Statystyczna Analiza Danych 2023

Projekt II, deadline: 11 czerwca 2023, godz. 23:59

Informacje wstępne

Opis danych

Narządy, takie jak trzustka, składają się z wielu typów tkanek, a te z kolei z wielu typów komórek. W obrębie trzustki możemy wyróżnić komórki typowe wyłącznie dla tego narządu, takie jak komórki alfa czy beta, ale także komórki związane z ukrwieniem czy układem immunologicznym.

Dane w tym zadaniu pochodzą z wielomodalnego sekwencjonowania pojedynczej komórki (ang. *multimodal single cell RNA sequencing*, scRNA-seq). Użycie *scRNA-seq* pozwala na studiowanie próbek w wysokiej rozdzielczości i oddzielenie od siebie komórek różnych typów. Możliwe jest między innymi porównanie komórek patologicznych, pobranych od pacjentów nowotworowych, z komórkami zdrowymi. W technologii multimodal *scRNA-seq* dla każdej komórki otrzymujemy dwa typy odczytów:

- **Zliczenia transkryptów RNA** odpowiadające ekspresji (aktywności) genów w danej komórce;
- **Ilość białek powierzchniowych (ang. *protein abundance)***, która jest wprost związana z typem danej komórki.

Wynikiem eksperymentu *scRNA-seq* są macierze, w których dla każdej komórki przypisany jest sygnał RNA z wielu tysięcy genów (w naszym zadaniu X) oraz sygnał pochodzący z kilkudziesięciu białek powierzchniowych (w naszym zadaniu dla uproszczenia wybraliśmy pojedyncze białko CD71[1], y).

Zgodnie z centralnym dogmatem biologii, wiemy, że informacja genetyczna przepływa z RNA na białka. Tym samym, należy spodziewać się korelacji między ilością białka a ekspresją genu, który to białko koduje. Z przyczyn technicznych i biologicznych, ta zależność niejednokrotnie ulega degeneracji. Problem w tym zadaniu polega na predykcji sygnału z białek powierzchniowych na podstawie ekspresji genów. Przewidywanie sygnału *protein abundance* jest kluczowe dla większości publicznie dostępnych zbiorów, dla których dostępna jest wyłącznie macierz RNA. Analiza sygnału o ekspresji genów i ilości białek powierzchniowych znacząco ułatwia proces identyfikowania i nazywania komórek w próbce.

Dane zostały pobrane z szpiku kostnego ludzkich dawców. Zebrane komórki to w większości komórki układu immunologicznego. Prawidłowe zidentyfikowanie limfocytów typu T w oparciu o oba typy odczytów w zbiorze takiego typu mogłoby być podstawą do rozwijania celowanych terapii nowotworowych (dla ciekawych: CAR T cell therapy).

[1] `https://en.wikipedia.org/wiki/Transferrin_receptor_1`

Sposób pobrania danych
Na przedmiotowej stronie Moodle znajduje się link do folderu z danymi dla każdej z grup laboratoryjnych. Ponieważ każda grupa pracuje na danych pochodzących z innego eksperymentu, wyniki pomiędzy grupami mogą się różnić. Dane są skompresowane oraz zapisane w formacie .csv. Udostępnione będą trzy pliki:

- `X_train.csv` oraz `X_test.csv`, zawierające macierze RNA. Każdy wiersz odpowiada komórce, kolumna genowi, natomiast wartości to poziom ekspresji. Kolumny tych macierzy to nasze *zmienne objaśniające*.
- `y_train.csv`, odpowiadający ilości białka powierzchniowego pewnego typu w komórkach (tych, których dotyczyły dane z pliku `X_train.csv`). Jest to nasza *zmienna objaśniana*.

W dalszej części opisu, dane z plików `X_train.csv` i `y_train.csv` będziemy nazywać treningowymi, a dane z pliku `X_test.csv` będziemy nazywać testowymi.

Sposób oddania projektu
Należy wysłać mailem na adres prowadzącego następujące pliki.

- Raport w formacie .pdf lub .html, realizujący opisane niżej polecenia (szablon nazwy pliku: `NrIndeksu_raport.ext`, gdzie `ext` to odpowiednie rozszerzenie).
- Kod zródłowy (np. .Rmd lub .R) generujący rozwiązania zadań (szablon nazwy pliku: `NrIndeksu_kod.ext`, gdzie `ext` to odpowiednie rozszerzenie).
- Wyniki predykcji na danych testowych (patrz zadanie 4) w formie pliku .csv, zawierającego kolumnę `Id` z numerami obserwacji oraz kolumnę `Expected` z wartościami predykcji (szablon nazwy pliku: `NrIndeksu_predykcja.csv`)

Dodatkowo plik z predykcją (opisany w ostatnim podpunkcie) powinien zostać wysłany na stronę konkursu Kaggle. Link do konkursu zostanie udostępniony na stronie Moodle przedmiotu do 26-ego maja. W ramach Kaggle każdy student będzie mógł zgłaszać wiele propozycji predykcji i w ten uzyskiwać pewne informacje o jej jakości, a także jej aktualną pozycję w rankingu. Przy ostatecznej ocenie uwzględniany jest jednak tylko plik wysłany do prowadzącego laboratoria.

Ocena
Za cały projekt można otrzymać 15 punktów. Maksymalne liczby punktów za każdy z poniższych podpunktów podane są w nawiasach. Ocena zadań od 1 do 4 będzie uwzględniała

- realizację przedstawionych poleceń,
- jakość raportu w formie .pdf lub .html (wizualizacje, czytelność tekstu, opis wyników),
- jakość wykorzystanego w tym celu kodu. Warto zadbać o to, by był on czytelny i reprodukowalny.

