Klasyfikacja w oparciu o metrykę budowaną poprzez dystrybuanty empiryczne na przestrzeni wzorców uczących

Cezary Dendek
Wydział Matematyki i Nauk Informacyjnych
PW

Plan prezentacji

- Plan prezentacji
- Wprowadzenie
- Własności transformacji przestrzeni poprzez dystrybuanty
- Metryka w oparciu o transformację i metrykę Mahalanobisa
- Klasyfikacja w oparciu o metrykę

Wprowadzenie

- Cel: transformacja przestrzeni obserwacji
  Dlaczego?
  - W naturalny sposób używamy przestrzeni R^n
  - Przestrzeń ta oddaje naturalną strukturę geometryczną modelowanego zjawiska
  - ale nie oddaje (w naturalny sposób) jego struktury probabilistycznej
- Jak?
  Przekształcając przestrzeń poprzez dystrybuantę.

Własności transformacji

- X – zmienna losowa w R^1
- *F* – dystrybuanta zmiennej losowej X
- F ciągła => zmienna losowa F(X) ma rozkład U[0,1]
  (por. metoda generowania próbki i.i.d. o zadanym rozkładzie z rozkładu U[0,1] poprzez odwrócenie dystrybuanty)

Własności transformacji

- Efekty:
  - Unormowanie przestrzeni (przedział [0,1])
  - Znana wartość oczekiwana (0.5)
  - Pomimo tych zysków praca z oryginalnym modelem probabilistycznym przestrzeni (*a nie z geometrią R^n*)
- Uwaga (istotna)
  Dystrybuanta może być postrzegana jako mapa przestrzeni probabilistycznej w punkcie E(X)

Własności transformacji

- Geometria zbioru, na którym jest określona X (X(\Omega)) może być odwzorowana przez rozkład U(X(\Omega))
- Jeśli X jest wielowymiarowe przyjmujemy
  F(x) = [F_1(x_1) ... F_n(x_n)]'
  (jeśli X_1...X_n są niezależne, to formalnie
  F(x)=F_1(x_1) * ...* F_n(x_n)
  ale zakładamy, że nic o zależności nie wiemy)

Transformacja próbki

- Dystrybuanta empiryczna X w R na podstawie próbki P={X_1, ... , X_n}:

$$F(x)= \#\{X_i \mid X_i<=x\}/\#P$$

- Najprostszy przypadek (można też estymować gęstość przy pomocy estymatora jądrowego i całkować; trudniej i tylko dla rozkładów ciągłych)

Transformacja próbki

- Przypadek wielowymiarowy:
  w i-tym wymiarze liczymy F_i(x_i) i budujemy F(x)
- Transformacja próbki poprzez przekształcenie każdego elementu przy pomocy F(x)
- Opisana transformacja często poprawia działanie klasyfikatorów

Transformacja próbki

- **PROPOZYCJA**
  Można łączyć różne dystrybuanty na danej przestrzeni:

  $$F_w(x)=\mu F_{prob}(x)+(1-\mu)F_{geom}(x)$$

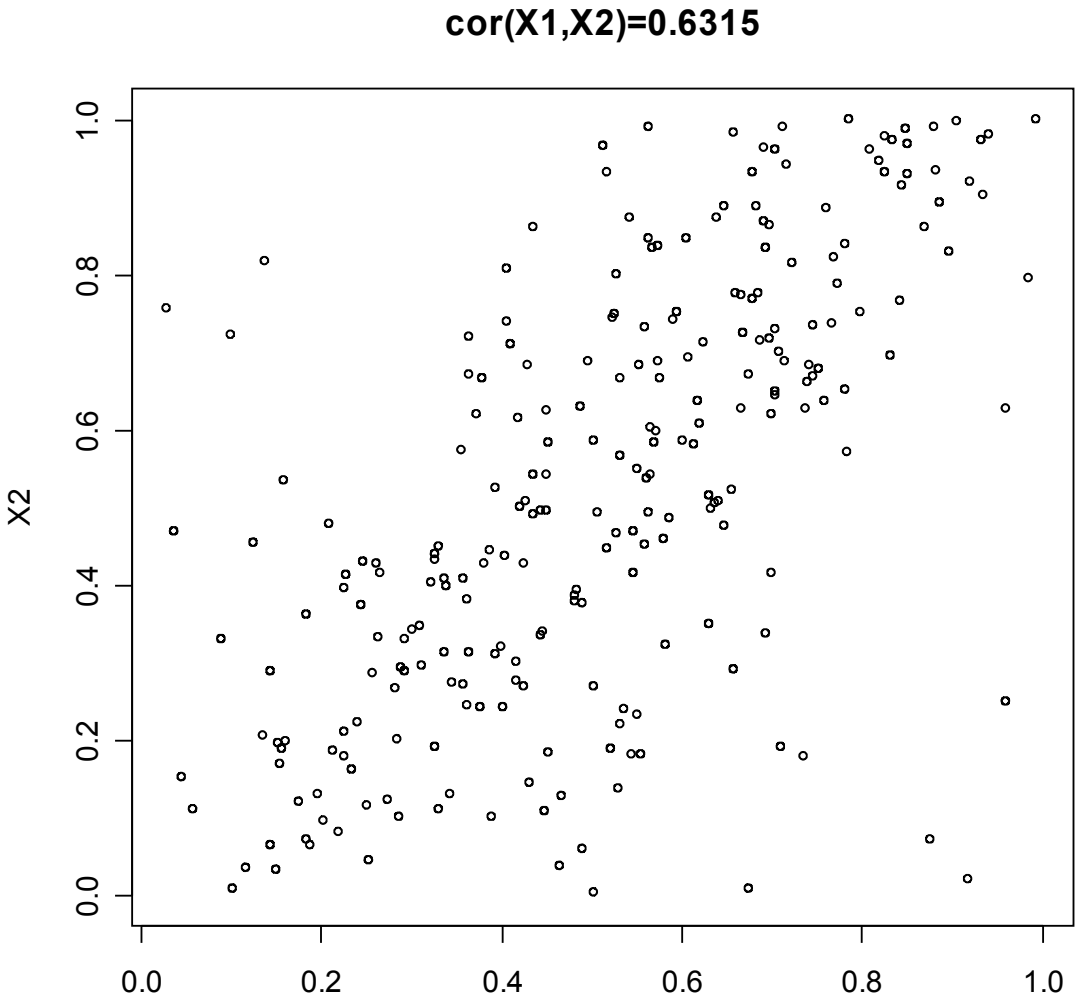
  (ogólnie: średnia ważona m dystrybuant)
- Umożliwia to dostosowywanie transformacji celem poprawy jakości klasyfikacji
- Czy przekształcenie istotnie zmienia przestrzeń?

