

## KARTA PRZEDMIOTU

**Nazwa przedmiotu:** Wnioskowanie statystyczne (BioAu>SM1WS20)

**Nazwa w języku polskim:**

**Nazwa w jęz. angielskim:** Statistical Inference

### Dane dotyczące przedmiotu:

**Jednostka oferująca przedmiot:** Wydział Automatyki, Elektroniki i Informatyki

**Przedmiot dla jednostki:** Politechnika Śląska

### Domyślny typ protokołu dla przedmiotu:

EGZ

### Język wykładowy:

polski

### Strona WWW:

<https://platforma2.polsl.pl/rau1/course/view.php?id=139>

### Skrócony opis:

Celem wykładu jest przekazanie studentom podstawowych wiadomości w zakresie budowy statystycznych systemów wnioskujących oraz metod analizy danych pozyskiwanych za pomocą technik wysokoprzepustowych (High-throughput screening HTS).

Celem ćwiczeń laboratoryjnych jest nabycie praktycznych umiejętności konstrukcji takich systemów i zastosowanie ich do rzeczywistych danych pozyskiwanych w biologii i medycynie.

### Opis:

ECTS: 3

Suma godzin: 80h (kontaktowo 50h / praca własna 30h)

Wykłady: 30h

Laboratorium: 15h

Inne (omówienie sprawozdań): 5h

Praca własna studenta: przygotowanie sprawozdań oraz do egzaminu

Liczba punktów ECTS uzyskanych na zajęciach z bezpośrednim udziałem nauczyciela akademickiego: 2

Liczba punktów ECTS uzyskanych na zajęciach o charakterze praktycznym (laboratoria, projekty): 1

### Wykład

1. Statystyczne systemy uczące się, podział metod uczenia statystycznego na metody nadzorowane i nienadzorowane, przykłady zadań klasyfikacji, klasyfikacja liniowa, metody Fisherowskie klasyfikacji, dyskryminacja logistyczna.
2. Klasyfikacja bayesowska, estymacja parametrów rozkładów w klasach, naiwny klasyfikator Bayesa, powiązanie klasyfikacji bayesowskiej z metodą największej wiarygodności, optymalność klasyfikacji bayesowskiej.
3. Nieparametryczne metody estymacji rozkładów w klasach, metoda najbliższych sąsiadów, metoda okien Parzena.
4. Ocena jakości klasyfikatora. Wskaźniki jakości: precyzja, czułość, specyficzność, krzywe ROC, pole pod krzywą ROC. Schematy walidacji klasyfikatora, resubstytucja, metoda hold-out, walidacja krzyżowa, bootstrap.
5. Komitety głosujące, drzewa klasyfikacyjne, pojęcie drzewa, reguły podziału, przycinanie drzewa, algorytmy baggingu i boostingu, lasy losowe.
6. Wybór modelu, kompromis bias-variance. Generalizacja, metody poprawy generalizacji, regularyzacja.
7. Uczenie nienadzorowane, analiza składowych głównych, estymacja gęstości prawdopodobieństwa wzdłuż wybranych kierunków, analiza skupień, metody kombinatoryczne grupowania, metody hierarchiczne grupowania.
8. Wprowadzenie do statystycznej analizy danych genomowych: estymatory częstości allelicznych i genotypowych, równowaga Hardy'ego-Weinberga i jej miary, funkcje mapujące Haldane'a i Kosambięgo.
9. Problem nierównowagi sprzężeń między lokusami, narzędzia statystyczne do jej oceny
10. Metody estymacji haplotypów oraz ich częstości w populacji: algorytmy Clarka oraz EM.
11. Odziedziczalność i jej miary.
12. Regresja logistyczna jako narzędzie do modelowania interakcji pomiędzy genami.
13. Problem wielokrotnego testowania hipotez statystycznych w danych biomedycznych i metody korekty błędu pierwszego rodzaju.

### Laboratorium

#### I semestr laboratorium

1. Częstości alleliczne i genotypowe.

Wylizanie częstości allelicznych, weryfikacja hipotez o równości częstości allelicznych w dwóch populacjach (testy:  $\chi^2$ , G, oraz test Fishera). Weryfikacja hipotez o równości częstości genotypowych w dwóch populacjach (test Fishera 3x2)

2. Częstości haplotypowe.

Estymacja częstości haplotypowych w danej populacji z niesfazowanych genotypów z użyciem algorytmu Clarka oraz algorytmu Expectation-Maximization.

