**Nazwa przedmiotu:**

Metody odkrywania wiedzy

**Koordynator przedmiotu:**

Paweł CICHOSZ

**Status przedmiotu:**

Fakultatywny dowolnego wyboru

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Informatyka

**Grupa przedmiotów:**

Przedmioty techniczne - zaawansowane

**Kod przedmiotu:**

MOW

**Semestr nominalny:**

2 / rok ak. 2014/2015

**Liczba punktów ECTS:**

4

**Liczba godzin pracy studenta związanych z osiągnięciem efektów uczenia się:**

1. udział w wykładach: 15 x 2 godz. = 30 godz.
2. przygotowanie do kolejnych wykładów i realizacji projektu (przejrzenie materiałów z wykładu i dodatkowej literatury, próba rozwiązania ćwiczeń domowych sformułowanych na wykładzie): 15 x 20 min. = 5 godz.
2. udział w konsultacjach związanych z realizacją projektu:15 godz.
3. realizacja zadań projektowych: 30 godz. (w tym zapoznanie się z literaturą, analiza zadania, implementacja algorytmów, strojenie parametrów, przeprowadzenie badań, sporządzenie dokumentacji)
4. przygotowanie do kolokwiów: 2 x 10 godz. = 20 godz.
Razem: 30 + 5 + 15 + 30 + 20 = 100 godz.

**Liczba punktów ECTS na zajęciach wymagających bezpośredniego udziału nauczycieli akademickich:**

1. udział w wykładach: 15 x 2 godz. = 30 godz.
2. udział w konsultacjach związanych z realizacją projektu:15 godz.
Razem: 30 + 15 = 45 godz. co daje ok. 2 ECTS

**Język prowadzenia zajęć:**

polski

**Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym:**

1. udział w konsultacjach związanych z realizacją projektu:15 godz.
2. realizacja zadań projektowych: 30 godz. (w tym zapoznanie się z literaturą, analiza zadania, implementacja algorytmów, strojenie parametrów, przeprowadzenie badań, sporządzenie dokumentacji)
Razem: 15 + 30 = 45 godz., co daje ok. 2 ECTS

**Formy zajęć i ich wymiar w semestrze:**

|  |  |
| --- | --- |
| Wykład: | 30h |
| Ćwiczenia: | 0h |
| Laboratorium: | 0h |
| Projekt: | 15h |
| Lekcje komputerowe: | 0h |

**Wymagania wstępne:**

podstawowa wiedza z zakresu metod probabilistycznych
podstawowa wiedza z zakresu informatyki
podstawowa umiejętność programowania

**Limit liczby studentów:**

60

**Cel przedmiotu:**

1. Zapoznanie studentów z najważniejszymi algorytmami stosowanymi do odkrywania wiedzy w danych, czyli odkrywania występujących w danych zależności i formułowania ich w postaci umożliwiającej wnioskowanie.
2. Zapoznanie studentów z zasadami stosowania algorytmów odkrywania wiedzy.
3. Ukształtowanie umiejętności w zakresie stosowania algorytmów odkrywania wiedzy do analizowania rzeczywistych zbiorów danych.

