PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 27 2008
Ekonometria 22

Piotr Michalski

BIAS-VARIANCE TRADE-OFF IN SUPERVISED CLASSIFICATION

1. Introduction

In the last decade of the 20th century much research was devoted to obtaining more accurate approximations of distributions used in classification rules. The main logic behind this was the belief that greater estimation accuracy leads to better predictive properties of classifiers. As was established later, in many cases, enhanced precision of estimation – contrary to intuition – does not necessarily bring better classification results. At issue here is the generalization property of a predictive model, i.e. the ability to retain a predictive power for observations outside a learning sample. It is not uncommon that conceptually simple models, like naive bayesian classifiers or linear probability models outperform some sophisticated regression methods in classification settings. The article presents the decomposition of the expected prediction error in classification introduced by J.H. Friedman [4], which can be used to explain this phenomenon. A simulation example of error calculation via Friedman's decomposition is also given.

2. Prediction models

In a traditional prediction problem it is most often assumed that a continuous dependent variable Y is stochastically associated with non-random explanatory vector $\mathbf{X} = \left[X_1 \;\; X_2 \;\; \ldots \;\; X_p\right]$ through a function $Y = f(\mathbf{X}) + \varepsilon$, $f \in C^1$ where ε is a random component, such that $E(\varepsilon|\mathbf{X}) = 0$. In order to minimise the expected prediction error (with the assumption of the squared loss function) it suffices to estimate a conditional expected value (regression function) $f(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ point-

wise. Then the prediction problem amounts to the approximation of the function $f(\mathbf{X})$ using a training sample $T=\{x_i, y_i\}_1^n$. In a classification problem, with Y assuming a finite set of values decoded as $G = \{1, 2, \ldots, g\}$, one models conditional probabilities $P(Y=k|\mathbf{X}) = p_k(\mathbf{X})$ $k = 1, 2, \ldots, g$ or their monotonic transformations. The squared loss function is replaced by a matrix-defined loss function of the form $[L(i, j)]_{g\times g}$, where $L(i, j)$ is a cost incurred when predicting $Y = j$ while in reality $Y = i$ It is usually assumed, that $L(i,i) \le 0$ and $L(i, j) > 0$, $\forall i, j \in G$. The optimal decision function is the Bayes classifier (see [1, p. 65]):

$$d^*(\mathbf{x}) = \arg\min_k \sum_{i=1}^{g} L(i,k) p_i(\mathbf{x}),$$

where, according to Bayes theorem, $p_i(\mathbf{x}) = \pi_i p(\mathbf{x}|i) \Big/ \sum_{r=1}^{g} \pi_r p(\mathbf{x}|r)$ and $\pi_i = P(Y = i)$, $i = 1, 2, \ldots, g$. Now the task is to estimate either $p_i(\mathbf{x})$ in a direct fashion, or $p(\mathbf{x}|i)$, π_i, and insert them into Bayes theorem equation. An estimator of the Bayes classifier will be further denoted by $\hat{d}(\mathbf{x})$.

3. Bias-variance trade-off in regression setting

The precision of distribution estimation is a function of model complexity, which depends on the number of parameters (parametric models), or some parameters assuming prespecified values (nonparametric models). The general rule is that an increase in model complexity results in expected prediction error decrease within learning sample. Enhanced data fit does not usually guarantee that the results would be as satisfactory in a test sample – an overfitted model often loses its generalization properties, leading to an increased prediction error. Described mechanism characterizes a phenomenon called bias-variance trade-off, which insists that one should seek an optimal degree of model complexity that minimizes the expected prediction error on independent test sample data. The most favorable point lies somewhere between two extreme models, as depicted in figure 1.

The tool that allows an analitical description of the phenomenon is the decomposition of the expected prediction error into three components: random component, bias and variance (noise-bias-variance decomposition). The expected prediction error will be further denoted by Err and in regression it can be written as

$$Err = E\left[L\left(Y, \hat{f}(\mathbf{X})\right)\right], \tag{1}$$

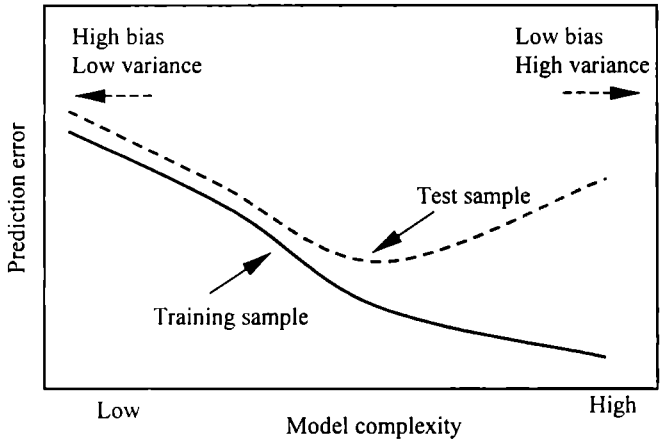

Fig. 1. Prediction error as model complexity function

Source: [5, p. 194].

