Metryka probabilistyczna, jej estymacja i zastosowanie do binarnej klasyfikacji k–NN

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

Politechnika Warszawska,
Wydział Matematyki i Nauk Informacyjnych

Abstrakt

Główny cel pracy

Poprawa binarnej klasyfikacji odległościowej (model *k–NN*) poprzez stworzenie modelu miary odległości opartej na przestrzeni probabilistycznej.

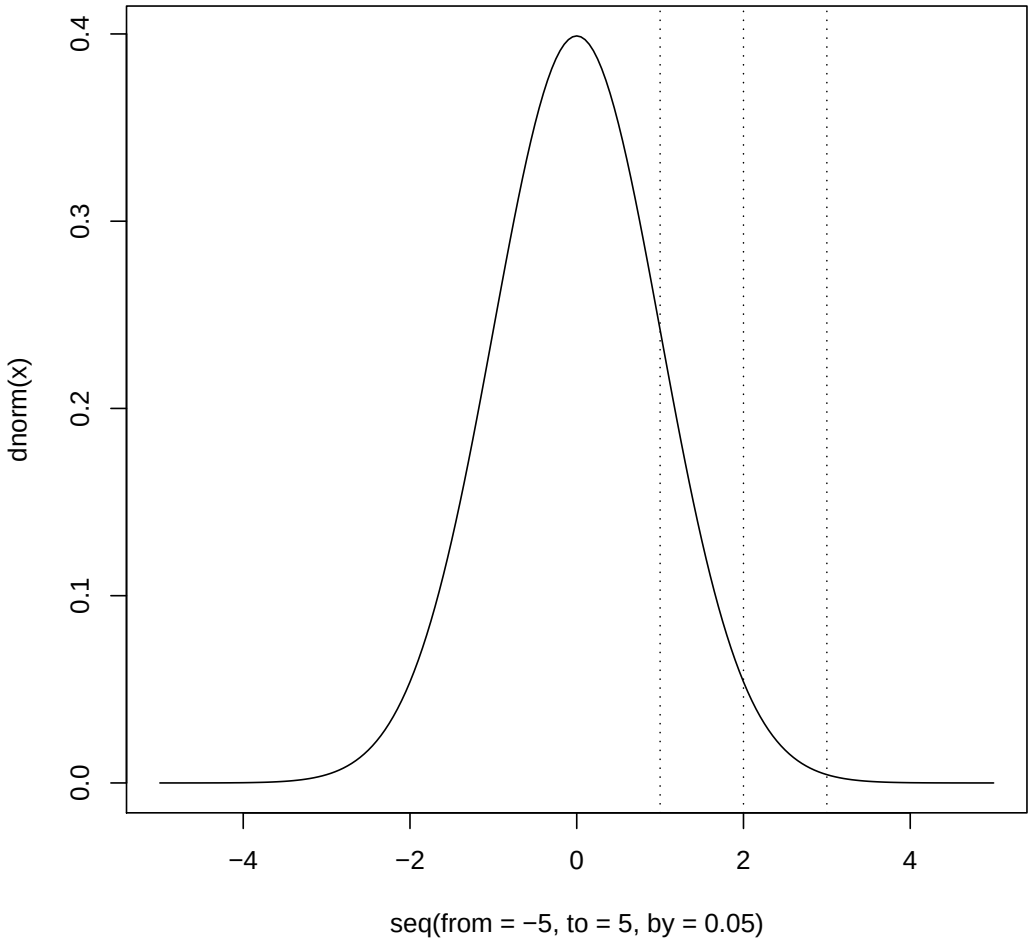
Outline

Klasyfikacja odległościowa w modelu *k–NN*

Założenie teoretyczne

W otoczeniu punktu $x \in X$ wyznaczanym przez metodę *k–NN* gęstość prawdopodobieństwa każdej z klas jest stała.

Najbliższe punkty... bezwzględnie? rangowo?

