Instytut Matematyczny
Uniwersytet Wrocławski

Komisja Egzaminów Dyplomowych

Zakres egzaminu magisterskiego dla specjalności

Analiza danych

(obowiązuje studentów studiujących wg programu z 1 X 2019 bądź późniejszego)

1 Wielowymiarowa Analiza Statystyczna

Wymagane pojęcia, fakty:

Dystrybuanta oraz dystrybuanty brzegowe i warunkowe wektora losowego. Gęstość rozkładu wektora losowego. Macierz losowa i jej rozkład. Wektor średnich wektora losowego. Macierz kowariancji wektora losowego. Funkcja charakterystyczna wektora losowego. Rozkład funkcji wektora losowego. Macierze blokowe. Różniczkowanie wektorów i ekstrema warunkowe. Wielowymiarowy rozkład normalny. Rozkład Wisharta. Rozkład Hotellinga. Rozkład wektora średnich próbkowych. Rozkład macierzy kowariancji próbkowej i jej podmacierzy. Test dla (wektora) średniej. Test dla kombinacji liniowej składowych (wektora) średniej. Test dla dwóch (wektorów) średnich.

Przykładowe zadania:

1. Wektor losowy $\boldsymbol{X} = (X_1, X_2)$ ma gęstość $f_{\boldsymbol{X}}(\boldsymbol{x}) = e^{-(x_1+x_2)}$, $x_1, x_2 > 0$. Wyznacz gęstość wektora losowego $\boldsymbol{U} = (U_1, U_2)$, gdzie $U_1 = X_1 + X_2$, $U_2 = X_1 - X_2$.

2. Niech $f(x, y) = 2 \cdot \mathbb{1}(0 < x < y, 0 < y < 1)$ będzie gęstością wektora losowego (X, Y). Wyznacz $E(Y|X)$, $E(X|Y)$ oraz korelację między tymi zmiennymi losowymi.

3. Niech $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ będzie próbą z rozkładu $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, gdzie $|\boldsymbol{\Sigma}| > 0$, $N > n$, a macierz kowariancji nie jest znana. Rozważamy problem testowania $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ przeciwko $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ na poziomie istotności α. Podaj postać statystyki testowej w powyższym zagadnieniu. Jaki jest jej rozkład przy H_0? Odpowiedź uzasadnij.

2 Wnioskowanie Statystyczne

Wymagane pojęcia, fakty:

Testy jednostajnie najmocniejsze. Testy nieobciążone. Testy w parametrycznym modelu normalnym. Testy oparte na ilorazie wiarogodności. Estymacja bayesowska. Estymacja minimaksowa. Estymacja nieobciążona o minimalnej wariancji.

Przykładowe zadania:

1. Niech $X_1, \ldots, X_n$ będzie próbą z rozkładu beta z parametrami α i β. Testujemy $H_0 : \alpha = 3, \beta = 1$ przeciwko $H_1 : \alpha = 5, \beta = 1$. Wyznacz test jednostajnie najmocniejszy na poziomie istotności $\alpha = 0.05$. Jaka jest moc testu przy alternatywie?

2. Niech $X_1, \ldots, X_n$ będzie próbą z rozkładu o gęstości $f(x, \theta) = \theta(1+x)^{-(\theta+1)} \mathbb{1}_{(0,\infty)}(x), \theta > 0$. Testujemy $H_0 : \theta = 1$ przeciwko $H_1 : \theta > 1$. Wyznacz test jednostajnie najmocniejszy na poziomie istotności $\alpha = 0.05$.

3. Niech $X_1, \ldots, X_n$ będzie próbą z rozkładu jednostajnego na odcinku $(0, \theta)$, $\theta \in (0, \infty)$. Wyznacz test ilorazu wiarogodności, na poziomie istotności α, w problemie testowania $H_0 : \theta = \theta_0$ przeciwko $H_1 : \theta \neq \theta_0$.

4. Niech $X_1, \ldots, X_n$ będzie próbą z rozkładu zero-jedynkowego $b(1, p)$, $p \in (0, 1)$. Wyznacz estymator bayesowski parametru p ze względu na rozkład *a priori* jednostajny $U(0, 1)$, gdy funkcja straty jest postaci $L(p, a) = (p - a)^2/[p(1 - p)]$. Oblicz jego ryzyko i ryzyko bayesowskie.

3 Analiza dużych zbiorów danych (*Theoretical Foundations of the Analysis of Large Data Bases, Statistical Learning, Metody Klasyfikacji i Redukcji Wymiaru*)

Wymagane pojęcia, fakty:

Theoretical Foundations of the Analysis of Large Data Sets.