where $f(\bullet)$ denotes regression function and $L(\bullet)$ is a loss function, that penalizes discrepancies between real value Y and its prediction $\hat{f}(\mathbf{X})$. Squared loss function is the most popular choice of statisticians. In classification one has

$$Err = E\left[L\left(Y, \hat{d}(\mathbf{X})\right)\right]. \tag{2}$$

Let's first consider the well-known case of regression. Suppose that the predictive dependency between features Y and $\mathbf{X}$ can be written as

$$Y = f(\mathbf{X}) + \varepsilon,$$

where $f(\mathbf{X})$ is a deterministic function, ε – random component, such that $E(\varepsilon \mid X) = 0$. The model has the property of $f(x) = E(Y \mid X = x)$. Having at one's disposal a training sample $\mathrm{T} = \{x_i, y_i\}_1^n$ the task is to find the best estimator of $f(X)$:

$$\hat{f}\left(\mathbf{x}|T\right) = \hat{E}\left(Y|\mathbf{X} = \mathbf{x}, \mathbf{x} \in T\right). \tag{3}$$

Let's note, that at point $\mathbf{X} = \mathbf{x}$ the function $\hat{f}\left(\mathbf{x}|T\right)$ is a random variable, as training sample T is also random. It is assumed that the value of the function at every point $\mathbf{x}$ follows a certain distribution $p\left(\hat{f}|\mathbf{x}\right)$ with known expected value and variance:

$$E\hat{f}(\mathbf{x}) = \int \hat{f} p\left(\hat{f}|\mathbf{x}\right) d\hat{f}, \tag{4a}$$

$$Var\hat{f}(\mathbf{x}) = \int \left(\hat{f} - E\hat{f}(\mathbf{x})\right)^2 p\left(\hat{f}|\mathbf{x}\right) d\hat{f}. \tag{4b}$$

The decomposition of the error (1) at a point x with the assumption of the squared loss function can be shown as:

$$\begin{aligned} Err(\mathbf{x}) &= E_{Y,T}\left[\left(Y - \hat{f}(\mathbf{x})\right)^2 |\mathbf{X} = \mathbf{x}\right] = E_Y\left[Y - f(\mathbf{x})\right]^2 + E_T\left[f(\mathbf{x}) - \hat{f}(\mathbf{x})\right]^2 = \\ &= E_Y\left[Y - f(\mathbf{x})\right]^2 + E_T\left[\hat{f}(\mathbf{x}) - E_T\hat{f}(\mathbf{x})\right]^2 + \left[E_T\hat{f}(\mathbf{x}) - f(\mathbf{x})\right]^2 = \\ &= Var\left(\varepsilon|\mathbf{X} = \mathbf{x}\right) + Var\hat{f}(\mathbf{x}) + Bias^2\left(\hat{f}(\mathbf{x})\right). \end{aligned} \tag{5}$$

The first term in the second and third line of (5) is an irreducible component of the prediction error (i.e. the variance of Y around its mean $f(\mathbf{X})$), resulting from the random nature of Y. The terms $E_T\left[\hat{f}(\mathbf{x}) - E_T\hat{f}(\mathbf{x})\right]^2$ and $\left[E_T\hat{f}(\mathbf{x}) - f(\mathbf{x})\right]^2$ are dependent only on the real mean $f(\mathbf{X})$ and its estimator $\hat{f}(\mathbf{X})$. $E_T\left[\hat{f}(\mathbf{x}) - E_T\hat{f}(\mathbf{x})\right]^2$ is the variance of $\hat{f}(\mathbf{X})$, characterizing sensitivity of $\hat{f}(\mathbf{X})$ to changes in a learning sample (new observations). $\left[E_T\hat{f}(\mathbf{x}) - f(\mathbf{x})\right]^2$ – the square of the bias – is the square of a value, by which the mean estimate $\hat{f}(\mathbf{X})$ differs from its actual mean $f(\mathbf{X})$. It is additionally assumed, that learning samples are of the same size and each time drawn from the same distribution $p(\mathbf{X}, Y)$.

