Wprowadzenie

Uczenie Maszynowe
uczeniemaszynowe@cs.uni.wroc.pl

14.10.2025, 10:00
pracownia

Uwagi ogólne:

- Proponuję zadania wykonać w języku python w notebookach jupyterowych, z wykorzystaniem odpowiednich pakietów, takich jak `matplotib` i/lub `seaborn` i/lub `plotly` (do robienia wykresów), `pandas` i/lub `numpy` (do wczytywania i przetwarzania danych).
- Proszę zadbać o czytelność i zrozumiałość tworzonych wykresów, m.in. odpowiedni opis osi, odpowiedni opis danych, odpowiedni dobór kolorów, odpowiedni dobór wielkości czcionki, itp.
- Na pytania zawarte w treści zadań należy umieć odpowiedzieć prowadzącemu zajęcia, w oparciu o przygotowane dane i wykresy.

Zadanie 1. [2.5 pkt]

0. Znajdź i pobierz zbiór danych IRIS z repozytorium danych UCI. Dowiedz się co opisują pobrane dane (wskazówka: przeczytaj plik `iris.names`, a terminy biologiczne rozjaśnią ilustracje z wikipedii).
1. Jakie wartości przyjmuje `sepal length`? Zrób wykres (typu scatter plot): na osi X numer próbki danych, na osi Y wartość `sepal length`. Czy jest sens łączyć kropki na tym wykresie?
2. Sprawdź ile jest próbek danych klasy iris setosa, dla których `sepal width` jest mniejszy niż 2.5. A między 2.5 i 3.0? A między 3.0 i 3.5? Itd. Zrób odpowiedni wykres (typu słupkowego). Czy zrobiłeś histogram?
3. Zrób wykres (typu scatter plot): na osi X `sepal length`, na osi Y `sepal width`. Czy jest sens łączyć kropki na tym wykresie? Pokoloruj kropki trzema różnymi kolorami w zależności od gatunku irysa. Ustaw rozmiar kropek proporcjonalnie do petal length (skalę rozmiarów dobierz tak, aby rysunek był czytelny). Co ciekawego widać na tym wykresie? Czy potrafisz wykorzystać wykres, żeby podać regułę odróżniania iris setosa od pozostałych dwóch gatunków irysa?
4. Zrób wykres z poprzedniego punktu dla każdej pary różnych atrybutów danych. Przemyśl, czy nie będzie wygodnie ustawić wykresy w jakiś sposób obok siebie, a nie jeden pod drugim.
5. Dowiedz się czym jest wykres skrzypcowy (ang. violin plot) i zrób go dla atrybutów danych.

Zadanie 2. [2.5 pkt]

0. Znajdź i pobierz zbiór danych BANK MARKETING z repozytorium danych UCI. Dowiedz się co opisują pobrane dane.
1. Zrób histogram atrybutu duration dla całości danych. Następnie zrób analogiczne dwa histogramy, osobno dla próbek pozytywnych i negatywnych (atrybut `y` równy `yes` albo `no`, odpowiednio). Co ciekawego widać na tych wykresach?
2. Zrób histogram atrybutu balance dla osób powyżej 25 roku życia (atrybut `age` większy niż 25). Zapoznaj się z pakietem `ipywidgets`, a następnie przekształć swój wykres w interaktywny z suwakiem do wybierania granicy wieku i zobacz jak zmienia się histogram.

3. Dla ustalonego progu, $t = 360$, czasu trwania rozmowy (atrybut `duration`), wyznacz próbki danych z atrybutem duration powyżej progu t. Policz ile procent z nich to próbki pozytywne. Powtórz obliczenia dla różnych t i zrób wykres: na osi X próg t, na osi Y procent próbek pozytywnych. Spróbuj wytłumaczyć zjawisko widoczne na wykresie.

4. Powtórz poprzedni punkt dla atrybutu `balance`. Wyobraź sobie, że przeprowadzasz kampanię marketingową lokat bankowych. Możesz kontaktować się z losowo wybranymi (z rozkładem jednostajnym na całych danych) klientami banku, albo jedynie z klientami o saldzie (atrybut `balance`) powyżej ustalonego progu (możesz sam ustalić ten próg). Która strategia jest lepsza? Jaki próg wybrać?

