

Statystyczna analiza danych w języku R

Zadanie zaliczeniowe

Zadanie

Zadanie zaliczeniowe polega na analizie dwóch zbiorów danych mikromacierzowych (linki poniżej). Analiza powinna zawierać metody klastrowania, wizualizacji danych (również te specyficzne dla mikromacierzy), klasyfikacji, selekcji biomarkerów i redukcji wymiaru. Wśród zastosowanych metod powinny się znaleźć te, które ćwiczyliśmy (lub będziemy ćwiczyć) na laboratorium, jak również inne metody (dostępne lub łatwe do zaimplementowania w języku R), które zostały użyte w literaturze. Każda grupa powinna znaleźć artykuł(y), w których były analizowane dane mikromacierzowe, zapoznać się z zastosowanymi tam metodami i powtórzyć je na naszych danych¹. Polecam szczególnie prace, w których już były analizowane dane Goluba. Proponuję sprawdzić następujące linki:

- prace na stronie laboratorium MIT, z którego pochodziły dane Goluba:
<http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi>
- cytowania artykułu Goluba w PubMed:
<http://www.pubmedcentral.nih.gov/tocrender.fcgi?action=citedtool=pubmedpubmedid=10521349>

Pracę źródłową trzeba wybrać do 10 maja i omówić ze mną w czasie laboratorium.

UWAGA: Przy ocenie brana będzie pod uwagę m.in. skuteczność zastosowanych metod. Istotne jest więc sprawdzenie różnych algorytmów i wybranie tych, które działają najlepiej dla danego zbioru danych. Jeśli natomiast ktoś zaimplementuje (zastosuje) ciekawą metodę z literatury, która akurat nie daje dobrych wyników dla „naszych” zbiorów danych, to może ratować swój wybór prezentując wyniki dla zbioru Goluba - zawsze trzeba jednak przedstawić również wyniki dla naszych zbiorów. (Dlatego szczególnie zachęcam do korzystania z prac analizujących właśnie dane Goluba - powinno być wtedy wiadomo, czy dana metoda dobrze działa przynajmniej na tych danych.)

Żeby zaliczyć zadanie należy zaimplementować wybrane metody w R-rze, przeprowadzić eksperymenty na danych oraz opracować raport (w formacie .pdf) zawierający analizę dwóch zbiorów danych (ew. również danych Goluba). Raport powinien zawierać opis przeprowadzonych eksperymentów, wybrane istotne wyniki (zarówno wizualizacje jak i tabelki z wynikami klasyfikacji i statystykami) oraz wnioski, a także referencje do wybranych prac źródłowych. Kod w

¹Nie trzeba koniecznie powtórzyć całej analizy przeprowadzonej w wybranej pracy - można sobie coś wybrać

R-rze musi umożliwiać automatyczne odtworzenie eksperymentów omówionych w raporcie. Po uruchomieniu program powinien wczytać dane z plików, przeprowadzić obliczenia i zapisać do plików wyniki poszczególnych eksperymentów (obrazki lub tabele).

Zadanie należy wykonać w dwuosobowych zespołach. Skład zespołów trzeba podać do 19 kwietnia. Zadanie należy oddać do 7 czerwca (wcześniej trzeba ustalić ze mną termin zaliczenia).

Punktacja

Praca będzie oceniona w trzech kategoriach: implementacja, raport oraz dobór metod i skuteczność klasyfikacji. Za każdą część można dostać maksymalnie 10 punktów. Poniżej przedstawiam krótki opis tego, na co będę zwracał uwagę przy przyznawaniu punktów.

- Implementacja 10 pkt.
 - Działanie
Kod w R-rze powinien po uruchomieniu powtórzyć eksperymenty opisane w raporcie i zapisać wyniki do plików (patrz wyżej).
 - Rozsądna ilość zaimplementowanych metod
Trzeba maksymalnie wykorzystać to co było na laboratorium oraz zaimplementować metodę z literatury.
 - Poprawność implementacji wybranych metod
 - Korzystanie z funkcji R'a
Korzystanie z wbudowanych funkcji zamiast np. używania pętli itp..
 - Modularność
Trzeba zaprojektować proces przetwarzania danych według schematu: wczytanie danych → analiza → zapis wyników do plików. Do wczytania danych mogą być osobne funkcje dla różnych zbiorów, ale funkcje na kolejnych etapach (implementujące poszczególne metody analizy i zapis danych) powinny już być ogólne, tak żeby nie powtarzać implementacji tego samego procesu.
- Raport 10 pkt.
 - Prezentacja analizy dwóch zbiorów danych (i ew. zbioru Goluba).
 - Poprawne wykorzystanie różnych metod wizualizacji danych.
 - Tabelki z wynikami walidacji krzyżowej dla metod klasyfikacyjnych.
 - Podsumowanie wyników i wnioski.