For a given bias, increase in a sample size usually leads to a drop in variance. As large samples are common, in practice it is bias that constitutes the main proportion of prediction error. This observation aroused interest in more flexible methods, that aim at bias reduction and simultaneously prevent a model from overfitting (increased variance). Such approach turned out to be successful in regression, but brought disappointing results in classification setting. Startling research results (see [2; 4]), stating that simple methods in a classification problem are no worse or often perform better than more sophisticated ones, encouraged new direction of research and explain their resilience. The next paragraph presents the decomposition of the expected prediction error in classification by J.H. Friedman that provides a coherent conceptual framework elucidating specifics of a supervised classification problem and indicating new ways of improving classifiers. A concise account of other approaches to bias-variance decomposition in supervised classification may be found in [6].

4. Bias-variance trade-off in classification setting – Friedman's decomposition

The decomposition of the expected prediction error was proposed by J.H. Friedman in 1997 (see [4, pp. 55-77]). It concerns the case of two categories decoded as a random variable Y, which assumes two values 0 and 1, $Y \in \{0, 1\}$, and zero-one loss function (the decomposition can be generalized into any loss function). It is assumed that at every point x the variable Y follows a distribution defined by probabilities

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = 1 - P(Y = 0|\mathbf{X} = \mathbf{x}) = p_1(\mathbf{x}). \tag{6}$$

The expected prediction error in independent test sample has the form

$$Err(\mathbf{x}) = \pi_0 P(\hat{d}(\mathbf{x}) = 1|0) + \pi_1 P(\hat{d}(\mathbf{x}) = 0|1) = P(Y \neq \hat{d}(\mathbf{x})|\mathbf{X} = \mathbf{x}). \tag{7}$$

Considering the fact that there are only two categories, Bayes classifier and its approximation can be written as

$$d^*(\mathbf{x}) = I(p_1(\mathbf{x}) > 1/2), \tag{8}$$

$$\hat{d}(\mathbf{x}) - I(\hat{p}_1(\mathbf{x}) > 1/2). \tag{9}$$

The classifier estimate (9) has the full form $\hat{d}(\mathbf{x}|T) = I(\hat{p}(\mathbf{x}|T) > 1/2)$, but the shorter notation will be kept for clarity.

Expanding (7) one obtains

$$\begin{aligned} Err(\mathbf{x}) &= P(Y \neq \hat{d}(\mathbf{x})) = \\ &= P(Y = d^*(\mathbf{x}))P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x})) + P(Y \neq d^*(\mathbf{x}))P(\hat{d}(\mathbf{x}) = d^*(\mathbf{x})) = \\ &= P(Y = d^*(\mathbf{x}))P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x})) + P(Y \neq d^*(\mathbf{x}))\left[1 - P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}))\right] = \\ &= P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}))\left[P(Y = d^*(\mathbf{x})) - P(Y \neq d^*(\mathbf{x}))\right] + P(Y \neq d^*(\mathbf{x})) = \\ &= P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}))\left[1 - 2P(Y \neq d^*(\mathbf{x}))\right] + P(Y \neq d^*(\mathbf{x})). \end{aligned} \tag{10}$$

Let $Err_B(\mathbf{x}) = P(Y \neq d^*(\mathbf{x}))$ denote the irreducible *bayesian error rate* at a point $\mathbf{x}$ – the analogue of the random component's variance in regression decomposition. Then one can write:

$$Err(\mathbf{x}) = P(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}))[1 - 2Err_B(\mathbf{x})] + Err_B(\mathbf{x}). \tag{11}$$

From (11) it is seen that the bias and the variance influence the expected prediction error not additively, as in regression, but in a multiplicative way. The expected prediction error consists of a noise component and the product of two elements: one which depends on the noise and $P\left(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}) | \mathbf{X} = \mathbf{x}\right)$ – the analogue of $E_T\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2$ from (5).

