Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Elektroniczne systemy diagnostyki medycznej i terapii

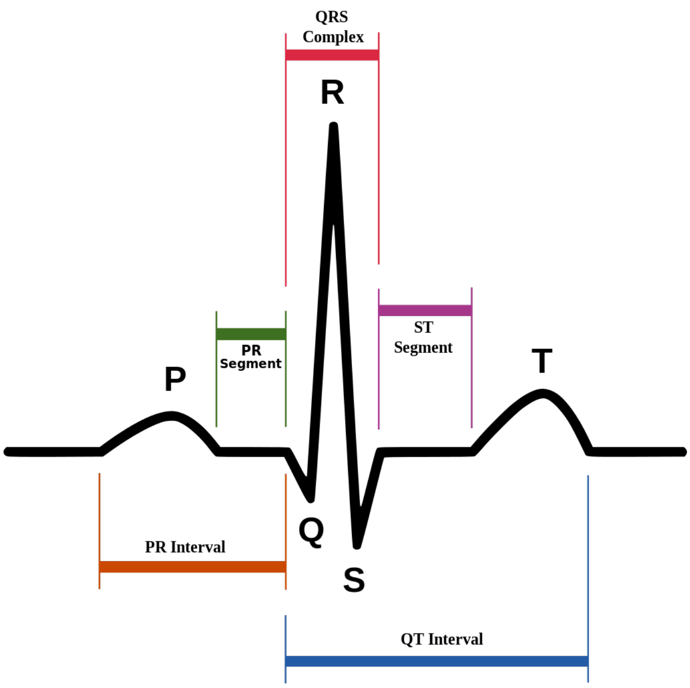
Klasyfikacja pulsu - Naiwny Bayes

Autorzy:
Piotr Janus
Kamil Piszczek

1 Wstęp

Naiwny klasyfikator Bayesa jest prostym klasyfikatorem probabilistycznym opartym na twierdzeniu Bayesa. Nazywany jest naiwnym ze względu na przyjęte założenie, które mówi, że poszczególne cechy są wzajemnie niezależne. Pomimo tak dużego uproszczenia, klasyfikator wypada niespodziewanie dobrze w wielu rzeczywistych problemach. Dużą zaletą tego klasyfikatora jest dobra skalowalność, metoda operuje jedynie na jawnych wzorach w przeciwieństwie do innych metod wykorzystujących podejście iteracyjne.

Jednym z zastosowań klasyfikatora jest diagnozowanie wad i dysfunkcji serca na podstawie sygnału EKG, a dokładniej występującego w nim zespołu QRS. Jest to zespół opisujący pobudzenie mięśni serca. Uproszczony przebieg EKG z zespołem QRS został umieszczony na rysunku 1.



Rysunek 1: Uproszczony zespół QRS - źródło **??**

W celu dokonania klasyfikacje konieczne jest zdefiniowanie wskaźników opisujących QRS. Na podstawie rysunku 1 możemy wyróżnić następujący cechy:

- Wartość szczytowa załamka R i moment jej wystąpienie
- Odstęp pomiędzy wcześniejszym a obecnie analizowanym załamkiem R
- Odstęp pomiędzy aktualnie analizowanym i kolejnym załamkiem R
- Początek/koniec oraz początkowa/końcowa wartość załamka P
- Wartość szczytowa załamka P i moment jej wystąpienia
- Początek/koniec i wartość początkowa/końcowa całego zespołu QRS
- Wartość szczytowa załamka T i moment jego wystąpienia
- Koniec i wartość końcowa załamka T

2 Algorytm

2.1 Założenia

Jak zostało wspomniane we wstępie, Naiwny Klasyfikator Bayesa zakłada wzajemną niezależność poszczególnych cech. Pierwszym etapem jest proces uczenia klasyfikatora. W tym celu definiuje się zbiór uczący zawierający wektory cech (w naszym przypadku są to wektory cech zespołu QRS) oraz odpowiadającą im klasę. Na podstawie zbioru uczącego, dla każdej z klas, tworzony jest rozkład prawdopodobieństwa, służący do klasyfikacji nowych próbek.