  $$cor(F_w(X), X)$$

Transformacja przestrzeni

- Jak zdefiniować odległość w przestrzeni wynikowej?
  - Można potraktować ją znów jako R^n ...
    ... ale nie uwzględnia się korelacji poszczególnych wymiarów, co wpływa na otoczenie danego zdarzenia

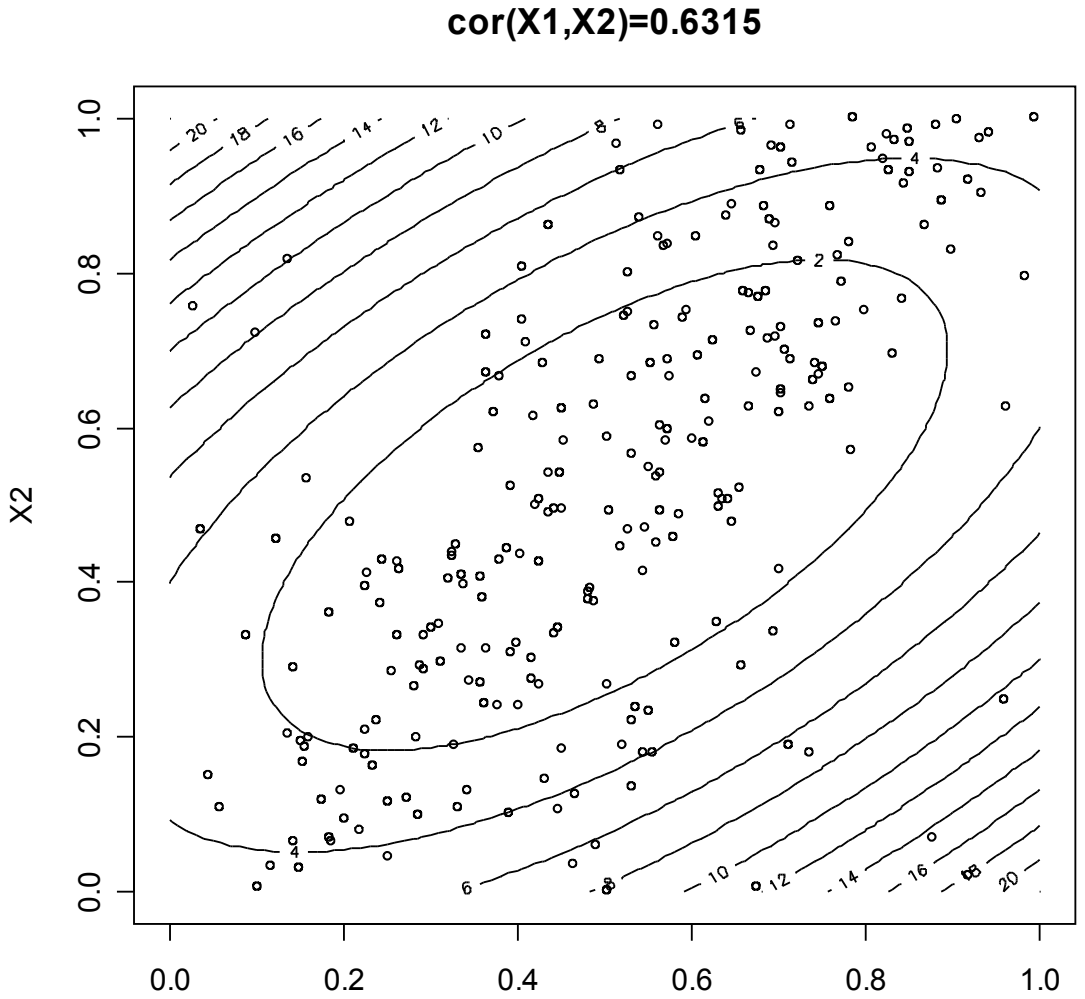
Odległość w przestrzeni wynikowej dlaczego korelacja?