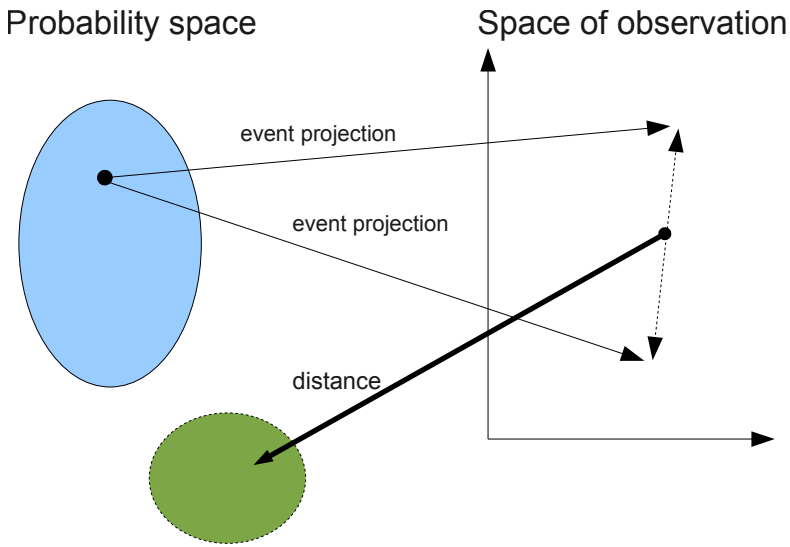


Najbliższe punkty

- klasycznie $d(x, y) = |x - y|$
- empirycznie $d(x, y) = |rank(x) - rank(y)|$
- probabilistycznie $d(x, y) = |F(x) - F(y)|$

Idea

$|F(x) - F(y)|$ to przecież....
... prawie (!) p-value hipotezy, że x oraz y to to samo zdarzenie



Uściślając...

Model przestrzeni probabilistycznej z błędem odwzorowania S_e:

- po "wylosowaniu" "odpowiedniego" zdarzenia, w procesie jego odwzorowywania do przestrzeni obserwacji następuje dodanie błedu
- klasycznie: $x = Sm(X) + Err$
- model: $x = Sm(X + Err)$
- istotne np. przy modelowaniu procesów biologicznych

Uściślając...

Probabilistyczna odległość wartość–zdarzenie:

- zakładamy, że przestrzenią, w której pracujemy jest S_e
- odległość pomiędzy punktem x a zdarzeniem u to wielkość błedu (rozkład jednostajny), jaki należałoby popełnić aby u odwzorowywało x

Uściślając...

Probabilistyczna odległość wartość–zdarzenie 1D:

$$d(x; v) = \int_{F^{-1}(x_c - |x_c - v_c|)}^{F^{-1}(x_c + |x_c - v_c|)} dF(x) =$$

$$= \min(1, x_c + |x_c - v_c|) - \max(0, x_c - |x_c - v_c|),$$

gdzie $x_c = F(x)$ oraz $v_c = F(v)$.

Model miary odległości

Pomiar odległości w przestrzeni wielowymiarowej:

1. pomiar odległości w poszczególnych wymiarach i przy pomocy miary D_i
2. połączenie wyników przy pomocy funkcji łączącej C

Pomiar odległości 1D

Probabilstyczna odległość zdarzenie–zdarzenie:

- wyprowadzana poprzez symetryzację $d(x; v)$
- przykład:

$$D_{\mathrm{ExpVal}}(u, v) = \frac{d(\frac{u+v}{2}; v) + d(\frac{u+v}{2}; u)}{2} \propto |F(u) - F(v)|$$

Funkcja łącząca

Probabilstyczna odległość zdarzenie–zdarzenie >1D:

- wprowadzana poprzez funkcję łączącą $R^n \rightarrow R$
- wielu przetestowanych kandydatów, np. łączenie kartezjańskie

$$C_{\text{std}}(x, y) = \sum_{i=1}^{n} D_i(x_i, y_i)^2$$

uśredniające

$$C_{\text{avg}}(x, y) = \frac{1}{n} \sum_{i=1}^{n} D_i(x_i, y_i)$$

Łączenie oparte o macierz korelacji

$$C_{\text{MahAvgSqrt}}(x, y) := \frac{1}{n} \sum_{i=1}^{n} \left(\Sigma^{-\frac{1}{2}} [D_i(x_i, y_i)]_{i=1}^{n} \right).$$

- wyprowadzone poprzez pierwiastkowanie jądra formy dwuliniowej
- złożoność obliczeniowa na podobnym poziomie (oszczędza się $n - 1$ mnożeń)

distance maeasure	**estimation method**	**BUPA**	**Pima**	**WDBC**	**Sonar**	**Ionosp.**
probability–based *with* outlier removal	leave-one-out CV	**26.67**	**21.88**	**2.28**	**11.06**	7.98
probability–based *without* outlier removal	leave-one-out CV	29.57	25.13	**2.28**	**11.06**	8.26
adaptive distance measure	leave-one-out CV	30.59	25.13	2.79	12.00	**4.29**
cam weighted distance	leave-one-out CV	35.3	24.7	3.5	*Non. avail.*	6.8
weighted distances	100 x 5–CV	36.22	27.33	*Non. avail.*	*Non. avail.*	*Non. avail.*
boosting distance estimation	100 x 20/80	33.58	28.91	4.67	25.67	16.27

Dziękuję za uwagę

Dziękuję za uwagę