- Testowanie hipotezy globalnej: procedura Bonfferoniego, test Fishera, test chi-kwadrat, test Simesa, testy oparte na dystrybuancie empirycznej, test Tukeya.
- Modele probabilistyczne dla hipotezy globalnej: igła w stogu siana, równomiernie rozłożone sygnały, rzadka mieszanina. Granice detektowalności w tych modelach.
- Wielokrotne testowanie: procedura Holma, zasada domknięcia, procedura Hochberga, procedura Benjaminiego-Hochberga, miary błędu pierwszego rodzaju w wielokrotnym testowaniu (Family Wise Error Rate, False Discovery Rate).
- Elementy statystyki Bayesowskiej : rozkład a priori, rozkład a posteriori, empiryczny estymator Bayesowski.
- Estymator Jamesa-Steina, nieobciążony estymator ryzyka Steina.

Statistical Learning

- Estymacja błędu predykcji w oparciu o sumę kwadratów resztowych.
- Czynnik Bayesa (Bayes factor), kryteria informacyjne – AIC, BIC, RIC, mBIC, wyznaczenie oczekiwanej liczby fałszywych i prawdziwych odkryć w przypadku, gdy $\mathbf{X}'\mathbf{X} = \mathbf{I}$.
- Metody regularyzacyjne (regresja grzbietowa, LASSO, SLOPE) – w przypadku regresji grzbietowej i LASSO - wyznaczenie oczekiwanej liczby fałszywych i prawdziwych odkryć oraz błędu średniokwadratowego w przypadku gdy$\mathbf{X}'\mathbf{X} = \mathbf{I}$.
- Grafowy model gaussowski – funkcja wiarogodności, estymacja za pomocą metody największej wiarogodności, LASSO i SLOPE.

Metody klasyfikacji i redukcji wymiaru

- Analiza składników głównych (PCA), rozkład wartości osobliwych (SVD), analiza składowych niezależnych (ICA), nieujemna faktoryzacja macierzy (NMF).
- Klasyczne klasyfikatory: najbliższych sąsiadów, naiwny klasyfikator Bayesa, mnożniki Lagrange'a - problemy pierwotne i dualne, maszyna wektorów nośnych (SVM), liniowa analiza dyskryminacyjna (LDA).
- Algorytmy grupowania: algorytm k-means, Gaussian Mixture Model (GMM), algorytm Expectation-Maximization (EM).
- Ukryte modele Markowa (HMM) z dyskretnymi obserwacjami: procedury forward i backward, algorytm Viterbiego.
- Metody Monte Carlo dla łańcuchów Markowa (Markov Chain Monte Carlo, MCMC) - próbnik Gibbsa, algorytm Metropolisa.

Przykładowe zadania

1. Znajdź maksimum funkcji $f(h,s) = 100h^{2/3}s^{1/3}$ przy ograniczeniu: $20h + 2000s = 20000$.

2. Niech $\mathbf{Z} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$. Znajdź (przez $\mathbf{x} \in \mathbb{R}^2$ rozumiemy wektor pionowy, tj. $\mathbf{x}^T = (x_1, x_2)$)

a) $\max\limits_{\mathbf{x}\in\mathbb{R}^2, \mathbf{x}\neq\mathbf{0}} \dfrac{\mathbf{x}^T\mathbf{Z}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$,

b) $\arg\max\limits_{\mathbf{x}\in\mathbb{R}^2, \mathbf{x}\neq\mathbf{0}} \dfrac{\mathbf{x}^T\mathbf{Z}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$.

3. Mamy dane 4 punkty z $\mathbb{R}^2$: $(-1,1), (1,-1), (2,2), (-2,-2)$. Wylicz składowe główne (algorytm PCA). Wskaż, którą składową należy wybrać, jeśli chcemy zredukować punkty metodą PCA do wymiaru 1.

4. Rozważmy nastepujący ukryty model Markowa:
(Ukryte) stany: $\mathcal{S} = \{0, 1\}$. Macierz prawdopodobieństw przejść między stanami jest nastepująca:

$$\mathbf{A} = \begin{bmatrix} 0.4 & 0.6 \\ 0.8 & 0.2 \end{bmatrix}.$$

Zbiór obserwacji: $\mathcal{V} = \{a, b\}$. Prawdopodobieństwa obserwacji będąc w każdym ze stanów $s \in \mathcal{S}$:

$$P(O_t = a|X_t = 0) = 0.9 \qquad P(O_t = b|X_t = 0) = 0.1$$

$$P(O_t = a|X_t = 1) = 0.5 \qquad P(O_t = b|X_t = 1) = 0.5$$

Znamy również rozkład początkowy łańcucha: $P(X_1 = 0) = 0.7, P(X_1 = 1) = 0.3$.

Używając procedury *forward* policz prawdopodobieństwo zaobserwowania B w chwili 1 oraz A w chwili 2 (tj. $O_1 = B, O_2 = A$).

4 Rachunek prawdopodobieństwa

Wymagane pojęcia, fakty:

Prawdopodobieństwo klasyczne i geometryczne; rozwiązywanie zadań. Rozkłady zmiennych i wektorów losowych; zmienne losowe ciągłe i dyskretne. Pojęcie ciągu niezależnych zmiennych losowych.Wartość oczekiwana z.l., momenty z.l., wariancja.Warunkowe prawdopodobieństwo i wartość oczekiwana. Przypadek gdy warunek ma dodatnie prawdopodobieństwo. Nierówność Czebyszewa. Zbieżność wg prawdopodobieństwa i wg rozkładu. Słabe prawo wielkich liczb. Twierdzenie graniczne Poissona i centralne twierdzenie graniczne.

