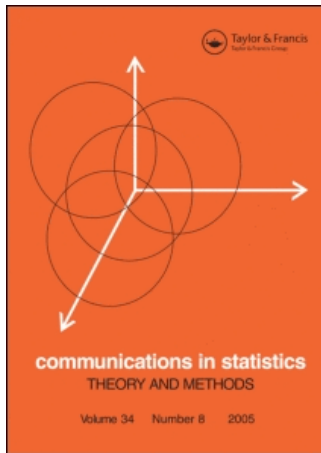


This article was downloaded by:[Juutilainen, Ilmari]  
On: 28 November 2007  
Access Details: [subscription number 782999223]  
Publisher: Taylor & Francis  
Informa Ltd Registered in England and Wales Registered Number: 1072954  
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Communications in Statistics - Theory and Methods

Publication details, including instructions for authors and subscription information:  
<http://www.informaworld.com/smpp/title~content=t713597238>

### A Method for Measuring Distance From a Training Data Set

Ilmari Juutilainen<sup>a</sup>; Juha Röning<sup>a</sup>

<sup>a</sup> Computer Engineering Laboratory, University of Oulu, Linnanmaa, Finland

Online Publication Date: 01 January 2007

To cite this Article: Juutilainen, Ilmari and Röning, Juha (2007) 'A Method for Measuring Distance From a Training Data Set', Communications in Statistics - Theory and Methods, 36:14, 2625 - 2639

To link to this article: DOI: 10.1080/03610920701271129

URL: <http://dx.doi.org/10.1080/03610920701271129>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

# A Method for Measuring Distance From a Training Data Set

ILMARI JUUTILAINEN AND JUHA RÖNING

Computer Engineering Laboratory, University of Oulu,  
Linnanmaa, Finland

*A new method is proposed for measuring the distance between a training data set and a single, new observation. The novel distance measure reflects the expected squared prediction error when a quantitative response variable is predicted on the basis of the training data set using the distance weighted  $k$ -nearest-neighbor method. The simulation presented here shows that the distance measure correlates well with the true expected squared prediction error in practice. The distance measure can be applied, for example, in assessing the uncertainty of prediction.*

**Keywords** Distance measure; Distance weighted  $k$ -nearest-neighbor; Model uncertainty; Novelty detection.

**Mathematics Subject Classification** Primary 62-07; Secondary 62G99, 51M05.

## 1. Introduction

In some applications, such as in evaluation of the reliability of prediction at a query point, it is interesting to measure the information given by the training data set about a new observation via the current prediction model. In this article, we propose a novel measure for the distance between a single observation or a query point and a training data set. We assume that the data set includes a response variable which is being predicted on the basis of the other variables. The distance measure reflects the expected uncertainty of the new observation being predicted on the basis of the training data set. The distance measure is a linear function of the approximated expected squared prediction error when the new observation is predicted with the distance weighted  $k$ -nearest-neighbor method.

There has been much discussion about measuring the distance between two observations. Different distance measures have been applied in the field of lazy

Address correspondence to Ilmari Juutilainen, Computer Engineering Laboratory, University of Oulu, P.O. Box 4500, FI-90014 Linnanmaa, Finland; E-mail: ilmari.juutilainen@ee.oulu.fi

learning. We refer to a review article of Wettschereck et al. (1997) that discusses the different methods. Often, Euclidean distance or Manhattan distance is used, and the problem lies in the scaling of the variables. The input variables that have a large effect on the response should have large weights in the distance measure. Unlike local distance measures, global distance measures use constant scaling, and some distance measures take the correlations between the explanatory variables into account.

The measurement of the distance between a set of observations and a single observation has also been widely discussed. Different distance measures have been applied in clustering and in prototype methods. In these applications, the aim in defining the distance has been to assign the observation to the nearest cluster or prototype. Examples of the different methods include the average pairwise distance, the Mahalanobis distance, and the Euclidian distance to the cluster centroid. We refer to Kaufman and Rousseeuw (1990) for these methods. However, these methods have been planned to measure the distance between a cluster and a single observation, and not the distance between a data set and a single observation.

Novelty detection aims to find abnormal observations from a data set. Abnormal observations can indicate that the modeled system is in an abnormal state, which needs to be reported. In classification, detection of novel observations is needed to identify new classes and observations that cannot be classified reliably. Novelty detection can be used to differentiate novel information from existing information when only the novel information needs to be shown to the learners. For novelty detection methods, we refer to the review by Markou and Singh (2003).

The usual approach in novelty detection is to somehow measure the similarity with the training data and to use some thresholds to interpret the observations as novel. The most common method is to model the joint density function of input variables to judge the observations with low density as novel (Markou and Singh, 2003). Our approach differs in that we do not construct any distribution model for the inputs. Our distance measure tries to measure the uncertainty about the expected response value at a new query point, which is quite a new approach to the problem.