One can also notice that for a given training sample T the error $Y \neq \hat{d}(\mathbf{x})$ is determined by an agreement between the decision (9) and the bayesian decision (8). In case of an agreement, the Bayesian error $Err_B(\mathbf{x}) = P\left(Y \neq d^*(\mathbf{x}|T)\right) = \min\{p_1(\mathbf{x}), 1 - p_1(\mathbf{x})\}$ is incurred, otherwise one can expect an increased error rate $P\left(Y \neq \hat{d}(\mathbf{x}|T)\right) = \max\{p_1(\mathbf{x}), 1 - p_1(\mathbf{x})\} = |2p_1(\mathbf{x}) - 1| + Err_B(\mathbf{x})$, which can be jointly written as

$$P\left(Y \neq \hat{d}(\mathbf{x}|T)\right) = |2p_1(\mathbf{x}) - 1| I\left(\hat{d}(\mathbf{x}|T)\right) + Err_B(\mathbf{x}). \tag{12}$$

Averaging (12) over all possible samples T one concludes that equivalently to (11) $Err(\mathbf{x})$ may be shown as

$$Err(\mathbf{x}) = |2p_1(\mathbf{x}) - 1| P\left(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}) | \mathbf{X} = \mathbf{x}\right) + Err_B(\mathbf{x}). \tag{13}$$

Friedman argues that it is reasonable to assume a normal distribution of the estimate $\hat{p}_1(\mathbf{x})$. One has therefore $\hat{p}_1(\mathbf{x}) \sim N\left(E\hat{p}_1(\mathbf{x}), Var\hat{p}_1(\mathbf{x})\right)$, where $E\hat{p}_1(\mathbf{x})$ and $Var\hat{p}_1(\mathbf{x})$ can be written in analogy to (4a-b), and the term $P\left(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}) | \mathbf{X} = \mathbf{x}\right)$ can be expressed as

$$P\left(\hat{d} \neq d^*\right) \approx I(p_1 < 1/2) \int_{0,5}^{\infty} c(\hat{p}_1) d\hat{p}_1 + I(p_1 \geq 1/2) \int_{-\infty}^{0,5} c(\hat{p}_1) d\hat{p}_1 . \tag{14}$$

Equation (14) restricts calculation to a specific point $\mathbf{X} = \mathbf{x}$ and $c\left(\hat{p}_1\right)$ denotes a density function of variable $\hat{p}_1$. Alternatively to (14) one has

$$P\left(\hat{d}(\mathbf{x}) \neq d^*(\mathbf{x}) | \mathbf{X} = \mathbf{x}\right) \approx \Phi\left(\frac{\text{sign}(1/2 - p_1(\mathbf{x}))(E\hat{p}_1(\mathbf{x}) - 1/2)}{\sqrt{Var\,\hat{p}_1(\mathbf{x})}}\right), \tag{15}$$

where $\Phi(\bullet)$ is the standard normal cumulative distribution function.

The value $\text{sign}(1/2 - p_1(\mathbf{x}))(E\hat{p}_1(\mathbf{x}) - 1/2)$ can be thought of as the *boundary bias*, for it is dependent on $p_1(\mathbf{x})$ through its location against the boundary 1/2. From (15) it is clear that if both $E\hat{p}_1(\mathbf{x})$ and $\hat{p}_1(\mathbf{x})$ are on the same side of the

boundary 1/2, then the boundary bias is negative and lowering the variance should cause the bias to decrease down to the minimal Bayes error rate. If $E\hat{p}_1(\mathbf{x})$ and $p_1(\mathbf{x})$ are on the opposite sides of the boundary, then the bias is positive and it seems advisable to increase the variance, as it leads to a drop in prediction error (see [5], p. 223). It is naturally preferable that the boundary bias be negative. If this is the case, then the classification error decreases as the value $|E\hat{p}_1(\mathbf{x}) - 1/2|$ increases, irrespective of the bias $p_1 - E\hat{p}_1$. This notice supports conclusion, that the main concern in a two-class supervised classification problem is not keeping the bias small, but retaining low variance, provided the boundary bias is predominantly negative.

Some highly biased methods (in the sense of the squared loss function) produce satisfying results in classification setting. Friedman points at a group of methods, which large bias is caused by excessive smoothing (oversmoothing). It is said, that a method is oversmoothing, when the estimate

$$\hat{p}_1(\mathbf{x}) = (1 - \alpha(\mathbf{x}))p_1(\mathbf{x}) + \alpha(\mathbf{x})\bar{y}$$

has a tendency to assume values close to the mean value of the dependent variable Y, which takes place when $\alpha(\mathbf{x})$ – where $\alpha(\mathbf{x}) \in [0,1]$ is a smoothing parameter – assumes values close to 1. As long as the decision boundary equals $\bar{y}$, the boundary bias remains negative (in our case the sample is balanced, i.e. $\bar{y} = 1/2$). One such method is the nearest neighbours method, in which the approximation of $p_1(\mathbf{x})$ consists in averaging class indicators of the k closest observations in a training sample. When $k \to n$ then $\alpha(\mathbf{x}) \to 1$, which entails $\hat{p}_1(\mathbf{x}) \to \bar{y}$.