3. Równowaga Hardy-Weinberga oraz nierównowaga sprzężeń.

Sprawdzanie odstępstw od równowagi Hardy-Weinberga poprzez stworzenie modelu, który stanowi punkt odniesienia do rzeczywistych danych i analizę porównawczą przy pomocy testu  $\chi^2$ . Badanie nierównowagi sprzężeń na podstawie otrzymanych danych sfazowanych genotypów przy pomocy statystyki  $\chi^2$ .

4. Metody weryfikacji biologicznych modeli matematycznych (model regresji logistycznej). Powiązanie wyników analiz statystycznych z wiedzą biologiczną dotyczącą badanych genów czy polimorfizmów. Wyciągnięcie odpowiednich wniosków z wyników analizy rzeczywistych danych biologicznych w postaci interpretacji miar efektów (OR).

5. Klasyfikator Bayesa.

Implementacja algorytmu naiwnego klasyfikatora Bayesa w środowisku Matlab. Badanie wpływu metod estymacji rozkładów cech w poszczególnych klasach na wynik klasyfikacji.

6. Lasy losowe

Implementacja lasów losowych Breimana w środowisku Matlab. Badanie wpływu parametrów klasyfikatora na jakość klasyfikacji.

### Literatura:

USOS: Szczegóły przedmiotu: BioAu>SM1WS20, w cyklu: <brak>, jednostka dawcy: <brak>, grupa przedm.: <brak>

1. Literatura: J. Koronacki, J. Mielniczuk. Statystyka. WNT 2001.
2. J. Koronacki, J. Ćwik: Statystyczne systemy uczące się, WNT, 2005
3. Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York: Springer-Verlag, 2001.
4. Richard O. Duda. Peter E. Hart. David G. Stork. Pattern Classification (2nd Edition). John Wiley & Sons 2000.

#### **Efekty uczenia się:**

Wiedza: zna i rozumie

Statystyczna analiza danych i konstrukcja statystycznych systemów wnioskujących. K2A\_W01, K2A\_W17

Zrozumienie potrzeby statystycznej analizy danych i konstrukcji statystycznych systemów wnioskujących. K2A\_W01, K2A\_W17

Definicja wektora cech, selekcji cech, klasyfikacji, klasteryzacji. K2A\_W01, K2A\_W17

Zdefiniowanie pojęcia statystycznego systemu wnioskującego oraz podziału stosowanych metod na nadzorowane i nienadzorowane. K2A\_W01, K2A\_W17

Podstawowe pojęcia genetyki populacji. Omówienie metod wnioskowania statystycznego w genomice populacji. K2A\_W01, K2A\_W17

Umiejętności: potrafi: oraz Kompetencje społeczne: Jest gotów do

Potrafi wybrać najlepszy klasyfikator (model statystyczny) oraz ocenić jego jakość. K2A\_U07, K2A\_U10, K2A\_K01, K2A\_K07

Potrafi przeprowadzić klasteryzację zadanego zbioru danych. Wybór najlepszego klasyfikatora (model statystyczny) oraz ocena jego jakości. K2A\_U07, K2A\_K01

Przeprowadzenie klasteryzacji zadanego zbioru danych. K2A\_U07, K2A\_K01

#### **Metody i kryteria oceniania:**

Egzamin z zakresu całego wykładu.

Laboratorium oceniane w formie sprawozdań.

Końcowa ocena z przedmiotu jest średnią oceny końcowej z laboratorium (waga 50%) oraz egzaminu (waga 50%).

Sylabus obowiązuje od roku akademickiego 2024/25 i jego treść nie może być zmieniana w trakcie semestru.

#### **Punkty przedmiotu w cyklach:**

##### **<bez przypisanego programu>**

| Typ punktów                                | Liczba | Cykl pocz.  | Cykl kon. |
|--------------------------------------------|--------|-------------|-----------|
| Europejski System Transferu Punktów (ECTS) | 3      | 2020/2021-L |           |