Wymagane umiejętności:

Proste zadania kombinatoryczne. Dystrybuanta, gęstość, funkcja prawdopodobieństwa. Obliczanie rozkładów przekształconych zmiennych lub wektorów losowych. Na przykład gdy jest znana gęstość zmiennej losowej X to jaka jest gęstość zmiennej losowej h(X). Podobnie dla wektorów. Wzory na rozkład normalny (zmiennej losowej i wektora losowego), rozkład dwumianowy i rozkład Poissona. Niezależność dwóch i więcej zmiennych losowych. Pojęcie ciągu niezależnych zmiennych losowych. Niezależność ciągu zdarzeń. Umiejętność liczenia wartości oczekiwanej zmiennej losowe dla przypadku ciągłego i dyskretnego. Wzór na wariancję sumy (przypadek zmiennych zależnych i niezależnych). Warianty twierdzenia Czebyszewa. Wykorzystanie centralnego twierdzenia granicznego do zadań praktycznych. Zadania na aproksymacje rozkładu dwumianowego i innych rozkładem Poissona. Zastosowanie transformat do obliczania rozkładów sum niezależnych zmiennych losowych i twierdzeń granicznych. Wykorzystanie twierdzeń granicznych (centralnego twierdzenia granicznego i twierdzenia Poissona) do zadań aplikacyjnych.

Przykładowe zadania

1. Niech (X, Y) będzie wektorem losowym z wektorem średniej $(2, 1)$ i macierzą kowariancji

$$\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

oraz $Z = X + 2Y$. Ciąg $Z_1, Z_2, \ldots, Z_{100}$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie jak Z. Oszacować

$$\mathbb{P}(Z_1 + \ldots + Z_{100} \in (400 - 10\sqrt{10}, 400 + 10\sqrt{10})).$$

2. a) Wykonujemy co sekundę doświadczenie Bernoulliego aż do chwili otrzymania pierwszego sukcesu. Niech X oznacza liczbę wykonanych doświadczeń, Y mierzony w sekundach czas oczekiwania na pierwszy sukces. Prawdopodobieństwo sukcesu wynosi p. Wyznaczyć rozkłady zmiennych losowych X i Y.
b) Przypuśćmy, że zamiast doświadczeń wykonywanych co sekundę, sa one dokonywane co $1/n$ sekundy. Niech prawdopodobieństwo sukcesu wynosi λ/n. Przez Y_n oznaczamy czas oczekiwania na pierwszy sukces mierzony w sekundach. Wyznaczyć ogon dystrybuanty Y_n oraz zbadać jego zachowanie gdy $n \to \infty$.

3. Wektor (X, Y) ma gęstość

$$f(x, y) = \begin{cases} Cxy, & x, y > 0, \quad x^2 + y^2 \leqslant 1, \\ 0, & \text{w przeciwnym razie.} \end{cases}$$

Wyznaczyć: a) stałą C, b) $P(X = Y)$, c) $P(X > Y)$, d) dystrybuantę X.

4. Na poczcie pojawia się 400 klientów dziennie, każdy z nich dokonuje wpłaty/wypłaty X_i, gdzie X_i są niezależnymi zmiennymi losowymi o jednakowym rozkładzie, zerowej średniej, i wariancji równej 100. Ile gotówki należy mieć w kasie rano, by z prawdopodobieństwem 0,99 na koniec dnia nie zabrakło gotówki.

Zakładamy, że w ciągu dnia naczelnik poczty uzupełnia ze swojej kieszeni, ale wieczorem odzyskuje, jeśli na koniec dnia jest wynik dodatni.

5. Wykonanie pewnej pracy zajmuje czas losowy z średnią 5 i wariancją 4. Zakłada się, że czasy pracy kolejnych prac są niezależne o jednakowym rozkładzie. Oszacować prawdopodobieństwo, że przynajmniej 50 prac zostanie wykonanych w przeciągu 240 godzin.

6. Niech dla każdego n zmienne losowe $\xi_{n,j}, j = 1, \ldots, n$ są niezależne o jednakowym rozkładzie jednostajnym na $[-n, n]$. Zdefiniujmy indykatorowe zmienne trafień w odcinek $(a, b) \subset [-n, n]$, gdzie $a < b$

$$I_{nj} = \mathbf{1}(\xi_{n,j} \in (a, b)), \qquad j = 1, \ldots, n$$

oraz liczbę trafień

$$S_n = \sum_{j=1}^{n} I_{nj}.$$

Oblicz średnią i wariancję S_n.
Udowodnić z powołaniem się na twierdzenia, że S_n zbiega do rozkładu Poissona z parametrem λ. Ile wynosi λ?