The following points recap some general conclusions that can be used in practice.

- Decomposition of the expected prediction error is much more complex in supervised classification setting and unveils complicated, non-additive interplay between its components;
- Friedman's analysis, despite a two-class limitation, helps to explain the competitivness of classifiers, which base on biased probability estimators (ex. naive Bayes classifier, linear probability model);
- More accurate probability estimation does not necessarily lead to better classification results;
- Imposition of a constraint $\hat{p}_1(\mathbf{x}) \in [0,1]$ might decrease estimation bias, but simultaneously it may pose the danger of a boundary bias rise;
- It seems that in practical applications (ex. credit scoring) one should use accurate classification methods to generate a *score*, but if there is only a need for a classification decision, the aforementioned, simple methods would suffice.

The next section gives an example of Friedman's decomposition for three econometric models.

5. Comparison of econometric models with different bias via Friedman's decomposition – simulated data example

In this paragraph Friedman's decomposition of the expected prediction error will be used to compare classification properties of three econometric models: linear probability model (LPM), linear logistic regression model (GLM-Logit) and generalized additive logistic regression model (GAM-Logit). Linear probability models represent a class of simple, highly biased models. Linear logistic regression models are more advanced and less biased models. Generalized additive logistic regression models represent a group of new methods with low bias and adjustable variance. Using the generalized linear models notation $\eta = g\left[E\left(Y|\mathbf{X}\right)\right]$ (in case of binary data $E\left(Y|\mathbf{X}=\mathbf{x}\right)=p_1(\mathbf{x})$), where η is a predictor (function of $\mathbf{X}$) and $g(\bullet)$ is a link function, the three models may be written as follows:

- LPM: $\eta=\alpha_0+\alpha_1 X$, $g\left[E\left(Y|X\right)\right]=E\left(Y|X\right)$ (identity function),
- GLM-Logit: $\eta=\alpha_0+\alpha_1 X$, $g\left[E\left(Y|X\right)\right]=\text{logit}\left[E\left(Y|X\right)\right]=\ln\left[E\left(Y|X\right)/\left(1-E\left(Y|X\right)\right)\right]$,
- GAM-Logit: $\eta=\alpha+f(X)$, $g\left[E\left(Y|X\right)\right]=\text{logit}\left[E\left(Y|X\right)\right]=\ln\left[E\left(Y|X\right)/\left(1-E\left(Y|X\right)\right)\right]$.

In GAM-Logit $f(\bullet)$ is a smooth and nonparametric function, while α, α_0, α_1 are parameters. In this simulation example the ordinary least squares estimator was used for LPM, maximum likelihood estimator for GLM-Logit and iteratively re-weighted least squares algorithm estimator (see [3, p. 240]) with a smoothing spline as a scatterplot smoother for GAM-Logit. Three degrees of freedom were assumed for GAM-Logit smoothing spline. LPM and GLM-Logit models were both estimated using *glm* package in *R-project* environment. *Gam* package was used to estimate generalized additive logit models.

Friedman's expected prediction error decomposition will be used to estimate classification properties of the three classifiers. A mean deviation of residuals will be employed to bias measurement.

Suppose that X is non-random and takes values on the real line [0; 1] at 0,001 intervals starting from 0 (i.e. $x_1=0$, $x_2=0{,}001$, $x_{1000}=0{,}999$), and the Y-generating mechanism has the following form ($Y\in\{0;1\}$):

$$P\left(Y=1|X\right)=0{,}8I\left(X\in[0;0{,}5)\right)+0{,}2I\left(X\in[0{,}5;1]\right),$$

then the best classifier is given by

$$\hat{p}_1(x)>0{,}5 \text{ if } x\in[0;0{,}5) \text{ and } \hat{p}_1(x)\le 0{,}5 \text{ if } x\in[0{,}5;1],$$

and Bayes classifier can be written as $d^*(x) = I\{x \in [0; 0,5)\}$. The calculations of *Err* will be carried out in the proximity of a boundary 1/2, at the point $x = 0,45$. At this point the Bayes error $Err_B(0,45) = 0,2$. The approximate value of $P\left[\hat{d}(0,45) \neq d^*(0,45)\right]$ may be obtained from (14). One then has

$$P\left[\hat{d}(0,45) \neq d^*(0,45)\right] \approx \int_{-\infty}^{0,5} c(\hat{p}_1) d\hat{p}_1 . \tag{16}$$

Following Friedman, the distribution of $\hat{p}_1(0,45)$ will be modelled by normal distribution with expected value and standard deviation estimated in a simulation fashion. The simulation results are based on 10.000 replications.

