Budowa modeli klasyfikacyjnych w oparciu o funkcję odległości w przestrzeni zdarzeń i uogólnienie pojęć probabilistycznych na przestrzenie metryczne

C. Dendek prof nzw. dr hab. J. Mańdziuk

Politechnika Warszawska,
Wydział Matematyki i Nauk Informacyjnych

Outline

Podstawowe pojęcia probabilistyczne

Podstawowe pojęcia probabilistyczne używane w budowie modeli

- wartość oczekiwana

$$E[X] = \int x dF(x)$$

- wariancja

$$Var[X] = \int (x - E[X])^2 dF(x)$$

- kowariancja

$$Cov[X, Y] = \int (x - E[X])(y - E[Y]) dF(x, y)$$

Konieczność uogólnienia

Cechy przedstawionych pojęć

- naturalnie osadzone w R^n
- konceptualnie wywodzą się z R^n
- wykorzystują wektorowość R^n (odejmowanie)
- można je stosować do przestrzeni metrycznych...
  ...po ich zanurzeniu w R^n

Uogólnienia pojęć

Aparat pojęciowy

W przestrzeni metrycznej do definiowania pojęć wykorzystać można jedynie *metrykę* $d(x, y)$

Wariancja

- w R jest to wartość minimalizująca funkcjonał

$$V(z) := \int (x - z)^2 dF(x)$$

$$Var[X] = \mathbf{min}_{z \in R} V(z)$$

- można ją wykorzystać do znalezienia wartości oczekiwanej!

Uogólnienia pojęć

Wariancja

Funkcjonał

$$V(z) := \int (x - z)^2 dF(x)$$

można zapisać jako

$$V(z) := \int d^2(x, z) dF(x)$$

Uogólnienia pojęć

Uogólniona wartość oczekiwana

- definicja
$$E_d[X] = \mathbf{argmin}_{z \in S} V(z)$$
- nie musi być jednoznacznie wyznaczona!
- można kontrolować zbiór, z którego pochodzi $E_d[X]$
- wariancja pozwala na wnioskowanie o skupieniu elementów wokół $E_d[X]$

Uogólniona kowariancja

Cechy kowariancji

- przechowuje eliptyczne przybliżenie rozkładu wokół wartości oczekiwanej
- Cov[X,Y] pozwala na przybliżenie Y przy znanym X

Uogólniona kowariancja

Kowariancja w przestrzeni z liniowym porządkiem

- definicje

$$Ord_d(x, y) = d(x, y)Ord(x, y)$$

$$Cov[X, Y] = \int Ord_d(X, E_d[X])Ord_d(Y, E_d[Y])dF(x, y)$$

- Cov[X,Y] pozwala na przybliżenie $Ord_d(Y, E_d[Y])$ przy znanym X
- zawiera informację o ”stronie”

Kowariancja w oparciu o punkty próbne

Idea

Wprowadzenie punktów próbnych $\{p_1, ...p_n\} \in S$, pomiar odległości od tych punktów i ich korelacji.
Dzięki temu można dokładniej przybliżać rozkład i uzyskiwać precyzyjniejsze estymaty ”przynależności” do obszaru

Prace

Prace trwają....

Klasyfikator odległościowy

Klasyfikator odległościowy oparty na odległości Mahalanobisa

- określenie odległości probabilistycznej na każdym wymiarze
- dla każdej klasy:
  - znalezienie wartości oczekiwanej
  - estymacja uogólnionej kowariancji
  - stworzenie metryki wewnątrzklasowej
  - określenie funkcji dyskrymiacyjnej

$$f(x) = Ord_d(x, E_d) Cov^{-1} Ord_d(x, E_d)$$

- punkt klasyfikowany jest do klasy o najmniejszej wartości funkcji dyskryminacyjnej

Nieoficjalne wyniki

Poprawa (nawet względem klasyfikacji k–NN)

- Pima
- BUPA

... najlepsze wyniki dla klasycznej różnicy dystrybuant.

Dziękuję za uwagę

Dziękuję za uwagę