5. Podziel losowo zbiór danych na dwie części porównywalnej wielkości. Używając pierwszej części danych, zrób wykres opisany w pkt. 4 i ustal próg (według własnego pomysłu). Przyjmij, że klienci z saldem powyżej ustalonego progu powinni być zainteresowani lokatą bankową (powinni być próbkami pozytywnymi). Oblicz ilu z nich, procentowo, nie było. Oblicz też ilu zainteresowanych lokatą klientów miało saldo nie większe od ustalonego progu. Następnie policz takie błędy modelu na drugiej części danych. Powtórz obliczenia dla różnych wartości progów. Zrób wykres (według własnego pomysłu) prezentujący wyniki. Ostatecznie, to jaki próg wybrać?

Zadanie 3. [2 pkt]

0. Znajdź i pobierz zbiór danych CAR EVALUATION z repozytorium danych UCI. Dowiedz się co opisują pobrane dane.

1. Na wykresie tortowym pokaż ile spośród wszystkich samochodów to samochody nieakceptowalne, akceptowalne, dobre i bardzo dobre.

2. Na wykresie słupkowym pokaż ile jest samochodów 2, 3 i 4 drzwiowych w różnych klasach bezpieczeństwa.

3. Dowiedź się czym jest wykres radarowy (ang. radar plot), nazywany też gwiazdowym lub pajęczynowym (ang. star plot lub spider plot), i zrób go dla 5 wybranych atrybutów danych.

4. Spróbuj określić jakie wartości jakich atrybutów decydują o tym, że samochód jest dobry lub bardzo dobry. Zrób to intuicyjnie, według własnego pomysłu - w dalszej części wykładu nauczysz się to robić metodami uczenia maszynowego. Sformułuj regułę, która na podstawie wartości kilku wybranych atrybutów mówi TAK (dobry lub bardzo dobry) lub NIE. Czy Twoja reguła działa prawidłowo dla wszystkich danych czy tylko dla części, i jak dużej części? Zrób rysunek (według własnego pomysłu) pokazujący gdzie reguła robi błędy.

Zadanie 4. [1.5 pkt]

0. Przypomnij sobie, albo posłuchaj przypomnienia na wykładzie, czym jest rozkład normalny.

1. Wygeneruj milion liczb z rozkładu normalnego o średniej 0 i wariancji 1 używając funkcji `numpy.random.randn(...)`. Narysuj histogram takich danych.

2. Na histogramie wygenerowanych danych narysuj funkcję gęstości rozkładu prawdopodobieństwa. Porównaj histogram z funkcją gęstości.

3. Przypomnij sobie pkt 3 z Zadania 1. Do atrybutów `sepal length` i `sepal width` dodaj zaburzenie losowe z rozkładem normalnym o średniej 0 i odchyleniu standardowym $s = 0.25$. Zobacz jak zmienia się wykres. Czy zaproponowana przez Ciebie reguła odróżniania iris setosa nadal działa? Spróbuj też użyć innych wartości s.

Zadanie 5. [1.5 pkt]

0. Przypomnij sobie, albo posłuchaj przypomnienia na wykładzie, czym jest wielowymiarowy rozkład normalny.

1. Wygeneruj milion punktów z dwuwymiarowego rozkładu normalnego o średniej (0, 0) i identycznościowej macierzy kowariancji używając funkcji `numpy.random.randn(...)`. Narysuj histogram takich danych (wskazówka: histogram będzie wykresem 3D).

2. Na histogramie wygenerowanych danych narysuj funkcję gęstości rozkładu prawdopodobieństwa (wskazówka: wykres będzie 3D). Porównaj histogram z funkcją gęstości (wskazówka: wygodnie będzie zrobić wykres dynamiczny, umożliwiający obracanie i powiększanie, umożliwia to na przykład pakiet plotly).