Figure 2 shows mean estimates $\hat{p}_1$ with two standard deviations confidence bands for the three models, depicting the variability of the estimates $\hat{p}_1$.

Figure 3 shows the plots of the estimated probability density for the three distributions and Table 1 summarizes the simulation results. The last column contains areas under the probability density function plots of $c(\hat{p}_1)$ on $(-\infty; 0,5]$. The distribution of $\hat{p}_1$ is characterized by the parameters given in the first two columns.

Table 1. Simulation results at point $x = 0,45$ (10.000 replicates)

Model	$\hat{E}\hat{p}_1$	$\sqrt{\hat{V}ar\,\hat{p}_1}$	$10^{-4}\sum_{i=1}^{10000}\lvert y_i - \hat{p}_{1i}\rvert$	$P\left[\hat{d}(0,45) \neq 1\right]$
LPM	0,5400	0,0405	0,4721	0,1619
GLM-Logit	0,5499	0,0568	0,4645	0,1903
GAM-Logit	0,6077	0,0821	0,4162	0,0948

Table 2 contains the estimates of the expected prediction error in classification and its components. As an example the error for the linear probability model was calculated using the equation (11):

$$Err(0,45) = 0,1619 \times (1 - 2 \times 0,2) + 0,2 = 0,2971.$$

Table 2. Expected prediction error in classification and its components at point $x = 0,45$

Model	Boundary bias $sign(1/2 - p_1(x))(E\hat{p}_1(x) - 1/2)$	$\sqrt{\hat{V}ar\,\hat{p}_1}$	Err_B	$Err(0,45)$
LPM	−0,0400	0,0405	0,2	**0,2971**
GLM-Logit	−0,0499	0,0568	0,2	**0,3142**
GAM-Logit	−0,1077	0,0821	0,2	**0,2569**

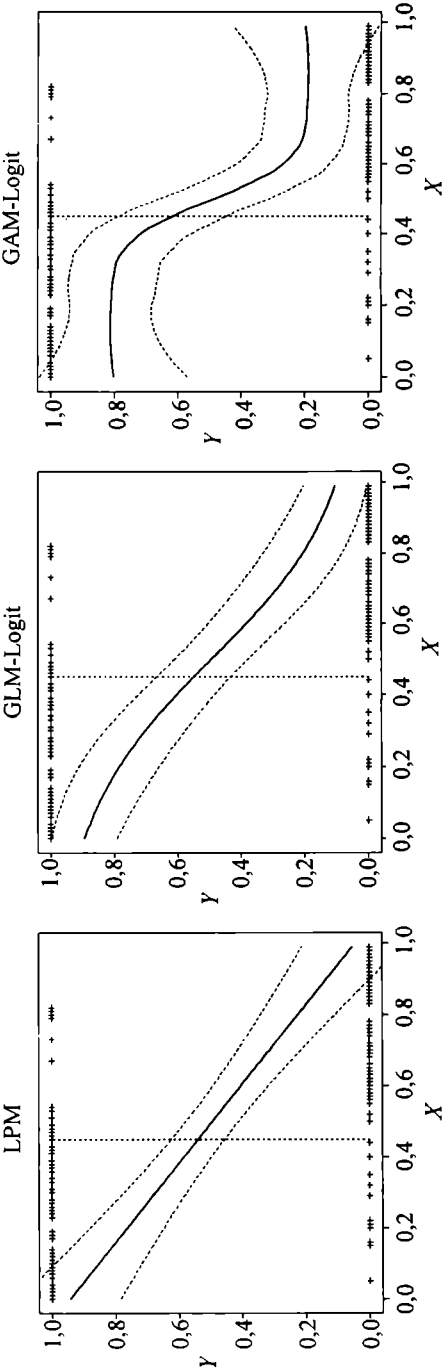

Fig. 2. Mean estimates $\hat{p}_1$ by linear probability model (LPM), linear logit model (GLM-Logit) and additive logit model (GAM-Logit) with two standard deviations confidence bands

From tables 1 and 2 one sees that at point $x = 0{,}45$ the boundary bias is negative in each case. The boundary bias of LPM and GLM-Logit are nearly equal and, as expected, LPM has the highest boundary bias, while GAM-Logit – the smallest. The GAM-Logit estimates are the most volatile, while the estimates of LPM – the most stable.