We find the measure of the distance from a training data set most useful when looking at observations or query points on the boundaries of a data set. The accuracy of prediction on the boundary can be either good or unacceptable, which makes a distance reflecting the uncertainty of prediction very useful. The standard errors of predictions measure the uncertainty with model variance. Our approach differs in such a way that we also take bias into account. The theory of regression models holds on the assumption that a correct model structure is used and the prediction is unbiased. In practice, this is not true, and bias has a significant role, especially on the boundaries of the training data set.

Angiulli and Pittuzi (2005) suggested a method for detecting outliers in a data set. To measure the exceptionality they calculated the sum of the Euclidean distances to the  $k$ -nearest neighbors, which is quite similar to our proposal. Mahamud and Hebert (2003) discussed the optimal distance measures in the  $k$ -nearest-neighbor prediction. We construct our distance measure using a similar optimality principle.

## 2. Distance Between Two Single Observations

Let  $x_{(j)}$  refer to the  $j$ th explanatory variable and  $x_{ij}$  denotes the  $i$ th observation of  $x_{(j)}$ ,  $y_i$  denotes the  $i$ th observation of the response, and  $T$  denotes the training data

set consisting of  $N$  observations  $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ ,  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ . Let  $(x_0, y_0)$  be a new test data observation and  $d_i = d(x_0, x_i)$  measure the distance between  $x_0$  and  $(x_i, y_i) \in T$ . We assume that the response depends on the inputs via a regression function  $f(\cdot)$ , and that the additive error term has a constant variance

$$y_i = f(x_i) + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2 \varepsilon_i \sim i.i.d. \quad (2.1)$$

A distance measure  $d$  is a metric distance if it satisfies positivity:  $d(x_i, x_j) \geq 0$ , with equality if  $x_i = x_j$ , symmetry:  $d(x_i, x_j) = d(x_j, x_i)$ , and triangle inequality:  $d(x_i, x_j) + d(x_j, x_k) \geq d(x_i, x_k)$ . A distance measure  $d$  is monotonic if the distance is always smaller to the point which is nearer in all directions:  $|x_{il} - x_{jl}| \leq |x_{il} - x_{kl}|$  AND  $\text{sign}(x_{il} - x_{jl}) = \text{sign}(x_{il} - x_{kl}) \forall l = 1 \dots p \Rightarrow d(x_i, x_j) \leq d(x_i, x_k)$ . We suggest that a reasonable distance measure should be monotonic. The distance measure  $d$  is global if  $d(x_i, x_j) = d(x_i + a, x_j + a) \forall a$ . A global distance measure can be defined as a function of directional distance  $d(x_i, x_j) = d^*(x_i - x_j)$ . We prefer global distance measures for interpretational and computational reasons.

Mahamud and Hebert (2003) discussed the optimal distance measures in nearest-neighbor classification. The authors defined that an optimal distance measure in the 1-nearest-neighbor prediction minimizes the expected loss function  $E_{y_0, x_0, T} L(y_0, y')$ , where  $y'$  is the measured response at  $x'$ , which is the nearest neighbor of  $x_0$  using a distance measure  $d$ . The distance measure  $d(x_0, x_i) = EL(y_0, y_i)$  is optimal, because the nearest neighbor  $x' = \arg \min_{x_i} EL(y_0, y_i)$  minimizes the expected loss  $L(y_0, y') \forall x_0 \forall T$  (Mahamud and Hebert, 2003). The same reasoning holds for the  $k$ -nearest-neighbor method, and the expected loss  $EL(y_0, y_i)$  is an optimal distance measure also in  $k$ -nearest neighbor prediction. All order-preserving transformations of the expected loss function are optimal, because the nearest neighbors remain the same. The expected loss function whose optimality is considered in this study is the expected squared error related to the true expectation  $\mu_0 = E(y_0 | x_0) = f(x_0)$ ,

$$EL(\mu_0, y_i) = E(\mu_0 - y_i)^2 = E(y_0 - y_i)^2 - \sigma^2. \quad (2.2)$$

The optimal distance measure cannot be used directly because the conditional expectation of the response is not known, and the true expected loss cannot be solved. The optimal distance measure is not monotonic, which implies an interpretational disadvantage: The nearest neighbors may lie far away from the query point on the scale of explanatory variables. To eliminate these problems, we must be content with a linear approximation of the expected loss: We use the sum of the expectations of squared differences in the true regression function, where one input variable,  $x_{(j)}$ , at a time is set to the measured values  $x_{0j}$  and  $x_{ij}$ , and the other input variables are drawn randomly,

$$E_{y_i, x_0, x_i} (\mu_0 - y_i)^2 = \sigma^2 + E_{x_0, x_i} [f(x_0) - f(x_i)]^2 \approx \sigma^2 + \sum_{j=1}^p E_x \{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\}^2. \quad (2.3)$$

