Natalia Machlus – opis projektu nr 2 healthy decades

1. Określenie zakresu problemu

Celem projektu jest zbudowanie modelu regresji liniowej przewidującego wiek pacjenta na podstawie własności obliczonych w ramach pierwszego projektu.

2. Wybranie metryki wydajności

W przypadku zagadnień regresyjnych klasyczną miarą wydajności jest pierwiastek błędu średniokwadratowego. Wyraża stopień w jakim model myli się w przewidywaniach, wraz ze wzrostem błędu waga metryki rośnie.

3. Pozyskanie danych

W moim projekcie korzystam z poprawionej tabeli, którą stworzyłam w ramach projektu 1, a następnie zapisałam do pliku formatu xls. Razem z projektem załączyłam już utworzoną tabelę, którą następnie wczytuję do programu Spyder rozpoczynając tym tworzenie projektu 2.

4. Podstawowe informacje o tabelach

Tabela, którą nazwałam tabela_gl zawiera dane ogólne. Dla każdego pacjenta wyliczone zostały poszczególne własności umieszczone w kolumnach. Druga tabela o nazwie tabela_sr analizuje każdego pacjenta w oknie(o długości 100 pomiarów) o największym maksymalnym średnim RR, natomiast tabela_std w oknie(o długości 100 pomiarów) o największym odchyleniu standardowym.

Dla każdej tabeli sprawdziłam typ danych w kolumnach oraz obecność braków danych. Każda z nich zawiera kolumnę nienumeryczną z oznaczeniem płci oraz posiada wszystkie dane kompletne.

5. Podział na dane testujące i trenujące

Używając funkcji StratifiedShuffleSplit przeprowadziłam próbkowanie warstwowe na podstawie wieku, aby zbiór testujący był reprezentatywny. Następnie sprawdzam proporcje wieku w zestawie testowym oraz dla porównania w całym zbiorze danych.

Kolejnym krokiem jest rozdzielenie czynników prognostycznych od etykiet zarówno w danych testujących jak i w trenujących.

W celu zbadania korelacji między poszczególnymi zmiennymi oraz ich wizualizacji tworzę kopię całego zbioru uczącego (zawierającego etykiety).

6. Poznanie i wizualizacja danych trenujących

Dla każdej z tabel, na skopiowanych danych trenujących dokonuje zamiany danych nienumerycznych w kolumnie z płecią na numeryczne a następnie wyświetlam wartości korelacji każdej kolumny ze zmienną celu „dekada”. Następnie wyświetlam wykres korelacji dla 4 zmiennych o największej korelacji dodatniej ze zmienną „dekada”. Widać że niektóre ze zmiennych są ze sobą skorelowane:

W tabeli bazowej:

P(0da) jest skorelowane dodatnio z P(da0), P(0d) oraz skorelowane ujemnie z P(aad).

P(da0) jest skorelowane dodatnio z P(0da), P(0d) oraz ujemnie z P(aad).

P(0d) jest skorelowane dodatnio z P(0da), P(da0) oraz ujemnie z P(aad).

P(aad) jest skorelowane ujemnie z P(0da), P(da0), P(0d).

W tabeli z największym średnim RR:

P(0) jest skorelowane dodatnio z P(0d), P(a0).

P(0d) jest skorelowane dodatnio z P(0), P(a0).

P(a0) jest skorelowane ujemnie z P(0), P(0d).

W tabeli bazowej o największym odchyleniu RR:

P(ada) jest skorelowane dodatnio z P(dad).

P(dad) jest skorelowane dodatnio z P(ada)

7. Przygotowanie danych pod algorytmy uczenia maszynowego

W każdej tabeli zamieniam dane nienumeryczne na numeryczne w danych uczących bez etykiet. Tworzę dla każdej tabeli osobny encoder, którym później będę przekształcać dane testujące.

Następnie każdą tabelę skaluję osobnym scalerem za pomocą transformatora Standardscaler.

8. Redukcja wielowymiarowości metodą PCA

Dla każdej tabeli tworzę najpierw dwa ogólne wykresy obrazujące jak ilość wybranych składowych głównych wpływa na wyjaśnianą wariancję. Z pierwszego wykresu odczytuję, że dla tabeli bazowej już mniej niż 10 komponentów wyjaśnia ponad 95% wariancji. Dla tabeli z oknem o max średnim RR wybrane 10 składowych wyjaśniałoby 90% wariancji. Dla tabeli z max odchyleniem, wybranie 10 składowych oznaczałoby wyjaśnienie około 85% wariancji.

Następnie rozważam wybór dwóch składowych dla każdej z tabel, co obrazuję na wykresach dwuwymiarowych. Na wykresach również wyświetla się procent wariancji wyjaśniany przez daną składową.

Dodatkowo rozważyłam również dla każdej tabeli redukcję do trzech komponentów, co przedstawiłam na wykresie trójwymiarowym.

Ostatecznie zdecydowałam się na wybranie tylu komponentów aby wyjaśniały 95% wariancji. Na takiej przetransformowanej tabeli będę uczyła model.

9. Wybór i uczenie modelu

Przedstawiane teraz kroki były wykonywane dla każdej z tabel osobno.

Na początku utworzyłam model regresji liniowej stochastycznego gradientu przy domyślnych parametrach i uczyłam go na danych trenujących. Następnie przetestowałam go na całym zbiorze uczącym i obliczyłam błąd RMSE.

Kierując się ciekawością innych rozwiązań porównałam wyniki modelu SGDRegressor przy krzyżowej walidacji na 4 zbiorach walidacyjnych z trzema różnymi modelami : LinearRegression, RandomForestRegressor oraz DecisionTreeRegressor.

Metodą przeszukiwania siatki szukałam najlepszych hiperparametrów dla modelu SGDRegressor.

Po znalezieniu ostatecznego modelu przeprowadziłam ostateczne uczenie na przygotowanych wcześniej danych trenujących.

Przygotowałam do testowania dane testujące używając wcześniej przygotowanych maszyn do kodowania, skalowania oraz redukcji wielowymiarowości. Na takich przygotowanych danych przetestowałam swój model oraz podałam końcowy wynik błędu RMSE.