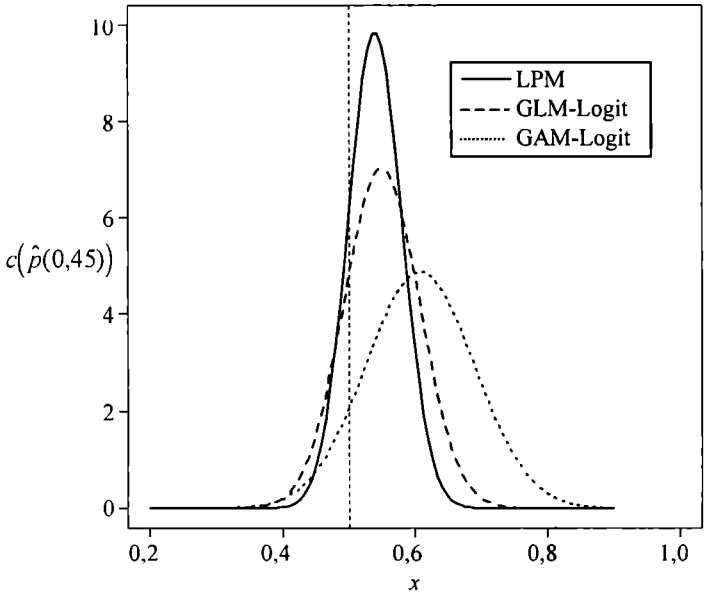

Fig. 3. Probability estimate density function $c(\hat{p}(0{,}45))$ of linear probability model (LPM), linear logistic regression model (GLM-Logit) and additive logistic regression model (GAM-Logit). The values (16) are equal to the areas under the function plot to the left of the dotted line

In the simulation experiment the linear probability model, which has the greatest bias (in the sense of the squared loss function), gave lower classification error rate than less biased linear logit model. This is caused by the smaller variance of the linear probability model, which neutralized the effect of a greater boundary bias. Although the GAM-Logit has the greatest variance, it gives lower boundary bias and the overall prediction error is the lowest of the three models. Figure 4 shows the expected prediction error estimated for the whole realm of X The conclusion is that the Err differences between the three models are immaterial.

Let now change the Y-generating mechanism to the form:

$$P(Y = 1|X) = 0{,}9I(X \in [0; 0{,}340]) + 0{,}2I(X \in [0341; 0{,}669]) + 0{,}9I(X \in [0{,}670; 0{,}999]),$$

so that probability $P(Y=1|X)$ is not a monotonically changing function of X. This is often the case in real classification problems, where some of the explanatory variables are nominants. The expected prediction error estimates, boundary bias and variance for the three models in the new scenario are shown in figure 5.

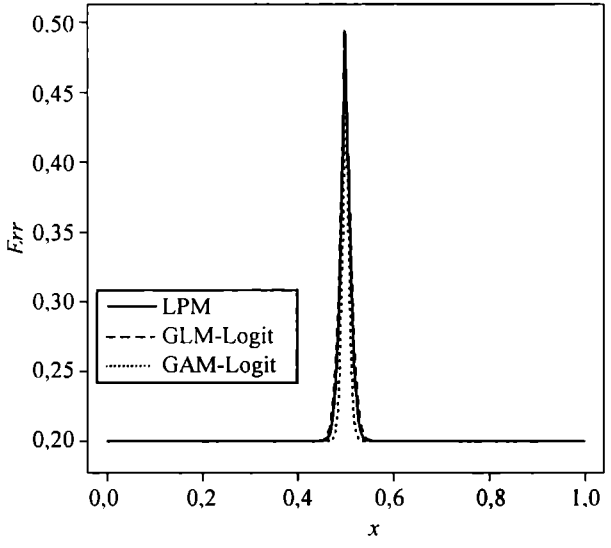

Fig. 4. Expected prediction error estimates for the whole realm of X in the first scenario

The problematic area for LPM and GLM-Logit is $X \in [0{,}341; 0{,}669]$, where the so called „masking" phenomenon occurs – the positive boundary bias is accompanied by low variance of the estimate of $\hat{p}_1$ and $Err = 1 - Err_B$. Two possible solutions to the problem are:

1) discretization of the variable X (most often used in business practice) within LPM or GLM-Logit;

2) using more accurate models like GAM-Logit, which – despite high variability – keep boundary bias negative.

The simulation results are, therefore, consistent with the claim, that there are classification situations, where simple probability estimators remain still competitive (figure 4 is an example). One of the prerequisites is that the variables are stimulants or destimulants. In other case one should resort to more flexible models to obtain negative bias.

