

WYDZIAŁ MATEMATYKI

KARTA PRZEDMIOTU

Nazwa w języku polskim: Teoretyczne podstawy analizy danych wielowymiarowych

Nazwa w języku angielskim: Theoretical foundations of large scale data analysis

Kierunek studiów (jeśli dotyczy): MATEMATYKA

Specjalność (jeśli dotyczy): Statystyka matematyczna

Stopień studiów i forma: 2 stopień, stacjonarna /niestacjonarna*

Rodzaj przedmiotu: obowiązkowy / wybieralny /ogólnouczelniany*

Kod przedmiotu MAP2068

Grupa kursów TAK / NIE

	Wykład	Ćwiczenia	Laboratorium	Projekt	Seminarium
Liczba godzin zajęć zorganizowanych w Uczelni (ZZU)	30		30		
Liczba godzin całkowitego nakładu pracy studenta (CNPS)	150				
Forma zaliczenia	Egzamin/ zaliczenie na ocenę				
Dla grupy kursów zaznaczyć kurs końcowy	X				
Liczba punktów ECTS	5				
w tym liczba punktów odpowiadająca zajęciom o charakterze praktycznym (P)	3				
W tym liczba punktów ECTS odpowiadająca zajęciom wymagającym bezpośredniego kontaktu (BK)	3				

WYMAGANIA WSTĘPNE W ZAKRESIE WIEDZY, UMIEJĘTNOŚCI I INNYCH KOMPETENCJI

Zna podstawowe pojęcia statystyki matematycznej, w tym pojęcie statystyki dostatecznej, zupełnej. Zna podstawowe graniczne własności estymatorów. Zna podstawowe metody wnioskowania statystycznego: estymacji (punktowej i przedziałowej) i testowania hipotez . Zna podstawowe kryteria optymalności estymatorów i testów statystycznych. Zna podstawowe modele liniowe (regresja wieloraka, analiza wariancji).

CELE PRZEDMIOTU

C1 Zrozumienie podstawowych problemów statystycznych związanych z analizą dużych zbiorów danych : problemy wielokrotnego testowania i ich związek z przeszacowaniem liczby istotnych predyktorów w modelach liniowych.

C2 Zrozumienie ograniczeń związanych z analizą dużych zbiorów danych – granica detektowalności sygnałów, związek między rzadkością modelu a siłą detektowalnych sygnałów.

C3 Opanowanie podstawowych współczesnych metod analizy danych wielowymiarowych : podstawowe procedury wielokrotnego testowania kontrolujące FWER i FDR, podstawowe kryteria wyboru modelu, LASSO, estymacja macierzy precyzji w gaussowskich modelach graficznych

PRZEDMIOTOWE EFEKTY KSZTAŁCENIA

Z zakresu wiedzy:

PEK_W01 zna podstawowe twierdzenia i metody dotyczące metod wielokrotnego testowania (metoda Bonferroniego, Simesa, Holma, Benjaminiego Hochberga)

PEK_W02 zna podstawowe pojęcia dotyczące kontroli błędu pierwszego rodzaju w procedurach wielokrotnego testowania: całkowity błąd pierwszego rodzaju (FWER), frakcja fałszywych odkryć (FDR)

PEK_W03 zna podstawowe twierdzenia asymptotyczne dotyczące detektowalności sygnałów w wielkoskalowym testowaniu

PEK_W04 zna podstawowe twierdzenia motywujące stosowanie estymatorów ściągających (estymator Jamesa-Steina) i rozumie ich związek z empirycznymi metodami Bayesowskimi

PEK_W05 zna podstawowe kryteria wyboru zmiennych w modelach liniowych i rozumie ich własności statystyczne

PEK_W06 zna podstawowe metody analizy modeli graficznych

Z zakresu umiejętności:

PEK_U01 potrafi dokonać wyboru właściwej dla danego problemu metody wielokrotnego testowania i zastosować ją w praktyce (w razie potrzeby zaimplementować w R).

