

Laboratorium Rozpoznawania Obrazów – Ćwiczenie #2

Klasyfikacja optymalna Bayesa

Termin oddawania: **12.03.2018, 15.03.2018, 19.03.2018, 22.03.2018**

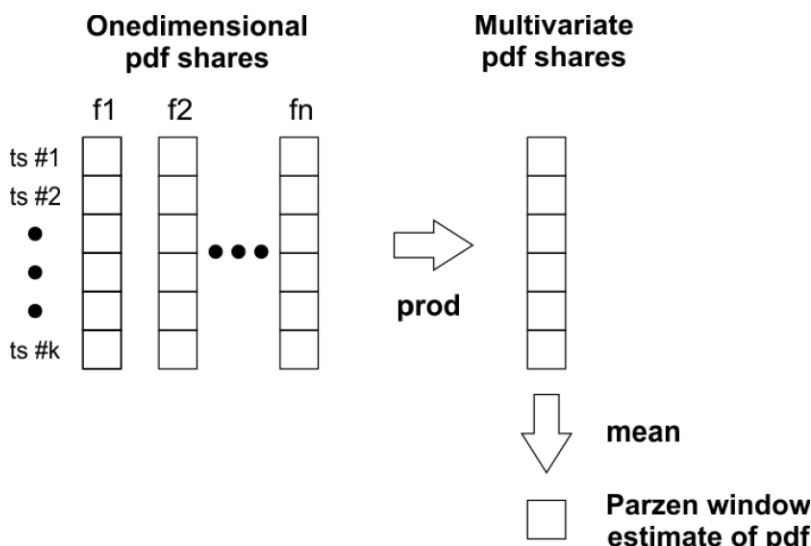
W tym ćwiczeniu Państwa zadaniem będzie przyjrzeć się klasyfikacji Bayesa, przy różnych metodach liczenia funkcji gęstości prawdopodobieństwa rozkładów warunkowych dla poszczególnych klas. Do porównania są trzy metody wyznaczania tej gęstości:

1. Przy założeniu, że cechy są niezależne, a rozkłady każdej cechy są normalne (w tym przypadku gęstość prawdopodobieństwa dla więcej niż jednej cechy jest liczona jako iloczyn gęstości dla poszczególnych cech).
2. Przy założeniu, że mamy do czynienia z wielowymiarowym rozkładem normalnym dla cech używanych do klasyfikacji.
3. Przy użyciu okna Parzena do wyznaczenia aproksymacji gęstości prawdopodobieństwa na podstawie zbioru uczącego.

Dane, które będziecie Państwo klasyfikowali to obrazy maści kart, reprezentowane przez niezmienniki momentowe Hu (http://en.wikipedia.org/wiki/Image_moment). Pierwszą kolumnę danych stanowią etykiety „finalne” (4 – pik, 3 – kier, 2 – karo, 1 – trefl). Ponieważ obrazy były drukowane na różnych urządzeniach, dla potrzeb klasyfikacji będziemy używać 8 klas.

Pierwszym krokiem do wykonania jest zaimplementowanie funkcji pdf (*probability density function*) i funkcji, które liczą parametry rozkładów dla stosownych funkcji pdf. W tym celu dobrze posłużyć się danymi z pliku pdf_test.txt (2 klasy i tylko 20 próbek w dwóch wymiarach).

Pewnego komentarza wymaga liczenie pdf z wykorzystaniem okna Parzena. Wartość gęstości „składamy” tutaj licząc udziały próbek ze zbioru uczącego w punkcie, dla którego mamy policzyć gęstość prawdopodobieństwa. Nie ma tu drogi na skróty: dla każdej próbki x trzeba policzyć jednowymiarowe pdf dla każdej cechy (tu mamy liczba_probek_w_klasie * liczba_cech wartości), a następnie właściwie zagregować:



Całkiem rozsądnym wyborem funkcji okna $\phi(u)$ jest rozkład normalny. Wartość konkretnej cechy punktu x podajemy jako wartość średnią, natomiast szerokość okna Parzena h_1 , dostosowaną do liczby próbek w klasie: $h_n = \frac{h_1}{\sqrt{n}}$ traktujemy jako odchylenie standardowe rozkładu.

Po uruchomieniu funkcji wyznaczających parametry oraz liczące gęstość rozkładu prawdopodobieństwa można zająć się analizą danych maści kart:

1. Ponieważ dane pochodzą z dwóch różnych populacji, dla potrzeb klasyfikacji będzie używanych 8 klas – w pliku mainscript.m jest już kod zmieniający odpowiednio etykiety.

