Naiwny klasyfikator Bayesa

Przemysław Klęsk & Joanna Kołodziejczyk

1 Problem do rozwiązania

Celem jest implementacja NBC w wersji dyskretnej dla zbioru „wine” z repozytorium UCI `https://archive.ics.uci.edu/ml/index.php`. Opis zadania polega na implementacji klasyfikatora w dwóch wariantach w języku Python.

Przed przystąpieniem do implementacji z repozytorium UCI pobierz zbiór danych o nazwie „wine” dotyczący klasyfikacji wina na podstawie składu chemicznego i zapoznaj się z nim. Zwróć uwagę, która ze zmiennych jest zmienną decyzyjną.

2 Kolejność zadań dla NBC dla zmiennych dyskretnych

1. Wczytaj dane z pobranego pliku tekstowego `wine.data` do macierzy `numpy` (wykorzystaj funkcję `numpy.genfromtxt`) i rozdziel tę macierz na dwie macierze `X` (o wymiarze 178×13) i `y` (178×1 — etykiety klas).

2. Dyskretyzację danych „wine” można wykonać wykorzystując gotowy obiekt `KBinsDiscretizer` (z pakietu `sklearn.preprocessing`) lub samodzielnie na poziomie opracowywanej klasy NBC (liczba przedziałów, na którą dyskretyzujemy cechy oryginalnie ciągłe, powinna być parametrem nastawialnym przez użytkownika).

3. Podziel dane na część uczącą i testową (wykorzystaj funkcję `train_test_split` z pakietu `sklearn.model_selection`).

4. Napisz klasę reprezentującą naiwny klasyfikator Bayesa w wariancie ze zmiennymi dyskretnymi. Klasę przygotuj zgodnie z ideą biblioteki `scikit-learn` — m.in.: wykonaj dziedziczenie po klasach `BaseEstimator` i `ClassifierMixin`, przygotuj metody `fit` (uczenie) i `predict` (klasyfikowanie) oraz pomocniczo `predict_proba`.

5. Zastanów się i zaplanuj wg własnego uznania wygodne struktury danych do przechowywania:

   - rozkładu a priori klas $P(Y = y)$,
   - rozkładów warunkowych $P(X_j = v|Y = y)$.

   Mogą to być tablice, słowniki, listy lub odpowiednie połączenia / zagnieżdżenia tych struktur. Do tego celu potrzebne będzie także ustalenie dyskretnych dziedzin zmiennych, tj. wykrycie, jakie unikalne wartości poszczególne zmienne mogą przyjmować, np. z wykorzystaniem funkcji `numpy.unique`. Przemyśl, czy informacje o dziedzinach należy zdobywać na poziomie funkcji `fit` na podstawie danych uczących, czy też lepiej przekazać je klasyfikatorowi już podczas konstrukcji.

6. Wyznacz dokładność ($Accuracy = \frac{NP}{LZ} * 100\%$, gdzie NP — liczba próbek poprawnie sklasyfikowanych w zadanym zbiorze, LZ — liczność zbioru) otrzymanego klasyfikatora na zbiorach uczącym i testowym.

7. Obliczenia powtórz uwzględniając poprawkę LaPlace'a (możesz do tego celu wprowadzić przełącznik w konstruktorze Twojej klasy). Zwróć uwagę, czy poprawka LaPlace'a podnosi dokładność testową dla tego zbioru danych.

2.1 Uczenie NBC

Uczenie NBC w wariancie dyskretnym polega na wyznaczeniu i zapamiętaniu (w pewnej strukturze danych, np. w tablicy lub słowniku) wszystkich możliwych prawdopodobieństw, które mogą być potrzebne jako czynniki we wzorze (3). Utożsamiając prawdopodobieństwa z częstościami występującymi w zbiorze uczącym.

2.2 Uzyskanie odpowiedzi z klasyfikatora

W ramach tego ćwiczenia obliczanie odpowiedzi klasyfikatora (w metodach `predict_proba`, `predict`) w wariancie dyskretnym może być realizowane zgodnie ze wzorem

$$y^* = \underset{y \in \{1,\dots,K\}}{\arg\max} \prod_{j=1}^{n} P\left(X_j = x_j \mid Y = y\right) P(Y = y) \tag{1}$$

tj. jako iloczyn prawdopodobieństw (bez zabiegu logarytmowania).