Odległość w przestrzeni wynikowej

- **PROPOZYCJA**
  Wprowadzona zostanie poprzez zastosowanie metryki Mahalanobisa:
  - ∑ = Cov(F_1,..., F_n)
  - (d(A,B))^2 = (F(A)-F(B))' * (∑)^(-1) * (F(A)-F(B))
  - (∑ dod. okr => forma kwadratowa ∑ dod. okr.)
- Zauważmy, że ta metryka jest zlokalizowana w punkcie E(X) (budowana w oparciu o mapę w E(X))

Odległość w przestrzeni wynikowej



Odległość w przestrzeni wynikowej

- Otoczenia są teraz elipsami
- Najlepiej przybliżone (w sensie probabilistycznym) otoczenie to otoczenie E(X)
- Co wyraża odległość?
  - F – dystrybuanta empiryczna; A, B - obserwacje
  - |F_i(A) - F_i(B)| wyraża p-wartość hipotezy, że... A=B
  - Wykorzystanie korelacji pozwala na uwzględnienie zależności pomiędzy zmiennymi
  - Być może warto zastosować normowanie (∑)^(-1), aby uzyskać normalizację odległości

Odległość w przestrzeni wynikowej

- Jak można lokalizować metrykę w innych punktach?
- **PROPOZYCJA**
  Inaczej obliczając kowariancję:
  $\sum_{\mathbf{A}} = Cov_{\mathbf{A}}(F_1 \ldots F_n) =$
  $\{ E\,(F_i(X)-F_i(\mathbf{A}))*(F_k(X)-F_k(\mathbf{A})) \}$
- Uzyskujemy wtedy „spojrzenie” na zbiór z innego punktu (punktu tożsamego ze zdarzeniem **A**) przestrzeni probabilistycznej

Odległość w przestrzeni wynikowej

- Interpretacja innej lokalizacji metryki:
  W przypadku, gdy
  - przechodząc przez ulicę
  - złamiemy nogę
  - i przejedzie zielony samochód marki X

  Będziemy ciągle (∑=rozsądek) blisko zdarzenia:
  - przechodząc przez ulicę
  - łamiemy nogę
  - i przejeżdza czerwony samochód marki Y

Odległość w przestrzeni wynikowej

- Łatwo to zrozumieć – wystarczy spojrzeć na świat (przestrzeń możliwych zdarzeń)
  oczami (z perspektywy)
  osoby ze złamaną na ulicy nogą (zdarzenie)
- Oczywiście z perspektywy świadka lub np. samochodu (w naturalny sposób zanurzanych w przestrzeń zdarzeń) istotność koloru (której wyrazem jest odległość w przestrzeni zdarzeń) prezentuje się w zupełnie inny sposób

Odległość w przestrzeni wynikowej

- **PROPOZYCJA**
  Uniezależniając się od E(X) metrykę można definiować poprzez formę:

  $$d(A, B)^2=(d_A(A, B)^2 + d_B(A, B)^2)/2$$

  czyli korzystając z 2 zlokalizowanych metryk

  Ma to dość prosty matematyczny wyraz formy odpowiadającej $((\Sigma_A)^{-1} + (\Sigma_B)^{-1})/2$. Widać, że konieczne jest tu stosowanie normowania!

Odległość w przestrzeni wynikowej

- Odległość można też wprowadzić (**PROPOZYCJA**) w przypadku, gdy
  - w niektórych wymiarach brakuje danych
  - celem jest obliczenie obejmujące grupę elementów (uogólnienie odległości Mahalanobisa od zbioru danych)

  Poprzez zastąpienie brakujących wartości warunkowymi wartościami oczekiwanymi.

Odległość w przestrzeni wynikowej

- **PROPOZYCJA**
  Lokalizacja metryki
  - poprzez obliczanie macierzy $\sum$ jedynie na (wystarczająco dużym) otoczeniu punktu, w którym ją lokalizujemy
  - obliczanie empirycznych dystrybuant tylko na tym otoczeniu

Klasyfikacja w oparciu o metrykę

- Model k-NN
  - Ile spośród *k* obserwacji najbliższych klasyfikowanemu punktowi należy do danej klasy?
- Model Mahalanobisa
  - Dla każdej z klas obliczyć estymator wartości oczekiwanej obserwacji, które do niej należą
  - Nieznany punkt klasyfikować na podstawie odległości od centrów każdej z klas (posługując się metryką Mahalanobisa z $\sum$ obliczoną dla każdej z klas osobno

Dziękuję za uwagę

- Pytania?
- Komentarze?
- Uwagi?