Laboratorium Rozpoznawania Obrazów – Ćwiczenie #2
Klasyfikacja optymalna Bayesa

Termin oddawania: **16.03.2015, 19.03.2015**

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.

Dość często wykorzystywaną metodą aproksymacji nieznanego rozkładu prawdopodobieństwa jest wykorzystanie okna Parzena. Polega ona na tym, że nieznany rozkład gęstości „budujemy" z punktów zbioru uczącego, składając gęstości cząstkowe, których źródłem są punkty zbioru uczącego.
Gęstości cząstkowe dostarcza nam funkcja, nazywana funkcją okna $\varphi(u)$. Nie ma specjalnych ograniczeń na funkcję okna, poza jednym: w całej dziedzinie musi całkować się do 1 (w końcu przy jednej próbce w zbiorze uczącym, to będzie funkcja gęstości prawdopodobieństwa).
Dla potrzeb tego ćwiczenia, przyjmiemy funkcję okna postaci:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

Przy liczeniu gęstości jest używany parametr h_1, nazywany szerokością okna (powiedzmy, że dla potrzeb ćwiczenia ograniczymy jego zakres do przedziału <0.0001, 0.01>).
Gęstość prawdopodobieństwa ze zbioru uczącego, zawierającego *n* punktów, będziemy liczyć z następującego wzoru:

$$h_n = \frac{h_1}{\sqrt{n}}$$

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n} \varphi\left(\frac{x - x_i}{h_n}\right)$$

Zbiór uczący dla tego zadania zawiera niezmienniki momentowe Hu (http://en.wikipedia.org/wiki/Image_moment) zeskanowanych obrazów maści kart. Pierwszą kolumną jest identyfikator klasy (4 – pik, 3 – kier, 2 – karo, 1 – trefl). Warto zwrócić uwagę, że znaki były drukowane różnymi technikami i na połowie obrazów widać raster drukarski, a na połowie nie. W zbiorze uczącym rozmiar czcionki oraz kąt obrotu były zmieniane systematycznie. W zbiorze testowym rozmiar czcionki może przyjmować dowolne wartości z pewnego przedziału, natomiast kąt obrotu znaków jest dowolny.

Materiały przygotowano dzięki wsparciu udzielonemu z funduszy norweskich i funduszy EOG, pochodzących z Islandii, Liechtensteinu i Norwegii oraz środków krajowych

Państwa zadaniem jest skonstruowanie kilku klasyfikatorów, przy różnych założeniach. Spodziewam się, że najlepsze z klasyfikatorów będą mieć zupełnie przyzwoitą jakość rozpoznawania (> 90% na całym zbiorze testowym). Plusem będzie „odchudzenie" zbioru uczącego, tzn. użycie tylko części próbek z całego zbioru. Ten punkt tyczy jedynie klasyfikacji z wykorzystaniem okna Parzena – w pozostałych dwóch i tak liczycie Państwo jedynie parametry rozkładu.

Teraz konkrety:

1. Proszę sprawdzić dane, a szczególnie zbiór uczący. Wartości odstające w tym zbiorze mogą mieć opłakane skutki w jakości klasyfikacji.
2. 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.
3. Proszę sprawdzić, jaki wpływ na klasyfikację zbioru testowego, ma dobór próbek w zbiorze uczącym (np. wzięcie 1/10, ¼, ½ i całego zbioru uczącego).
4. Proszę sprawdzić, jaki wpływ na klasyfikację zbioru testowego, ma dobór parametru h_1 (to oczywiście tylko w przypadku klasyfikatora z oknem Parzena).
5. Jak zmienią się wyniki klasyfikacji jeśli prawdopodobieństwo *a priori* będzie dwukrotnie większe dla maści czerwonych (0.17, 0.33, 0.33, 0.17)?
6. Jak mają się wyniki klasyfikatorów Bayesa, do klasyfikatora 1-NN z pierwszego ćwiczenia? (Chodzi oczywiście o to, żeby uruchomić klasyfikatora 1-NN na danych kart. Przy okazji musicie Państwo rozstrzygnąć, czy dane kart należy dla tego klasyfikatora normalizować, czy nie.)

Uwaga: Oczekuję sprawozdania na piśmie - zwięzłego, ale zawierającego najważniejsze informacje. Do tekstu sprawozdania trzeba dołączyć kod Octave użyty w ćwiczeniu (ale w oddzielnych plikach .m).

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

1. Proszę uważnie popatrzeć na równania na poprzedniej stronie i porównać je z gęstością prawdopodobieństwa rozkładu normalnego (przyda się funkcja normpdf).
2. Funkcje, które przygotujecie powinny nadawać się zarówno do użycia w środowisku laboratoryjnym, jak i w rzeczywistym działaniu. To oznacza, że tak dużo, jak się da należy policzyć w fazie konstruowania klasyfikatora (dotyczy to szczególnie pierwszych dwóch klasyfikatorów; przy klasyfikatorze Parzena zbyt dużo zoptymalizować się nie da).
3. Zbiór **testowy** powinien być zgodny z założonym 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".

Materiały przygotowano dzięki wsparciu udzielonemu z funduszy norweskich i funduszy EOG, pochodzących z Islandii, Liechtensteinu i Norwegii oraz środków krajowych