Estymacja błędu predykcji i jej zastosowania

Jan Mielniczuk

Instytut Podstaw Informatyki PAN i
Wydział Matematyki i Nauk Informacyjnych PW
e-mail: miel@ipipan.waw.pl, miel@mini.pw.edu.pl

W przeglądowym wykładzie zostanie omówiony problem estymacji błędu predykcji i jej zastosowań w selekcji modelu i konstrukcji estymatorów postselekcyjnych. Podstawowym rozpatrywanym obiektem będzie prosta próba losowa $\mathcal{U} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, gdzie poszczególne obserwacje sa wektorami z R^{p+1}, a problemem estymacja funkcji regresji $f(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ na jej podstawie. Dla ustalonego estymatora $\hat{f}(\mathbf{x}, \mathcal{U}) = \hat{f}(\mathbf{x})$ i funkcji straty $L(f(y, f(\mathbf{x}))$ zostaną rozpatrzone: warunkowy błąd predykcji $Err_{\mathcal{U}} = E(L(Y^0, \hat{f}(X^0))|\mathcal{U})$, bezwarunkowy błąd predykcji $Err = E(Err_{\mathcal{U}})$ i błąd wewnątrzpróbkowy (*in-sample error*) Err_{in}, gdzie $(X^0, \mathbf{Y}^0)$ jest kopią $(X_1, \mathbf{Y}_1)$ niezależną od $\mathcal{U}$. $Err_{in} = n^{-1} \sum_{i=1}^n E_{\mathbf{Y}^0}(L(Y_i^0, \hat{f}(\mathbf{X}_i))|\mathcal{U})$, gdzie $\mathbf{Y}^0 = (Y_1^0, \ldots, Y_n^0)$ i Y_i^0 są niezależnie generowane z rozkładów $P_{Y|\mathbf{X}=\mathbf{X}_i}$ $i = 1, \ldots, n$. Przedstawione będą podstawowe estymatory tych wielkości, w szczególności estymator oparty na powtórnym podstawieniu $\bar{err}$ i estymator kroswalidacyjny. Postać $E_{(Y_1, Y_2, \ldots, Y_n)}(\bar{err})|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ prowadzi do tzw. poprawki kowariancyjnej i funkcji kryterialnej ze szczególną postacią funkcji kary. Własność ta motywuje podejście do problemu selekcji modelu przy użyciu funkcji kryterialnych oraz konstrukcję estymatorów postselekcyjnych. W dalszej części zostaną omówione własności tych estymatorów, w szczególności ich zgodność i konserwatywność oraz własności ryzyka. Podstawowym przykładem, dla którego będzie analizowane przedstawione podejście, jest model liniowy z losowymi wartościami atrybutów.