2.2 Klasyfikacja

2.3 Zbiór testowy i uczący

3 Opis matematyczny

3.1 Twierdzenie Bayesa

Twierdzenie w teorii prawdopodobieństwa określające zależność między prawdopodobieństwem warunkowym wystąpienia zdarzeń $A|B$ i $B|A$. Przyjmijmy zbiór zdarzeń X, w którym zdarzenia $B_i \in X$, $P(B_i) > 0$ $(i = 1, 2, \dots, n)$ tworzą układ zupełny (iloczyn każdych dwóch zdarzeń jest zdarzeniem niemożliwym, natomiast suma wszystkich zdarzeń jest zdarzeniem pewnym). Wówczas dla dowolnego $A \in X$ zachodzi następująca zależność:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} \tag{1}$$

Wykorzystując dodatkowo wzór na prawdopodobieństwo całkowite, powyższa zależność może zostać przekształcona do następującej postaci:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{k=1}^{n} P(A|B_k)P(B_k)} \tag{2}$$

3.2 Model probabilistyczny

Zdefiniujmy k-elementowy zbiór klas $C = \{C_1, C_2, \dots, C_k\}$ oraz dane do klasyfikacji opisane jako wektor $x = \{x_1, x_2, \dots, x_n\}$ zawierający n niezależnych cech. Prawdopodobieństwo przynależności do danej klasy może zostać zapisane z wykorzystaniem twierdzenia Bayesa (rozdział 3.1):

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)} \tag{3}$$

Prawdopodobieństwo $P(x)$ występujące w mianowniku wzoru (3) nie zależy od C i jest stałe, licznik może natomiast zostać przekształcony poprzez wykorzystanie definicji prawdopodobieństwa warunkowego:

$$P(x_1, \dots, x_n|C_i)P(C_i) = P(x_1, \dots, x_n, C_i) \tag{4}$$

$$P(x_1, \dots, x_n, C_i) = P(x_1|x_2, \dots, x_n, C_i)P(x_2|x_3, \dots, x_n, C_i) \dots P(x_{n-1}|x_n, C_i)P(x_n|C_i)P(C_i) \tag{5}$$

Wykorzystując przyjęte na początku założenie, że cechy $x_1, \dots, x_n$ są niezależne można wyprowadzić następującą zależność:

$$P(x_j|x_{j+1}, \dots, x_n, C_i) = P(x_j|C_i) \quad dla \quad j = \{1, 2, \dots, n-1\} \tag{6}$$

Podstawiając (6) do równania (5) otrzymujemy:

$$P(x_1, \dots, x_n, C_i) = P(C_i) \prod_{j=1}^{n} P(x_j|C_i) \tag{7}$$

Ostatecznie wzór (3) można zapisać w postaci:

$$P(C_i|x) = \frac{P(C_i) \prod_{j=1}^{n} P(x_j|C_i)}{P(x)} \tag{8}$$

3.3 Rozkłady prawdopodobieństwa

3.3.1 Rozkład normalny

$$p(x_i|C_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_{ij})^2}{2\sigma_{ij}^2)}\right), \quad -\infty < x < \infty,\ -\infty < \mu_{ij} < \infty,\ \sigma_{ij} > 0 \tag{9}$$

3.3.2 Rozkład lognormalny

$$p(x_i|C_j) = \frac{1}{x\sigma_{ij}(2\pi)^{1/2}} \exp\left(\frac{-(log(x/m_{ij}))^2}{2\sigma_{ij}^2}\right), \quad 0 < x < \infty,\ m_{ij} > 0,\ \sigma_{ij} > 0 \tag{10}$$

3.3.3 Rozkład Gamma

$$p(x_i|C_j) = \frac{(x/b_{ij})^{c_{ij}-1}}{b_{ij}\Gamma(c_{ij})} \exp\left(\frac{-x}{b_{ij}}\right), \quad 0 \leqslant x < \infty,\ b_{ij} > 0,\ c_{ij} > 0 \tag{11}$$

3.3.4 Rozkład Poissona

$$p(x_i|C_j) = \frac{\lambda_{ij}^x \exp\Big(-\lambda_{ij}\Big)}{x!}, \quad 0 \leqslant x < \infty,\ \lambda_{ij} > 0,\ x = 0, 1, 2, \ldots \tag{12}$$

4 Dodatek A: Instrukcja uruchomienia programów

Repozytorium z projektem dostępne jest pod linkiem: Klasyfikacja pulsu - Naiwny Bayes. Model programowy algorytmu został napisany przy pomocy Pythona. Do skonfigurowania środowiska uruchomieniowego dla projektu służą poniższe instrukcje:

1. Kod programu jest kompatybilny z interpreterem języka Python w wersji 2.7.x i znajduje się w folderze `Model`.
2. Aby włączyć program, należy uruchomić skrypt `main.py`.
3. Wynik działania programu jest przekierowany na standardowe wyjście oraz zapisany do pliku `bayes_logger.txt`.