Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

Elektroniczne systemy diagnostyki medycznej i terapii

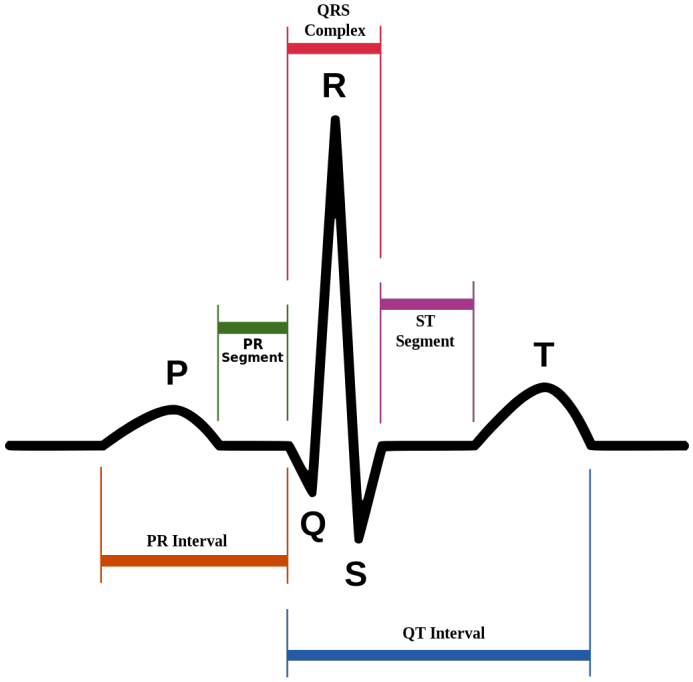
Klasyfikacja pulsu - Naiwny Bayes

Autorzy:
Piotr Janus
Kamil Piszczek

1 Wstęp

Naiwny klasyfikator Bayesa jest prostym klasyfikatorem probabilistycznym opartym na twierdzeniu Bayesa. Jest nazywany naiwnym ze względu na przyjęte założenie, mówiące, że poszczególne cechy są wzajemnie niezależne. Pomimo tak dużego uproszczenia, klasyfikator wypada niespodziewanie dobrze w wielu rzeczywistych problemach. Jego dużą zaletą jest dobra skalowalność, operuje on 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 [1]

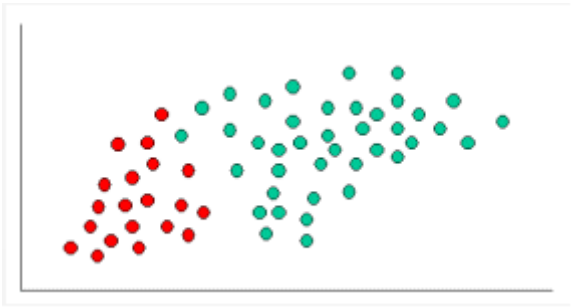
W celu dokonania klasyfikacje konieczne jest zdefiniowanie wskaźników opisujących QRS. Na podstawie rysunku 1 możemy wyróżnić następujący cechy:

- 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

Idee metody Naiwnego Bayesa można łatwo wyjaśnić na prostym przykładzie, precyzyjny opis matematyczny został zamieszczony w rozdziale 3. Na rysunku 2 przedstawiony został zbiór punktów, podzielony na dwie klasy (czerwone i zielone). Zadaniem klasyfikatora jest przydzielenie nowego obiektu do jednej z tych klas. W tym przypadku klasyfikacja będzie dokonana na podstawie położenia i elementów znajdujących się w sąsiedztwie nowo dodanego obiektu. Podany zbiór punktów, pełni w tym przypadku rolę zbioru uczącego.

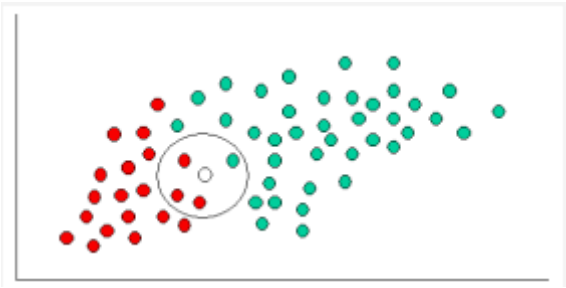


Rysunek 2: Prosty przykład – zbiór punktów (źródło [3])

2.2 Klasyfikacja

W omawianym przykładzie możemy zauważyć, że obiektów zielonych jest dwa razy więcej niż czerwonych. W związku z tym możemy założyć "z góry", że nowy obiekt ma dwa razy większe prawdopodobieństwo bycia zielonym niż czerwonym. Obliczone w ten sposób prawdopodobieństwo nazywane jest prawdopodobieństwem *a priori*.

Wszystkich obiektów jest 60 w czym 40 zielonych i 20 czerwonych. Prawdopodobieństwo *a priori* wylicza się jako iloraz liczby obiektów danego koloru do liczby wszystkich obiektów. Następnie przystępujemy do kolejnego etapu klasyfikacji, przyjmijmy pewne sąsiedztwo nowego punktu (rysunek 3). Możemy założyć, że im więcej obiektów danego koloru w otoczeniu nowego obiektu, tym bardziej prawdopodobne, że jest on tego koloru. Wyznaczone w ten sposób prawdopodobieństwo nazywane jest szansą, oblicza się je jako stosunek liczby obiektów danego koloru w sąsiedztwie, do całkowitej liczby obiektów tego koloru.