2. Proszę sprawdzić dane, a szczególnie zbiór uczący. Wartości odstające w tym zbiorze mogą mieć opłakane skutki dla jakości klasyfikacji (prowadzący znaleźli dwie wartości odstające). Informacja o tym, które próbki zostały usunięte, musi znaleźć się w sprawozdaniu.
3. Proszę wybrać dwie cechy i zbudować dla nich klasyfikator optymalny Bayesa, wyznaczając gęstość prawdopodobieństwa zgodnie z punktami 1-3 na poprzedniej stronie. Prawdopodobieństwa *a priori* przyjąć równe 0.25.
To czy traficie Państwo w cechy najlepsze, nie jest szczególnie istotne, ale warto przy wstępnej analizie danych zwrócić uwagę na „potencjał” klasyfikacyjny poszczególnych cech i wybrać dwie najbardziej obiecujące.
4. Proszę sprawdzić, jaki wpływ na klasyfikację zbioru testowego, ma dobór próbek w zbiorze uczącym (np. wzięcie $1/10$, $1/4$, $1/2$ i całego zbioru uczącego).
Uwaga: stosowną część próbek ze zbioru uczącego należy wylosować niezależnie dla poszczególnych klas; ponieważ wprowadzamy element losowy, to eksperyment trzeba powtórzyć (minimum 5 razy) i podać uśrednione wyniki (prócz tego warto obejrzeć wartości minimalne, maksymalne i odchylenie standardowe)
Tu powinniście Państwo zaimplementować funkcję `reduce`, która zostawia stosowną część poszczególnych klas. W tym punkcie redukcja dotyczy jedynie zbioru uczącego.
5. Proszę sprawdzić, jaki wpływ na klasyfikację zbioru testowego, ma dobór parametru h_1 (to oczywiście tylko w przypadku klasyfikatora z oknem Parzena).
6. Jak zmieniają się wyniki klasyfikacji jeśli prawdopodobieństwo *a priori* będzie dwukrotnie większe dla maści czerwonych (0.085, 0.165, 0.165, 0.085, 0.085, 0.165, 0.165, 0.085)? Stosowną redukcję wykonujecie tutaj **tylko** na zbiorze testowym. Uwzględnijcie też uwagę z punktu czwartego.
7. Jak mają się wyniki klasyfikatorów Bayesa, do klasyfikatora 1-NN z pierwszego ćwiczenia? (Chodzi oczywiście o to, żeby uruchomić klasyfikator 1-NN na danych kart. Przy okazji musicie Państwo rozstrzygnąć, czy dane kart należy dla tego klasyfikatora normalizować, czy nie.)
Ponieważ mamy tu przyzwoitej wielkości zbiór uczący i zbiór testowy, należy sklasyfikować zbiór testowy funkcją `c1s1nn` i obliczyć współczynnik błędu.

Uwaga:

Oczekujemy sprawozdania na piśmie – zwięzłego, ale zawierającego najważniejsze informacje, w szczególności wyniki eksperymentów. Do tekstu sprawozdania trzeba dołączyć kod Octave użyty w ćwiczeniu.

Zdecydowanie nie chcemy dostawać danych: ani uczących, ani testowych. Za przysłanie z rozwiązaniem danych będziemy odejmować 1 punkt.

Parę uwag, które mamy nadzieję, mogą pomóc w realizacji ćwiczenia:

1. Klasyfikator (`bayesc1s`) jest już zaimplementowany.
2. W klasyfikacji używamy 8 klas (taka uroda danych), ale nasz klient jest zainteresowany tylko etykietami maści. Prócz jakości klasyfikacji dla 8 klas, proszę podać jakość po powrocie do czterech etykiet „klienta”. W których punktach z listy powyżej warto to zrobić?
3. Zbiór **testowy** powinien być zgodny z założonymi prawdopodobieństwami *a priori*. Dla równych, liczba próbek wszystkich klas w zbiorze testowym powinna być równa (tak jest). Kiedy prawdopodobieństwa *a priori* klas są różne, trzeba zapewnić, żeby w zbiorze testowym było dwa razy więcej znaków czerwonych (kar i kierów) niż czarnych (pików i trefli). Przyda się tutaj funkcja `randperm` (wywołana z parametrem N daje wektor losowo permutowanych wartości z zakresu $1..N$). Można jej użyć do losowego odrzucenia stosownej części klas „czarnych”.
4. Sporo pożytecznych informacji jest w plikach skryptów stanowiących część tego pakietu.