

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

Cykl dydaktyczny: Semestr letni 2022/2023

Koordynator przedmiotu cyklu: Dr inż. Joanna Żyła

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:

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-Weinberga i jej miary, funkcje mapujące Haldane'a i Kosambię.
9. Problem nierównowagi sprzężeń między lokusami, narzędzia statystyczne do jej oceny
10. Metody estymacji haplotypów oraz ich częstości w populacji: algorytmy Clarka, EM oraz Gibbsa.
11. Techniki rekonstrukcji fazy w wielolokusowych oznaczeniach genotypowych – algorytm ELB.
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.

Wyliczanie częstości allelicznych, weryfikacja hipotez o równości częstości allelicznych w dwóch populacjach (testy: χ^2 , G, oraz test Fishera). Weryfikacja hipotez o równości częstości genotypowych w dwóch populacjach (test Fishera 3×2)

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 G. Badanie nierównowagi sprzężeń na podstawie otrzymanych danych sfazowanych genotypów przy pomocy statystyki χ^2 .

4. Metody weryfikacji biologicznej modeli matematycznych.

Powiązanie wyników analiz statystycznych z wiedzą biologiczną dotyczącą badanych genów. Wyciągnięcie odpowiednich wniosków z wyników analizy rzeczywistych danych biologicznych.

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.

II semestr, laboratorium

1. Przetwarzanie wstępne.

Kontrola jakości obrazów mikromacierzowych, usuwanie artefaktów obrazów. Znajdowanie sond odstających przy pomocy testu Dixona.

2. Re-adnotacja sond i normalizacja danych mikromacierzowych

Poznanie różnych metod przypisania sond mikromacierzy do odpowiednich zestawów, powiązanych z ekspresją genów danego organizmu.

Badanie problemu niejednorodności sygnału pomiędzy macierzami. Porównanie metod RMA, MAS 5.0, GeneChipRMA oraz MBEI.

3. Selekcja cech

Implementacja różnych metod redukcji wymiarowości danych statystycznych (t-test, U-test) oraz metod opartych na klasyfikatorach (RFE, MCFS).

4. Statystyczne metody selekcji cech i problem wielokrotnego testowania.
Zapoznanie się ze wskaźnikami kontroli błędu I-go rodzaju oraz metodami ich kontroli. Wprowadzenie modelu mieszaniny w zastosowaniu do oceny FDR.

5. Klasyfikacja danych.
Wprowadzenie nadzorowanych metod klasyfikacji, tj. naiwny klasyfikator Bayesa, liniowa analiza dyskryminacyjna, lasy losowe, maszyny wektorów podpierających.

6. Ocena jakości klasyfikatora.
Badanie różnych metod krosvalidacji (leave-one-out, bootstrap) oraz różnych miar jakości klasyfikatora (średni błąd klasyfikacji, krzywe ROC).

Liczba punktów ECTS: 3
Liczba punktów ECTS uzyskanych na zajęciach z bezpośrednim udziałem nauczyciela akademickiego: 2
Liczba punktów ECTS uzyskanych na zajęciach o charakterze praktycznym (laboratoria, projekty): 1

Wykłady: kontaktowych 30 / pracy studenta 15
Laboratorium: kontaktowych 15 / pracy studenta 15
Inne: kontaktowych 30 / pracy studenta 30
Suma wszystkich godzin: 135

Literatura:

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_K01
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.

Dane dotyczące przedmiotu cyklu:

Domyślny typ protokołu dla przedmiotu cyklu:

EGZ

Szczegóły zajęć i grup

wykład (30 godzin)

Dane grup zajęciowych

Grupa numer 1

Prowadzący grupy:

Mgr inż. Joanna Tobiasz
Dr inż. Joanna Żyła
Dr hab. inż. Sebastian Student
Prof. dr hab. inż. Krzysztof Fajurewicz

laboratorium (15 godzin)

Dane grup zajęciowych

Grupa numer 1

Prowadzący grupy:

Dr inż. Joanna Żyła

Przynależność do grup przedmiotów w cyklach:

Opis grupy przedmiotów	Cykl pocz.	Cykl kon.
Biotechnologia S2 semestr 1 specj. BIO (BioAu>SM1-BIO-20)	2020/2021-L	

Punkty przedmiotu w cyklach:

<bez przypisanego programu>

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