PEK_U02 potrafi dokonać wyboru właściwego dla danego problemu kryterium wyboru modelu i zastosować je w praktyce (w razie potrzeby zaimplementować w R).

PEK_U03 potrafi wykorzystywać praktycznie poznane metody analizy modeli graficznych przy wspomaganii profesjonalnych komputerowych pakietów statystycznych

Z zakresu kompetencji społecznych:

PEK_K01 potrafi wyszukiwać i korzystać z literatury zalecanej do kursu oraz samodzielnie zdobywać wiedzę

PEK_K02 potrafi twórczo współpracować w grupie studenckiej, budować pozytywne więzi emocjonalne z jej członkami

TREŚCI PROGRAMOWE

Forma zajęć - wykłady		Liczba godzin
Wy1	Procedury wielokrotnego testowania i wyboru modelu – motywacja z analizy danych genetycznych.	2
Wy2	Wielokrotne testowanie, globalna hipoteza zerowa – szukanie igły w stogu siana (pojedynczy sygnał), optymalny test ilorazu wiarygodności, korekta Bonferroniego, granice detekcji.	2
Wy3	Testowanie globalnej hipotezy zerowej – wiele umiarkowanych sygnałów, test chi-kwadrat, analiza wariancji, kombinowany test Fishera, test Simesa, podniesiony krytycyzm Tukeya.	2
Wy4	Wielokrotne testowanie w rzadkich mieszaninach, globalna hipoteza zerowa – granice detekcji, optymalność testów.	2
Wy5	Wielokrotne testowanie – słaba, mocna kontrola całkowitego błędu pierwszego rodzaju, procedura Holma, zasada domknięcia.	2
Wy6	Procedura Hochberga, generyczne procedury zstępujące, frakcja fałszywych odkryć (FDR) i procedury ją kontrolujące.	2
Wy7	Asymptotyczne własności procedury Benjaminiego-Hochberga (BH), dodatnia regresyjna zależność, kontrola FDR z punktu widzenia procesów empirycznych.	2

Wy8	Martynałowy dowód kontroli FDR przez BH, Bayesowskie FDR, BH jako empiryczna metoda Bayesowska.	2
Wy9	Estymator Jamesa-Steina.	2
Wy10	Nieobciążony estymator ryzyka Steina (SURE).	2
Wy11	Kryteria wyboru modelu liniowego, błąd predykcji i błąd w próbie treningowej.	2
Wy12	Kryteria związane z kontrolą FDR.	2
Wy13	LASSO i SLOPE	2
Wy14	Estymacja macierzy kowariancji. Analiza składowych głównych.	2
Wy15	Modele graficzne.	2
	Suma godzin	30

Forma zajęć - laboratorium		Liczba godzin
La1	Symulacja danych genetycznych.	2
La2	Wielokrotne testowanie – porównania różnych testów dla globalnej hipotezy zerowej, empiryczna weryfikacja granicy detektowalności.	6
La3	Wielokrotne testowanie – porównania różnych procedur z punktu widzenia kontroli liczby fałszywych odkryć i mocy.	4
La 4	Procedura Benjaminiego-Hochberga – empiryczna weryfikacja własności teoretycznych	4
La 5	Estymacja wektora wartości oczekiwanych w wielowymiarowym rozkładzie normalnym – estymator Jamesa-Steina. Nieobciążony estymator ryzyka Steina.	4
La 6	Kryteria wyboru modelu w zastosowaniu do rzadkiej regresji.	6
La 7	Estymacja macierzy kowariancji i modeli graficznych	4
	Suma godzin	30

STOSOWANE NARZĘDZIA DYDAKTYCZNE

1. Wykład informacyjny, problemowy, metoda tradycyjna i częściowo prezentacja multimedialna
2. Laboratorium
3. Konsultacje
4. Praca własna studenta – przygotowanie do laboratorium