**Treści kształcenia:**

Wprowadzenie.
Informacje o przedmiocie. Sformułowanie zadania odkrywania wiedzy. Charakterystyka procesu odkrywania wiedzy. Rodzaje wiedzy. Terminologia i notacja. Przykładowe zastosowania.
Narzędzia ze statystyki i teorii informacji.
Charakterystyka rozkładów atrybutów dyskretnych i ciągłych. Estymacja parametrów rozkładów. Identyfikacja wartości odstających. Estymacja przedziałowa. Testy statystyczne. Statystyczne i teorioinformacyjne miary zależności.
Zadanie klasyfikacji.
Sformułowanie zadania. Podstawowe miary jakości modeli. Nadmierne dopasowanie. Przykładowe zastosowania.
Indukcja drzew decyzyjnych.
Reprezentacja modelu. Zstępująca konstrukcja drzewa. Kryteria stopu. Kryteria wyboru testu. Przycinanie drzew decyzyjnych. Probabilistyczne drzewa decyzyjne. Przetwarzanie przykładów z brakującymi wartościami atrybutów.
Klasyfikacja probabilistyczna.
Twierdzenie Bayesa. Naiwny klasyfikator bayesowski. Minimalizacja prawdopodobieństwa pomyłki. Minimalizacja kosztów pomyłek. Zasada minimalnej długości kodu.
Ocena modeli klasyfikacji.
Analiza kosztów pomyłek. Wskaźniki jakości oparte na macierzy pomyłek. Analiza ROC. Procedury oceny modeli.
Zadanie regresji.
Sformułowanie zadania. Podstawowe miary jakości modeli. Nadmierne dopasowanie. Przykładowe zastosowania.
Drzewa regresji.
Reprezentacja modelu. Kryteria stopu. Kryteria wyboru testu. Drzewa modeli.
Metody pamięciowe.
Algorytm najbliższych sąsiadów. Miary odległości. Lokalne modelowanie.
Modele parametryczne.
Reprezentacja parametryczna. Algorytm spadku gradientu. Model liniowy. Metoda najmniejszych kwadratów. Klasyfikacja liniowa. Przezwyciężania ograniczenia liniowości.
Maszyny wektorów nośnych i metody jądrowe.
Margines klasyfikacji liniowej. Maksymalizacja marginesu klasyfikacji. Algorytmy SVM i SVR. Sztuczka jądrowa. Typy funkcji jądrowych.
Agregacja modeli.
Koncepcja i motywacja. Techniki agregacji: bagging, boosting, las losowy.
Modyfikacja reprezentacji.
Dyskretyzacja atrybutów ciągłych. Numeryczne kodowanie atrybutów dyskretnych. Selekcja atrybutów przez filtrowanie. Selekcja atrybutów przez opakowywanie algorytmów modelowania.
Zadanie grupowania.
Sformułowanie zadania. Grupowanie płaskie i hierarchiczne. Typy algorytmów grupowania. Przykładowe zastosowania.
Grupowanie na podstawie niepodobieństwa.
Miary niepodobieństwa. Rodzina algorytmów k środków. Hierarchiczne grupowanie wstępujące. Ocena modeli grupowania.
Sieci bayesowskie.
Definicja sieci bayesowskiej. Założenia o warunkowej niezależności. Reprezentacja łącznego rozkładu prawdopodobieństwa. Wnioskowanie w sieciach bayesowskich. Nie-całkiem-naiwny klasyfikator bayesowski.
Odkrywanie wiedzy w tekście.
Reprezentacja wektorowa. Klasyfikacja tekstu. Grupowanie tekstu.
Zakres projektu:
Proponowane tematy projektów są podzielone na dwa rodzaje: analityczne oraz implementacyjno-analityczne. Środowiskiem do realizacji tematów obydwu rodzajów jest język R, zapoznanie się z którym jest efektem ubocznym wykonania projektu. Realizacja tematów analitycznych obejmuje pełen proces wnikliwej i szeroko zakrojonej analizy rzeczywistych danych oraz budowy modeli z wykorzystaniem szeregu algorytmów, których implementacje są dostępne w pakietach języka R. Zakres niezbędnych prac może obejmować przygotowanie danych, statystyczny opis danych, zmianę reprezentacji (modyfikację zestawu atrybutów) oraz tworzenie modeli i ocenę ich jakości. Tematy implementacyjno-analityczne obejmują część implementacyjną i część analityczną. Część implementacyjna polega albo na samodzielnej implementacji pewnego algorytmu, albo na modyfikacji istniejącej implementacji pewnego algorytmu. Część analityczna polega na eksperymentalnym zademonstrowaniu działania zaimplementowanego/zmodyfikowanego algorytmu w zastosowaniu do kilku wybranych zbiorów danych, w porównaniu z innymi algorytmami dostępnymi w R realizującymi to samo zadanie.

**Metody oceny:**

Ocena sumatywna oparta na wynikach kolokwiów, wstępnej dokumentacji oraz realizacji projektu.
Ocena formatywna na podstawie interakcji ze studentami w czasie wykładu, rozwiązań ćwiczeń domowych formułowanych na wykładzie oraz pytań w ramach konsultacji.

