Modelowanie Statystyczne w Zarządzaniu Wierzytelnościami Masowymi

Laboratorium 7.

1. Przygotuj dane aplikacyjne do analizy eksploracyjnej oraz wytypuj cechy do analizy korelacji.
2. Do zbioru cech aplikacyjnych dodaj cechę wydzielającą klientów, którzy dokonali przynajmniej 3 wpłat lub przekroczyli skuteczność 0.5% w pierwszych 6 miesiącach obsługi (klienci dobrzy).
3. Stwórz i dokonaj analizy wykresów uzyskanych za pomocą funkcji `corrgram` oraz `corrplot` – czy widzisz grupy zmiennych powielających informację?
4. Stwórz tabelę zawierającą współczynnik VIF dla każdej ze zmiennych – dokonaj korekty listy potencjalnych zmiennych objaśniających i przelicz tabelę.

Rozgrzewka

- wygeneruj losowy zbiór obserwacji z dwuwymiarowego rozkładu normalnego
- zdekomponuj zbiór danych na czynniki obliczając wektory własne macierzy kowariancji lub korelacji
- zaprezentuj wylosowane dane na wykresie i zaznacz kierunki główne wskazywane przez wektory własne

5. Zestandaryzuj cechy objaśniające. Dlaczego jest to ważne w kontekście PCA?
6. Oblicz wartości własne oraz wyznacz wektory własne macierzy korelacji za pomocą funkcji `eigen` lub `prcomp`.
7. Oblicz współrzędne danych w przestrzeni głównych składowych.
8. Stwórz oraz zinterpretuj wykres przedstawiający udział wariancji całkowitej wyjaśnianej przez poszczególne główne składowe.
9. Oblicz macierz korelacji oryginalnych zestandaryzowanych danych z głównymi składowymi – czy cechy najsilniej skorelowane z dobrocią klienta są silnie skorelowane z pierwszymi głównymi składowymi?
10. Zaznacz na wykresie cechy objaśniające w przestrzeni korelacji cech z dwiema pierwszymi głównymi składowymi.
11. Przedstaw zbiór danych na wykresach w przestrzeniach wybranych par głównych składowych w podziale na dobroć klienta – która para głównych składowych najlepiej dyskryminuje grupy?
12. Dla każdej pary głównych składowych oblicz wybraną miarę siły dyskryminacji klientów dobrych i złych – czy najlepsze są pierwsze główne składowe?
13. Zinterpretuj główne składowe.