OCENA OSIĄGNIĘCIA EFEKTÓW KSZTAŁCENIA

Oceny (F – formująca (w trakcie semestru), P – podsumowująca (na koniec semestru))	Numer efektu kształcenia	Sposób oceny osiągnięcia efektu kształcenia
F1	PEK_U01- PEK_U03	Odpowiedzi ustne, raporty

F2	PEK_W01- PEK_W06	Kolokwium zaliczające
F3		
P=0,7*F1+0,3*F2		

LITERATURA PODSTAWOWA I UZUPEŁNIAJĄCA

LITERATURA PODSTAWOWA:

- [1] Efron, B. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, IMS Monographs, 2012
 [2] Johnstone, Gaussian estimation: Sequence and wavelet models, draft version, 2013

LITERATURA UZUPEŁNIAJĄCA:

- [1] Dembo, A. i Zeitouni, O. Large deviations techniques and applications, second edition, Springer, Application of Mathematics, vol. 38, 1998.
 [2] Shorack, G. i Wellner, J. Empirical Processes With Applications to Statistics, Classics in Applied Mathematics, 1986
 [3] Adler, R. i Taylor, J. Random Fields and Geometry, Springer Monographs in Mathematics Springer, New York, 2007.

OPIEKUN PRZEDMIOTU (IMIE, NAZWISKO, ADRES E-MAIL)

Dr hab. Malgorzata Bogdan (malgorzata.bogdan@pwr.edu.pl)

MACIERZ POWIĄZANIA EFEKTÓW KSZTAŁCENIA DLA PRZEDMIOTU
Teoretyczne podstawy analizy danych wielowymiarowych
Z EFEKTAMI KSZTAŁCENIA NA KIERUNKU MATEMATYKA
I SPECJALNOŚCI Statystyka matematyczna

Przedmiotowy efekt kształcenia	Odniesienie przedmiotowego efektu do efektów kształcenia zdefiniowanych dla kierunku studiów i specjalności (o ile dotyczy)	Cele przedmiotu**	Treści programowe**	Numer narzędzia dydaktycznego**
PEK_W01 (knowledge)	K2MAT_W03, K2MAT_W08- K2MAT_W11, K2MAT_W13, K2MAT_W15S3STM	C1, C3	Wy1 – Wy8,	1,3,4
PEK_W02	K2MAT_W03, K2MAT_W08- K2MAT_W11, K2MAT_W13, K2MAT_W15S3STM	C1,C3	Wy5 – Wy8	1,3,4
PEK_W03	K2MAT_W03, K2MAT_W08- K2MAT_W11, K2MAT_W13, K2MAT_W15S3STM	C1-C3	Wy2, Wy4	1,3,4
PEK_W04	K2MAT_W03, K2MAT_W08- K2MAT_W11, K2MAT_W13, K2MAT_W15S3STM	C3	Wy9-Wy10	1,3,4
PEK_W05	K2MAT_W03, K2MAT_W08- K2MAT_W11, K2MAT_W13, K2MAT_W15S3STM	C1,C3	Wy11 – Wy13	1,3,4
PEK_W06	K2MAT_W03, K2MAT_W08- K2MAT_W11, K2MAT_W13, K2MAT_W15S3STM	C1,C3	Wy14-Wy15	1,3,4
PEK_U01) (skills)	K2MAT_U01-K2MAT_U06, K2MAT_U08, K2MAT_U13S3STM.	C1-C3	La2 – La4	2-4
PEK_U02	K2MAT_U01-K2MAT_U06, K2MAT_U08, K2MAT_U13S3STM.	C1,C3	La5-La6	2-4
PEK_U03	K2MAT_U01-K2MAT_U06, K2MAT_U08, K2MAT_U13S3STM.	C1, C3	La7	2-4
PEK_K01 (competences)	K2MAT_K01-K2MAT_K07	C1 – C3	Wy1 – Wy15, La1 – La7	1, 2, 3, 4
PEK_K02	K2MAT_K01-K2MAT_K07	C1 – C3	Wy1 – Wy15, La1 – La7	1, 2, 3, 4

** - z tabeli powyżej