

Statystyka II. Egzamin komputerowy 09.09.09.

1. Za pomocą *glm* na danych `Pima.tr` wyestymuj parametry regresji logistycznej, a następnie dla danych `Pima.te` wyestymuj $\log(P(y=="Yes"|x)/P(y=="No"|x))$, gdzie $y=Pima.te[, "type"]$.
2. Wczytaj z archiwum `MacierzePot11_3kol.zip` na stronie <http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/DaneEgzaminacyjne/> potencjały kontaktowe (PK), czyli 12 symetrycznych macierzy 20×20 opisujących oddziaływania 210 par aminokwasów w białkach. PK są zapisane w plikach trójkolumnowych tak, że w pierwszej kolumnie znajduje się numer wiersza, w drugiej – numer kolumny, a w trzeciej - wartość odpowiedniego elementu macierzy PK o tych współrzędnych. Napisz ogólną funkcję `three2Vec`, która przekształca plik trójkolumnowy na wektor złożony z elementów macierzy PK należących do jej górnej części (ewentualnie dolnej) oraz przekątnej. Ze zbioru wektorów x_1, \dots, x_{12} otrzymanych z 12tu PK, zbuduj macierz danych $X=[x_1, \dots, x_{12}]$. Sprawdź, czy w zbiorze X (210 obserwacji 12- wymiarowych) są obserwacje odstające.
3. Wykonaj klasteryzację danych X z poprzedniego zadania metodą k-średnich. Ile jest klastrów (odpowiedź uzasadnij wykresem separowalności)? Narysuj oddziaływania między aminokwasami na płaszczyźnie dwóch pierwszych składowych głównych i policz udział wariancji tych składowych w całkowitej wariancji danych.
4. Dla modelu regresji $y = \beta^T X + \varepsilon$ dany jest wąski rozkład $X=QR$ oraz X jest macierzą pełnego rzędu $n \times p$. Napisz funkcję, która bezpośrednio (bez użycia `lm`, wykorzystując tylko y i Q) oblicza estymator nieobciążony wariancji y kosztem $O(n \cdot p)$ działań arytmetycznych.
5. Niech czas życia żarówki (C) ma rozkład gamma z parametrami: skala=1, kształt=3. Wylosuj 1000 niezależnych obserwacji z tego rozkładu i wyestymuj prawdopodobieństwo, że C jest nie większy niż $d=1.5 \cdot z_{0.75}$, gdzie $z_{0.75}$ jest kwantylem próbkowym rzędu 0.75 tego rozkładu (wskazówka: funkcja `quantile` oblicza kwantyle próbkowe). Ponadto policz d numerycznie (bez losowania).
6. Zbuduj model regresji liniowej wielu zmiennych dla danych `bodyfatJohnson.txt` (bez 1szej kolumny) ze strony <http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/>. Cechą przewidywaną ma być "Bodyfat".