(pieczęć wydziału)

KARTA PRZEDMIOTU

1. Nazwa przedmiotu: WNIOSKOWANIE STATYSTYCZNE	**2. Kod przedmiotu:**
3. Karta przedmiotu ważna od roku akademickiego: 2012/2013	
4. Forma kształcenia: studia drugiego stopnia	
5. Forma studiów: studia stacjonarne	
6. Kierunek studiów: BIOTECHNOLOGIA; (WYDZIAŁ AEII)	
7. Profil studiów: ogólnoakademicki	
8. Specjalność: BIOINFORMATYKA	
9. Semestr: 1, 2	
10. Jednostka prowadząca przedmiot: Instytut Automatyki, RAu1	
11. Prowadzący przedmiot: prof. dr hab. inż. Joanna Polańska, dr hab. inż. Krzysztof Fujarewicz	
12. Przynależność do grupy przedmiotów: przedmioty wspólne	
13. Status przedmiotu: obowiązkowy	
14. Język prowadzenia zajęć: polski	
15. Przedmioty wprowadzające oraz wymagania wstępne: Przedmioty podstawowe dla pierwszego stopnia studiów, w tym analiza matematyczna, statystyka. Zakłada się, że przed rozpoczęciem nauki niniejszego przedmiotu student posiada przygotowanie w zakresie uczenia maszynowego i metod sztucznej inteligencji.	
16. Cel przedmiotu: 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.	

17. Efekty kształcenia:

Nr	Opis efektu kształcenia	Metoda sprawdzenia efektu kształcenia	Forma prowadzenia zajęć	Odniesienie do efektów dla kierunku studiów
1	Zna pojęcie statystycznego systemu wnioskującego oraz rozumie podział stosowanych metod na nadzorowane i nienadzorowane.	EP	WT, WM	K_W01
2	Zna podstawowe pojęcia: wektor cech, selekcja cech, klasyfikacja, klasteryzacja.	EP	WT, WM	K_W01
3	Zna podstawowe pojęcia genetyki populacji	EP	WT, WM	K_W17
4	Zna metody wnioskowania statystycznego w genomice populacji	EP	WT, WM	K_W17
5	Potrafi zbudować klasyfikatory różnych typów: Fishera, Bayesa, komitet głosujący, dla zadanego zbioru danych.	EP, CL, PS	L	K_U07, K_U10
6	Potrafi wybrać najlepszy klasyfikator (model statystyczny) oraz ocenić jego jakość.	EP, CL, PS	L	K_U07, K_U10
7	Potrafi przeprowadzić klasteryzację zadanego zbioru danych.	EP, CL, PS	L	K_U07, K_U10
8	Potrafi zastosować metody wnioskowania statystycznego dla rzeczywistych danych biologicznych	EP, CL, PS	L	K_U07, K_U10
9	Rozumie potrzebę statystycznej analizy danych i konstrukcji statystycznych systemów wnioskujących.	EP	WT, WM	K_K07, K_K01

18. Formy zajęć dydaktycznych i ich wymiar (liczba godzin)

W. : 30 L.: 30

19. Treści kształcenia:

Wykład

1. Statystyczne systemy uczące się, podział metod uczenia statystycznego na metody nadzorowanie 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 nienadzorowanie, 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 Kosambiego.
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 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 G. Badanie nierównowagi sprzężeń na podstawie otrzymanych danych sfazowanych genotypów przy pomocy statystyki χ2.

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 kroswalidacji (leave-one-out, bootstrap) oraz różnych miar jakości klasyfikatora (średni błąd klasyfikacji, krzywe ROC).

20. Egzamin: tak; pisemny.

21. Literatura podstawowa:

1. Literatura: J. Koronacki, J. Mielniczuk. Statystyka. WNT 2001.
2. J.Koronacki, J. Ćwik: Statystyczne systemy uczące się, WNT, 2005

22. Literatura uzupełniająca:

1. Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York: Springer-Verlag, 2001.
2. Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification (2nd Edition), John Wiley & Sons 2000.

23. Nakład pracy studenta potrzebny do osiągnięcia efektów kształcenia

Lp.	Forma zajęć	Liczba godzin kontaktowych / pracy studenta
1	Wykład	30/15
2	Ćwiczenia	0/0
3	Laboratorium	30/15
4	Projekt	0/0
5	Seminarium	0/0
6	Inne	30/30
	Suma godzin	90/60

24. Suma wszystkich godzin: 150

25. Liczba punktów ECTS: 5

26. Liczba punktów ECTS uzyskanych na zajęciach z bezpośrednim udziałem nauczyciela akademickiego: 3

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

26. Uwagi:

Zatwierdzono:

…… ……

(data i podpis prowadzącego) (*data i podpis dyrektora instytutu/kierownika katedry/ Dyrektora Kolegium Języków Obcych/kierownika lub dyrektora jednostki międzywydziałowej)*