Dodatkowe informacje na temat szczegółów punktacji można uzyskać u swojego prowadzącego laboratorium.

Uwaga
Zadania oddane po terminie nie podlegają ocenie i otrzymują 0 punktów.

Treści zadań

1. Eksploracja *(3 pkt.)*

(a) Sprawdź, ile obserwacji i zmiennych zawierają wczytane dane treningowe oraz testowe. Przyjrzyj się typom zmiennych i, jeśli uznasz to za słuszne, dokonaj odpowiedniej konwersji przed dalszą analizą. Upewnij się, czy dane są kompletne.

(b) Zbadaj rozkład empiryczny zmiennej objaśnianej (przedstaw kilka podstawowych statystyk, do analizy dołącz histogram lub wykres estymatora gęstości).

(c) Wybierz 250 zmiennych objaśniających najbardziej skorelowanych ze zmienną objaśnianą. Policz korelację dla każdej z par tych zmiennych. Zilustruj wynik za pomocą mapy ciepła (*heatmap*).

Uwaga: opisany tu wybór zmiennych jest tylko na potrzeby niniejszego podpunktu, analizę opisaną w kolejnych zadaniach należy przeprowadzić na **pełnym** zbiorze danych treningowych.

2. ElasticNet *(3 pkt.)*

Pierwszy model, który należy wytrenować, to *ElasticNet*. Podczas wykładu spotkaliśmy się z jego szczególnymi przypadkami: regresją grzbietową (*ridge regression*) oraz lasso.

(a) Wyszukaj i przedstaw w raporcie informacje o modelu ElasticNet. Opisz parametry, które są w nim estymowane, optymalizowaną funkcję oraz hiperparametry, od których ona zależy. Dla jakich wartości hiperparametrów otrzymujemy regresję grzbietową, a dla jakich lasso?

(b) Zdefiniuj siatkę (*grid*) hiperparametrów, opartą na co najmniej trzech wartościach każdego z hiperparametrów. Zadbaj o to, by w siatce znalazły się konfiguracje hiperparametrów odpowiadające regresji grzbietowej i lasso. Użyj walidacji krzyżowej do wybrania odpowiednich hiperparametrów (o liczbie podzbiorów użytych w walidacji krzyżowej należy zdecydować samodzielnie oraz uzasadnić swój wybór).

(c) Podaj błąd treningowy i walidacyjny modelu (należy uśrednić wynik względem wszystkich podzbiorów wyróżnionych w walidacji krzyżowej).

3. Lasy losowe *(3 pkt.)*

W tej części projektu należy wytrenować model lasów losowych i porównać jego działanie z utworzonym wcześniej modelem ElasticNet.

(a) Spośród wielu hiperparametrów charakteryzujących model lasów losowych wybierz trzy różne. Zdefiniuj trójwymiarową siatkę przeszukiwanych kombinacji hiperparametrów i za pomocą walidacji krzyżowej wybierz ich optymalne (w kontekście wykonywanej predykcji) wartości. Wykorzystany przy walidacji krzyżowej podział danych powinien być taki sam, jak w przypadku ElasticNet.

(b) Zrób podsumowanie tabelaryczne wyników, jakie otrzymywały metody w walidacji krzyżowej w obu rozważanych modelach. (Porównanie to jest powodem, dla którego zależy nam na zastosowaniu tych samych podziałów). Określ, który model wydaje Ci się najlepszy (uzasadnij swój wybór). Do porównania dołącz podstawowy model referencyjny, który dowolnym wartościom zmiennych objaśniających przypisuje średnią arytmetyczną zmiennej objaśnianej.

4. Predykcja na zbiorze testowym *(6 pkt.)*

Ta część projektu ma charakter otwarty. W oparciu o dane treningowe należy dopasować dowolnie wybrany model, a następnie zastosować go do przewidywania wartości zmiennej objaśnianej w zbiorze testowym. Sposób wyboru i budowy modelu, a także motywacje stojące za takim wyborem powinny zostać opisane w raporcie. Wygenerowane predykcje należy wysłać do prowadzącego w osobnym pliku, którego format został opisany wcześniej, a także umieścić na stronie Kaggle (odpowiedni link zostanie udostępniony na Moodle). Liczba uzyskanych punktów zależeć będzie od jakości predykcji, mierzonej pierwiastkiem błędu średniokwadratowego, RMSE.

Wskazówka: Warto rozważyć modele zaimplementowane w pakiecie `caret`. Polecamy również wypróbować różne techniki wspomagające uczenie maszynowe (np. odpowiedni dobór podzbioru zmiennych objaśniających, transformacje zmiennych lub redukcja wymiaru).

Szczegóły punktacji:

(1 pkt.) – za błąd niższy od pochodzącego z opisanego wcześniej, podstawowego modelu referencyjnego.

(2 pkt.) – za błąd niższy od pochodzącego z modelu ElasticNet wytrenowanego przez prowadzących laboratoria.

(3 pkt.) – ten bonus obliczany jest według wzoru $\frac{1}{2}\lfloor 6\widehat{F}(e)^3 \rfloor$, gdzie e to błąd testowy predykcji studenta, $\widehat{F}$ jest dystrybuantą empiryczną błędów wszystkich zgłoszonych predykcji w grupie laboratoryjnej studenta, natomiast $\lfloor \cdot \rfloor$ to część całkowita.