Egzamin ze Statystycznej Analizy Danych

23.06.2014

Czas trwania: 2h. Wszystkie odpowiedzi piszemy na kartkach podpisanych imieniem, nazwiskiem, kierunkiem studiów i numerem indeksu. Można pisać na odwrocie kartki z zadaniami. Uzasadnienia odpowiedzi powinny być zwięzłe, ale powinny zawierać istotne informacje, np. nazwę użytego testu, czynione założenia, badane parametry.

1. (4pkt) Szpital ma pacjentów z pomiarami ilościowymi (np. poziom cukru we krwi), którzy biorą udział w teście nowego leku, który ten poziom ma obniżać. Jeśli mamy po 20 pacjentów w dwóch grupach (placebo i lek) i po jednym pomiarze dla każdego pacjenta, to jak przetestować statystycznie, czy użycie leku ma wpływ na zmierzony poziom? Opisz procedurę badania takiej hipotezy statystycznej i odpowiedź uzasadnij (2pkt). Co zmieniłoby się, gdyby zamiast poziału na grupy, dla każdego pacjenta znana byłaby dawka leku i poziom po zażyciu? (1pkt) A jak postąpić w przypadku, gdy zamiast dwóch grup pacjentów mamy ich więcej (np. lek A, lek B, lek C i kontrola)? (1pkt).

2. (2 pkt) Lokalny browar chciałby przetestować nowe szczepy drożdży do produkcji mocnego piwa, aby sprawdzić które mutacje mogą być odpowiedzialne za tolerancję na wysokie stężenie alkoholu. Piwowarzy mają sto szczepów tolerujących wysokie procenty i sto szczepów, które tych warunków nie tolerują. Dla każdego z nich zbadano 1000 mutacji (załóżmy że mutacje występują niezależnie od siebie i są binarne: dana mutacja w danym szczepie jest lub jej nie ma). Jak wybrać mutacje istotnie powiązane z tolerancją alkoholu? Jak określić prawdopodobny odsetek błędów I rodzaju dla zadanego poziomu istotności? Czy założenie o niezależności jest istotne i jeśli tak to dlaczego?

3. (2pkt) Podaj przykład danych gdzie sensowne byłoby zastosowanie analizy wariancji (anova). Na jakie pytania (hipotezy) można w ten sposób szukać odpowiedzi?

4. (2pkt) Ostatnio analizowałem pewien zbiór danych przy użyciu regresji liniowej. Testowałem modele wyjaśniające zmienną Y przy użyciu zmiennych X_1 lub X_2. Zapamiętałem, że w modelu wg wzoru $Y \sim \alpha_i X_i + \beta$ współczynnik α_1 był dwa razy większy od α_2. Co z tego wynika dla korelacji zmiennych X_1 i X_2 z Y? Jakie parametry mają większe znaczenie i jaka jest ich interpretacja?

5. (3 pkt) Załóżmy, że mamy do czynienia z zagadnieniem klasyfikacji. Kiedy lepiej jest zastosować metodę k-najbliższych-sąsiadów (kNN), a kiedy lepiej sprawdzą się metody dyskryminacyjne (typu LDA i QDA)? Jak można sprawdzić, czy zbiór danych będzie lepiej nadawał się do jednej lub drugiej metody (np. gdy zbiór danych jest duży i nie chcemy znajdować optymalnego modelu obiema metodami)?

6. (3 pkt) Załóżmy, że w zbiorze danych do klasyfikacji znajdują się dwie obserwacje o takich samych wszystkich cechach a różnych decyzjach (np. dwaj bliźniacy jednojajowi, zdrowy i chory, w zagadnieniu klasyfikacji chorych na podstawie genotypu). Jakie będzie to miało konsekwencje dla miar jakości znalezionego modelu (czułość, specyficzność, AUC)? Czy można znaleźć klasyfikator o 100 procentowej skuteczności? Czy odpowiedź może się zmienić się jeśli użyjemy walidacji krzyżowej?

7. (2 pkt) Mam dane opisowe nt. wielu szczepów tej samej bakterii (np. komórki duże/małe, kolor, grubość kolonii, szybkość wzrostu itp.). Chciałbym je pogrupować, aby ew. testować podatność różnych podtypów bakterii na interesujący mnie antybioktyk. Czy mogę użyć metod k-średnich, k-medoidów czy metod hierarchicznych? Jaki wpływ na wynik będzie miał wybór funkcji podobieństwa.

8. (2 pkt) Podaj przykłady dwóch zbiorów danych dla których dla redukcji wymiaru:

   - lepiej jest zastosować analizę składowych głównych niż skalowanie wielowymiarowe
   - można zastosować skalowanie wielowymiarowe, ale nie analizę składowych głównych.

Powodzenia!