W metodzie `predict_proba` konieczne jest obliczenie prawdopodobieństwa natomiast obliczenia ze wzoru (3) dają wartości wiarygodności (likelihood). Aby z tych wartości uzyskać prawdopodobieństwo należy podzielić obliczone dla danej wartości klasy

$$Prob(Y = y) = \frac{Likelihood(Y = y)}{\sum_{y=1}^{K} Likelihood(Y = y)},$$

gdzie

$$Likelihood(Y = y) = \prod_{j=1}^{n} P\left(X_j = x_j \mid Y = y\right) P(Y = y).$$

2.3 Poprawka LaPlace'a

Przypuśćmy, że w m próbach zaobserwowaliśmy k wystąpień pewnego zdarzenia A dotyczącego zmiennej o q unikalnych wartościach. Szacując prawdopodobieństwo na podstawie częstości, powinniśmy napisać $P(A) \approx k/m$. Stosując poprawkę LaPlace'a, oszacowanie przybiera postać

$$P(A) \approx \frac{k+1}{m+q}. \tag{2}$$

3 Kolejność zadań dla NBC dla zmiennych ciągłych

Jako rozszerzenie opracuj nowy klasyfikator bayesowski realizujący klasyfikację danych z winem w wariancie ciągłym, czyli bez wykonywania dyskretyzacji danych.

1. Napisz klasę reprezentującą naiwny klasyfikator Bayesa w wariancie ze zmiennymi ciągłymi. Przyjmując takie same założenia jak w zadaniu z danymi dyskretnymi.
2. Zastosuj estymaty funkcji gęstości oparte na rozkładach normalnych. W szczególności zaplanuj odpowiednie struktury danych do przechowywania średnich i odchyleń standardowych dla poszczególnych gęstości warunkowych.
3. Porównaj dokładność otrzymanego klasyfikatora z jego dyskretnym odpowiednikiem.
4. Porównaj także zgodność działania otrzymanego klasyfikatora (Twojej implementacji) z gotową implementacją `GaussianNB` dostępną w pakiecie `sklearn.naive_bayes`.

3.1 Uzyskanie odpowiedzi z klasyfikatora

W ramach tego ćwiczenia obliczanie odpowiedzi klasyfikatora (w metodach `predict_proba`, `predict`) może być realizowane zgodnie ze wzorem:

$$y^* = \underset{y \in \{1,\ldots,K\}}{\arg\max} \prod_{j=1}^{n} p_j(x|Y=y) \tag{3}$$

3.2 Funkcje gęstości

Przypuśćmy że wszystkie rozpatrywane zmienne wejściowe są ciągłe, i przypomnijmy notację dla zbioru danych postaci: $D = \big\{(\mathbf{x}_i, y_i)\big\}_{i=1,\ldots,m}$, gdzie $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \in \mathbb{R}^n$ są wektorami cech rzeczywistoliczbowych, zaś y_i etykietami klas. A zatem chcąc przygotować NBC w wariancie ciągłym gaussowskim, musimy wyznaczyć $2 \cdot n \cdot K$ parametrów — oznaczmy je jako μ_{jy} i σ_{jy} (z użyciem pary indeksów) — będących średnimi i odchyleniami standardowymi, dla wszystkich warunkowych rozkładów zmiennych $X_j|Y=y$, gdzie $j = 1,\ldots,n$, $y = 1,\ldots,K$. Oznaczając gęstość takiego wybranego rozkładu jako

$$p_j(x|Y=y) = \frac{1}{\sigma_{jy}\sqrt{2\pi}} e^{-\frac{(x-\mu_{jy})^2}{2\sigma_{jy}^2}}, \tag{4}$$

stosuje się poniższe wzory do wyznaczenia estymat odpowiednio średniej i odchylenia standardowego:

$$\mu_{jy} = \frac{1}{m} \sum_{\substack{i=1\\ y_i=y}}^{m} x_{ij}, \tag{5}$$

$$\sigma_{jy} = \sqrt{\frac{1}{m-1} \sum_{\substack{i=1\\ y_i=y}}^{m} (x_{ij} - \mu_{jy})^2}. \tag{6}$$

Uwaga — czynnik normalizujący $\frac{1}{m-1}$ widoczny w drugim wzorze nie jest pomyłką, a wynika z posłużenia się tzw. *estymatorem nieobciążonym*.

4 Przekazanie zadań

Kod z rozwiązaniem proszę podpiąć w Teams. Proszę w nazwach plików źródłowych zawierać swoje nazwisko celem łatwiejszej identyfikacji.