787944$$

Zatem, aby oszacować całkowity błąd wykorzystujemy formułę:

$$e = 0.632 \cdot e_{test} + 0.368 \cdot e_{train}.$$



### 0.34 Przygotowywanie przykładów

Typowe zadania przy przygotowywaniu próbki do uczenia:

- Zadbanie o brakujące wartości.
- Wyrównywanie i wygładzanie (aligning and presmoothing).
- Dyskretyzacja.
- Wybór atrybutów.

### 0.35 Brakujące wartości

Wyróżniamy dwa podstawowe typy brakujących wartości:

- Niosące informacje.
- Nieniosące informacji.

### 0.36 Uzupełnianie brakujących wartości

Proces wypełniania brakujących wartości nazywamy *uzupełnianiem*.

Metody bezpośrednie:

- Usuwanie obiektów z brakującymi wartościami. Dość drastyczne i nie zawsze dopuszczalne.
- Zastępowanie braków pewną **specjalną** wartością np. -1 dla atrybutu o wartościach naturalnych czy  $\infty$  dla rzeczywistych. Działa dobrze tak długo jak metoda uczenia nie zaczyna produkować nonsensownych hipotez.

$$(hair = blonde) \wedge (temp. = -1) \Rightarrow (Heat = Yes)$$

### 0.37 Uzupełnianie w oparciu o dane

Metody uzupełniania w oparciu o dane:

- Metody lokalne. Zastęp braki wykorzystując informację o rozkładzie jednego atrybutu.
- Metody globalne. Zastęp braki wykorzystując informację o rozkładzie wszystkich atrybutów (całych danych).

### 0.38 Metody lokalne

Najpopularniejsze metody lokalne:

- Ustalamy **domyślną** wartość dla atrybutu.
- Zastępujemy brakujące wartości medianą po wartościach atrybutu występujących w próbce. Ta metoda działa pod warunkiem, że mediana ma jakąś interpretację. Wzięcie zwykłej średniej nie jest zalecanym posunięciem. Dla atrybutów dyskretnych/tekstowych możemy zamiast mediany wstawić najczęściej występującą wartość.

### 0.39 Metody lokalne

Zastępowanie wielokrotne (uzupełnianie kombinatoryczne). Uzupełniany obiekt jest zastępowany przez zbiór nowych, po jednym na każdą możliwą wartość brakującego atrybutu.

To podejście sprawdza się tylko dla danych o stosunkowo niewielkiej liczbie brakujących wartości i atrybutów o małej liczbie możliwych wartości. Dla większych danych i skomplikowanych atrybutów może dojść do **eksplozji kombinatorycznej**.

### 0.40 Metody globalne

Estymujemy globalny rozkład przykładów na podstawie próbki treningowej. Następnie wstawiamy w brakujące miejsce wartość, która jest najbardziej prawdopodobna według tak wyznaczonego rozkładu. Możemy mieć kontrolę nad parametrami estymatora, a także nakładać inne dodatkowe ograniczenia.

Jedna z najbardziej popularnych metod estymacji jest algorytm EM (Expectation Maximization) - technika znajdowania estymacji dla wielowymiarowych złożonych rozkładów gaussowskich. Technikę tą omówimy później w kontekście grupowania pojęciowego.

### 0.41 Wyrównywanie, formatowanie

Wyrównywanie i formatowanie danych (ang. aligning) – w wielu przypadkach dokonujemy pewnych prostych transformacji atrybutów, aby zapewnić sobie właściwe działanie algorytmu uczącego.

Typowe przykłady:

- Normalizacja - sprowadzenie do przedziału  $[0,1]$ .

- Przesunięcie wartości o stałą.
- Zmiana skali na logarytmiczną.

## 0.42 Wygładzanie

Wygładzanie (ang. presmoothing) - gdy wartości atrybutu są nieprecyzyjne, zakłócone lub podlegają nadmiernym wahaniom. W takich sytuacjach zbiór wartości powinien zostać “wygładzony”. Nazwa pochodzi ze statystyki.

Typowe zastosowanie - analiza szeregów czasowych.

## 0.43 Dyskretyzacja

Niektóre metody uczenia albo wprost zakładają, że zbiór wartości atrybutu jest dyskretny, albo działają lepiej gdy tak jest. Dlatego pojawia się zapotrzebowanie na metody odpowiedniej zamiany zbiorów wartości atrybutów.

Dyskretyzacja – zwykle rozumiana jako zamiana atrybutów ciągłych na dyskretnie (symboliczne).

Grupowanie lub kwantyzacja - zmniejszanie rozmiaru przestrzeni wartości atrybutu (zwykle dyskretnego) przez sklejenie pewnych jego wartości.

## 0.44 Dyskretyzacja

Dwie metody kategoryzacji algorytmów dyskretyzacji (grupowania):

- Arbitralne vs. oparte na mierze.
- Lokalne vs. globalne.

## 0.45 Nadzorowane vs. oparte na mierze

Bez nadzoru (arbitralne) - Dzielimy przestrzeń atrybutu na mniejsze kawałki w sposób arbitralny, bez oglądania się na zależności między atrybutami i na decyzję (etykietę).

Nadzorowane (oparte na mierze) - Wykorzystujemy pewną numeryczną miarę wyliczaną dla zbioru danych, aby ocenić czy dany podział atrybutu poprawia sytuację.

Przykład: dyskretyzacja wykorzystująca miarę entropii i regułę MDL (Minimal Description Length).

#### 0.46 Lokalne vs. globalne

Lokalne – jeden atrybut na raz. Zmienia się zbiór wartości tylko jednego atrybutu.

Globalne – różne (czasem wszystkie) atrybuty na raz. Przebudowa przykładów, możliwa jest konstrukcja nowych i/lub usuwanie istniejących atrybutów.

#### 0.47 Wybór atrybutów

Fakt – niewłaściwe atrybuty obniżają jakość działania algorytmów uczących się.

Fakt – nie da się wywnioskować sensownej hipotezy z bezsensownych danych.

#### 0.48

*Theory in when we know everything  
and nothing works.*

*Practice is when everything works  
and no one knows why.*

*We combine theory with practice,  
nothing works and no one knows why.*