Rysunek 3: Prosty przykład – sąsiedztwo (źródło [3])

Mając dane prawdopodobieństwo *a priori* oraz szansę, możemy przystąpić do ostatniego etapu klasyfikacji. Końcowe prawdopodobieństwo czy nowy obiekt należy do danej klasy (jest

danego koloru) obliczane jest jako iloczyn dwóch wyznaczonych wcześniej prawdopodobieństw. Obiekt zostaje oczywiście przypisany do klasy o większym prawdopodobieństwie.

2.3 Wykorzystanie w detekcji pulsu

Omówiony w poprzednich punktach przykład, był bardzo uproszczony i miał na celu jedynie pokazać idee klasyfikatora. Klasyfikacja zespołu QRS jest problemem wielowymiarowym (dokładnie 18 wymiarowym, gdyż taka jest długość wektora cech). W takim przypadku nieco inaczej oblicza "szansa". Osobno wyznaczane jest prawdopodobieństwo przynależności do danej klasy na podstawie każdego elementu z wektora cech. Ostatecznie pod uwagę brany jest iloczyn wszystkich prawdopodobieństw.

Kolejnym zagadnieniem jest sposób obliczania prawdopodobieństwa na podstawie cechy. W przypadku niniejszego projektu wykorzystywany jest w tym celu rozkład normalny (rozdział 3.3). Dla każdej klasy, poszczególne cechy mają przypisaną wartość średnią i odchylenie standardowe. Są one obliczane na podstawie zbioru uczącego w trakcie procesu uczenia.

2.4 Zbiór testowy i uczący

Metoda Naiwnego Bayesa należy do grupy algorytmów nadzorowanego uczenia maszynowego (z nauczycielem). Dane, które podlegają klasyfikacji za pomocą tej metody należy podzielić na dwa zbiory: uczący oraz testowy. Pierwszy z nich wykorzystywany jest w procesie uczenia klasyfikatora. Drugi weryfikuje, czy nauczony klasyfikator ma zdolność generalizacji, a także sprawdza efektywną skuteczność klasyfikacji. Warto zaznaczyć, że algorytmy nadzorowanego uczenia maszynowego mogą stracić zdolność do generalizacji w wyniku zjawiska overfittingu. Następuje ono wtedy, gdy algorytm nadmiernie dostosowuję się do zbioru uczącego - tzn. błąd zbioru uczącego maleje, a błąd zbioru testowego zaczyna rosnąć.

Istnieje wiele sposobów podziału danych na zbiór uczący i testowy. Generalnie metody te nazywane są sprawdzianem krzyżowym [2]. Poniżej opisane zostały najbardziej znane metody.

Pierwszy ze sposobów to metoda Prostej Walidacji. Jest ona obecnie używana w niniejszym projekcie. Metoda ta polega na losowym podziale danych na dwa rozłączne zbiory: uczący i testowy. W tym wariancie zbiór testowy stanowi co najwyżej 1/3 całkowitej próbki danych. Dodatkowo, zastosowana implementacja gwarantuje, że proporcje ilościowe klas w obu zbiorach będą takie same jak w początkowym zbiorze (przed podziałem).

Drugi ze sposobów to K-krotna Walidacja. W tym wariancie dane dzielone są na K różnych podzbiorów. Następnie kolejno każdy z nich bierze się jako zbiór testowy, a pozostałe podzbiory stanowią zbiór uczący. Cały proces powtarzany jest K razy. Otrzymane rezultaty łączone są w jeden za pomocą wybranego sposobu. Najczęściej są one po prostu uśredniane.

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 $B_i \in X$, $P(B_i) > 0$ $(i = 1, 2, \ldots, n)$ 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 $C = \{C_1, C_2, \ldots, C_k\}$ oraz dane do klasyfikacji opisane jako wektor $x = \{x_1, x_2, \ldots, x_n\}$ 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:

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

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

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

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

Podstawiając (6) do równania (5) otrzymujemy:

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

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

Występujące w równaniu (8) prawdopodobieństwo, że dany obiekt należy do klasy j na podstawie cechy i może zostać wyznaczone na różne sposoby, w zależności od analizowanego problemu. Jak zostało napisane w rozdziale 2.3, w niniejszym projekcie wykorzystywany jest rozkład normalny. Warto pamiętać, że istnieje także możliwość wykorzystania innych funkcji gęstości prawdopodobieństwa.

3.3.1 Rozkład normalny

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

Gdzie: μ_{ij} – średnia, σ_{ij} – odchylenie standardowe

3.3.2 Rozkład lognormalny

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

Gdzie: m_{ij} – parametr skali, σ_{ij} – parametr kształtu

3.3.3 Rozkład Gamma

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

Gdzie: b_{ij} – parametr skali, c_{ij} – parametr kształtu

3.3.4 Rozkład Poissona

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

Gdzie: λ_{ij} – średnia

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`.

Literatura

[1] Wikipedia, hasło: *QRS complex*, https://en.wikipedia.org/wiki/QRS_complex (ostatni dostęp 15.12.2016)

[2] Wikipedia, hasło: *Sprawdzian krzyżowy*, https://pl.wikipedia.org/wiki/Sprawdzian_krzy%C5%BCowy (ostatni dostęp 15.12.2016)

[3] Internetowy Podręcznik Statystyki, rozdział: Naiwny klasyfikator Bayesa, http://www.statsoft.pl/textbook/stathome.html (ostatni dostęp 15.12.2016)