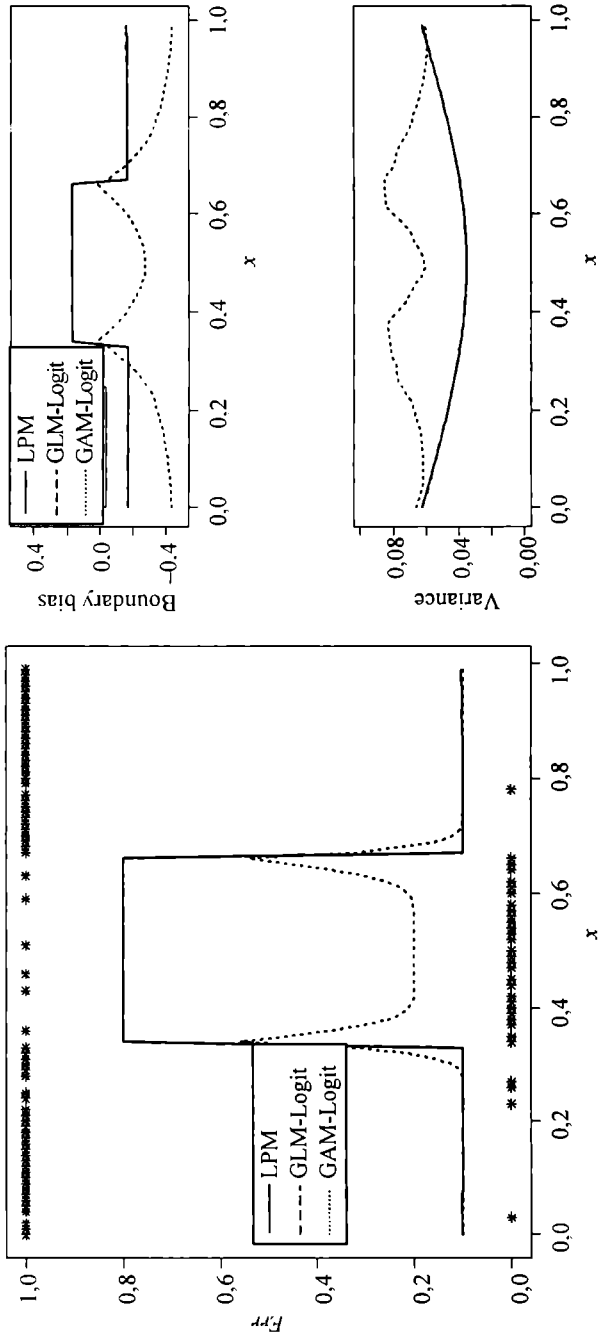

Fig. 5. Expected prediction error estimates for the whole realm of X and its components in the second scenario

Literature

[1] Ćwik J., Koronacki J., *Statystyczne systemy uczące się*, Wydawnictwo Naukowo-Techniczne, Warszawa 2005.

[2] Domingos P., *A Unified Bias-Variance Decomposition for Zero-One and Squared Loss*, Austin (USA): AAAI Press, Proceedings of the Seventeenth National Conference on Artificial Intelligence 2000, pp. 564-569.

[3] Faraway J.J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC Press, London 2006.

[4] Friedman J.H., *On Bias, Variance, 0/1-loss, and the Curse-of-dimensionality*, Kluwer Academic Publishers: Data Mining and Knowledge Discovery 1 1997, pp. 55-77.

[5] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York 2003.

[6] Le Borgne Yann-Ael, *Bias-Variance Trade-off Characterization in a Classification Problem. What Differences with Regression?* Machine Learning Group, Université Libre de Bruxelles 2005.

WYMIENNOŚĆ WARIANCJI I OBCIĄŻENIA W MODELU KLASYFIKACJI POD NADZOREM

Streszczenie

W artykule zaprezentowano podejście do dekompozycji oczekiwanego błędu predykcji w klasyfikacji według J.H. Friedmana. Dekompozycja ta ujawnia multiplikatywną wymienność wariancji i obciążenia w modelu klasyfikacji pod nadzorem oraz pozwala wyjaśnić klasyfikacyjną konkurencyjność prostych, obciążonych modeli, takich jak np. liniowy model prawdopodobieństwa. W artykule przedstawiono również symulacyjny przykład obliczenia oczekiwanego błędu predykcji w klasyfikacji za pomocą dekompozycji Friedmana, porównujący trzy modele ekonometryczne o różnym obciążeniu.

Piotr Michalski – mgr, doktorant w Katedrze Ekonometrii Uniwersytetu Ekonomicznego we Wrocławiu.