- Dobór metod i skuteczność klasyfikacji 10 pkt.
 - Skuteczność klasyfikacji w teście walidacji krzyżowej.
 - Dobór metod klasyfikacji i selekcji biomarkerów/redukcji wymiaru
Powinny zostać użyte zarówno metody z laboratorium jak i z wybranej pracy źródłowej.
 - Poprawność oceny wyników.
Poprawność testu walidacji krzyżowej – selekcja atrybutów i trenowanie klasyfikatora tylko na części treningowej danych.
 - Omówienie pracy źródłowej do 10 maja

Oceny

| Punkty | Ocena |
|----------|-------|
| 27 - 30+ | 5 |
| 24 - 26 | 4+ |
| 21 - 23 | 4 |
| 18 - 20 | 3+ |
| 15 - 17 | 3 |

Bonus

Są trzy sposoby zarobienia dodatkowych punktów:

1. Najlepszy wynik klasyfikacji 2,4,6 pkt.
Za najlepszy wynik klasyfikacji (na jednym z dwóch zbiorów danych) 6 pkt., za drugi wynik 4 pkt., za trzeci 2 pkt.. Wynik klasyfikacji liczony ma być jako średnia skuteczność z 10-ciu przebiegów walidacji krzyżowej z podziałem na 10 części.
2. Ciekawa metoda 1-6 pkt.
Zastosowanie ciekawej metody z wybranej publikacji wymagającej więcej pracy implementacyjnej. Wybór i liczbę dodatkowych punktów można konsultować.
3. Eksperyment z pracy Simona 3 pkt.

Przeprowadzenie i zaprezentowanie eksperymentu omówionego w pracy Simona (<http://jnci.oxfordjournals.org/cgi/content/extract/95/1/14>) porównującego wyniki klasyfikacji w teście walidacji krzyżowej, przy zastosowaniu wyboru biomarkerów na wszystkich danych (również testowych) oraz wyłącznie na treningowej części danych.

Uwaga: nie wymaga wiele pracy.

Zbiory danych

1. Dane mikromacierzowe prof. Jarzab - dotyczą raka tarczycy. Dane i dodatkowe informacje są tu:

<http://www.genomika.pl/thyroidcancer/PTCCancerRes.html>

Zbiór danych składa się 50 próbek od 23 pacjentów chorych na raka tarczycy (PTC) i 10 pacjentów z innymi chorobami tarczycy (benign). Dla 16 pacjentów (initial_set) pobrano próbki tkanki chorej i zdrowej (contralateral). Szczegóły w pracy prof. Jarzab (link ze strony wymienionej powyżej). Klasyfikację (i wybór biomarkerów) należy zrobić na 2 klasy (contralateral + benign) vs. PTC (podział na initial_set i validation_set należy pominąć, gdyż wyniki testujemy przy użyciu walidacji krzyżowej). Klasteryzacja i wizualizacje powinny uwzględniać podział na 3 grupy (PTC, contralateral i benign). Należy zwrócić uwagę na próbki pochodzące od tych samych pacjentów i opisać ewentualne zależności (jeżeli takie będą).

2. Dane mikromacierzowe prof. Ostrowskiego - dotyczą choroby refluksowej: <http://bioputer.mimuw.edu.pl/janusz/sadR/dane>

Dane przetworzone są na dwa sposoby (można sobie wybrać wersję, na której się pracuje). Login i hasło podaję na laboratorium.

Zbiór danych składa się z obserwacji ekspresji 22277 genów u 91 pacjentów z 3 klas (oddzielne arkusze w pliku .xls):

- Nonerosive reflux disease (NERD)
- Erosive reflux disease (ERD)
- Barrett's esophagus (BE)

Powyższe stany chorobowe wymienione są w kolejności od najmniej do najbardziej poważnej. Tzw. przełyk Barretta powodowany jest przez przewlekłą chorobę refluksową. Więcej na ten temat można się dowiedzieć np. tu:

<http://www.epacjent.pl/arttykul.php?idartykul=648poddzial=Gastrologia>

W idealnym przypadku klasyfikator powinien dzielić pacjentów na trzy grupy. W razie niepowodzenia, można też spróbować klasyfikacji na dwie grupy (NERD+ERD vs. BE) lub (NERD vs. ERD + BE).

FAQ ?