**Egzamin:**

nie

**Literatura:**

1. Witten, I.H., Frank, E., Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
2. Cichosz, P. (2000, 2007). Systemy uczące się. WNT.
3. Publikacje w czasopismach (m.in. Machine Learning, Journal of Machine Learning Research, Artificial Intelligence, Journal of Artificial Intelligence Research) i materiałach konferencji (m.in. International Conference on Machine Learning).

**Witryna www przedmiotu:**

http://www.ise.pw.edu.pl/~cichosz/mow

**Uwagi:**

## Efekty przedmiotowe

### Profil ogólnoakademicki - wiedza

**Efekt MOW\_W1:**

ma wiedzę na temat najważniejszych zadań i etapów procesu odkrywania wiedzy

Weryfikacja:

kolokwium

**Powiązane efekty kierunkowe:** K\_W04, K\_W06

**Powiązane efekty obszarowe:** T2A\_W02, T2A\_W04

**Efekt MOW\_W2:**

ma wiedzę dotyczącą podstawowych technik opisu danych i wykrywania zależności

Weryfikacja:

kolokwium

**Powiązane efekty kierunkowe:** K\_W04, K\_W06, K\_W08

**Powiązane efekty obszarowe:** T2A\_W02, T2A\_W04, T2A\_W07

**Efekt MOW\_W3:**

ma wiedzę na temat mechanizmów działania i zasad stosowania algorytmów tworzenia modeli predykcyjnych

Weryfikacja:

kolokwium

**Powiązane efekty kierunkowe:** K\_W04, K\_W06

**Powiązane efekty obszarowe:** T2A\_W02, T2A\_W04

**Efekt MOW\_W4:**

ma wiedzę dotyczącą metod oceny jakości modeli predykcyjnych

Weryfikacja:

kolokwium

**Powiązane efekty kierunkowe:** K\_W06

**Powiązane efekty obszarowe:** T2A\_W04

### Profil ogólnoakademicki - umiejętności

**Efekt MOW\_U1:**

potrafi planować proces odkrywania wiedzy, dostosowując zakres analizy i dobierając algorytmy w zależności od potrzeb właściwości danych

Weryfikacja:

projekt

**Powiązane efekty kierunkowe:** K\_U01, K\_U06, K\_U12, K\_U13

**Powiązane efekty obszarowe:** T2A\_U01, T2A\_U08, T2A\_U09, T2A\_U17, T2A\_U18

**Efekt MOW\_U2:**

potrafi stosować algorytmy odkrywania wiedzy do rzeczywistych danych, intepretować uzyskane wyniki i poszukiwać możliwości ich poprawienia

Weryfikacja:

projekt

**Powiązane efekty kierunkowe:** K\_U01, K\_U06, K\_U09, K\_U11

**Powiązane efekty obszarowe:** T2A\_U01, T2A\_U08, T2A\_U09, T2A\_U11, T2A\_U16

**Efekt MOW\_U3:**

potrafi posługiwać się językiem R i jego pakietami do analizy danych oraz implementacji i stosowania algorytmów odkrywania wiedzy, a także pozyskiwać informacje niezbędne do samodzielnego rozwiązywania związanych z tym problemów technicznych

Weryfikacja:

projekt

**Powiązane efekty kierunkowe:** K\_U01, K\_U04, K\_U06, K\_U13

**Powiązane efekty obszarowe:** T2A\_U01, T2A\_U05, T2A\_U08, T2A\_U09, T2A\_U18

**Efekt MOW\_U4:**

potrafi przedstawić plan, przyjęte założenia i wyniki procesu odkrywania wiedzy w formie ustnej i pisemnej

Weryfikacja:

projekt

**Powiązane efekty kierunkowe:** K\_U02, K\_U03

**Powiązane efekty obszarowe:** T2A\_U02, T2A\_U03

### Profil ogólnoakademicki - kompetencje społeczne

**Efekt MOW\_K1:**

potrafi przezwyciężać problemy związane ze złożonością rozwiązywanych zadań odkrywania wiedzy oraz stosowanymi algorytmami i narzędziami

Weryfikacja:

projekt

**Powiązane efekty kierunkowe:** K\_K01

**Powiązane efekty obszarowe:** T2A\_K06