In the formula,  $x$  is a randomly drawn input observation,  $w_0^{(j)}(x)$  is otherwise identical with  $x$  but the  $j$ th element is altered  $w_0^{(j)}(x) = x_{0j}$ , and  $w_i^{(j)}(x)$  is otherwise identical with  $x$  but the  $j$ th element  $w_i^{(j)}(x) = x_{ij}$ . The equation (2.3) means that our

pairwise distance measure between two single observations is interpreted as the sum of expected loss caused by single input variables added by error variance. In the case of linear model the linear approximation holds exactly because  $E_{x_0, x_i, x} \{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\} = 0, \forall j = 1 \dots p$ .

For the continuous input variables  $x_{(1)}, x_{(2)}, \dots, x_{(p)}$ , the squared differences in  $y$  are approximated with the squared differences in the input variable values

$$E_x \{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\}^2 \approx \alpha_j (x_{0j} - x_{ij})^2, \quad j = 1, 2, \dots, p_1. \quad (2.4)$$

Mahamud and Hebert (2003) proposed estimating the  $\alpha$ -coefficients by fitting a regression model to a data set of pairs of training data instances using the response  $L(y_i, y_j)$ . The advantage of their direct method is that the regression function needs not be estimated. We propose a different method. Let our prediction model be

$$\hat{y} = \hat{f}(x) = \hat{f}(x_{(1)}, x_{(2)}, \dots, x_{(p)}), \quad (2.5)$$

and let  $\hat{\sigma}^2$  be the corresponding error variance estimate. Now let  $(x_c, y_c) \in T$  denote a training data observation lying near  $x_0$ , and let

$$\hat{f}'(x_c) = \left( \frac{\partial \hat{f}(x)}{\partial x_{(1)}}, \frac{\partial \hat{f}(x)}{\partial x_{(2)}}, \dots, \frac{\partial \hat{f}(x)}{\partial x_{(p)}} \right)_{(x=x_c)} \quad (2.6)$$

denote the gradient of the fitted response surface at point  $x_c$ . From the first-order Taylor approximation  $\hat{f}(x_0) \approx \hat{f}(x_c) + \hat{f}'(x_c)(x_0 - x_c)$ , we have the result

$$[\hat{f}(x_0) - \hat{f}(x_c)]^2 \approx [\hat{f}'(x_c)(x_0 - x_c)]^2 \quad (2.7)$$

which motivates us to suggest that  $\alpha$ -coefficients are defined as the average squared partial derivative over the training data set

$$\alpha_j = \frac{1}{N} \sum_{i=1}^N \left( \frac{\partial \hat{f}(x)}{\partial x_{(j)}} \right)_{(x=x_i)}^2, \quad j = 1, 2, \dots, p_1. \quad (2.8)$$

For large  $N$ , it is enough to calculate the average over a sample. The regression function can be fitted using any learning method, for example, neural networks or additive models. The partial derivatives of the fitted response surface with respect to each input variable  $x_{(j)}$  are approximated numerically with

$$\frac{\partial \hat{f}(x)}{\partial x_{(j)}} \Big|_{(x=x_i)} = \frac{\hat{f}(x_i) - \hat{f}(x_i + o_j)}{|o_j|}, \quad (2.9)$$

where  $o_j$  is a vector of zeros elsewhere, but a small constant at the  $j$ th element. We used the value  $|o_j| = \text{sd}(x_{(j)})/100$ .

When  $x_{(j)}$  is a categorical input variable with class levels  $\gamma_{j1}, \gamma_{j2}, \dots, \gamma_{jq_j}$ , we are interested in the distance between the class levels  $\gamma_{jl}$  and  $\gamma_{jm}$ . When one of the two observations belongs to the class  $\gamma_{jl}$  and the other to class  $\gamma_{jm}$ ; ( $x_{ij} = \gamma_{jl}$  AND  $x_{i0} =$

$\gamma_{jm}$ ) OR ( $x_{i0} = \gamma_{jl}$  AND  $x_{ij} = \gamma_{jm}$ ), we can estimate the expected squared difference in the response between each two class levels using the fitted prediction model with

$$E_x \{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\}^2 \approx \frac{1}{|J^{(\gamma_{jl}, \gamma_{jm})}|} \sum_{j \in J^{(\gamma_{jl}, \gamma_{jm})}} [\hat{f}(x_i) - \hat{f}(\tilde{w}^{(\gamma_{jl}, \gamma_{jm})}(x_i))]^2. \quad (2.10)$$

In the notation, the input vectors  $\tilde{w}^{(\gamma_{jl}, \gamma_{jm})}(x_i)$  are otherwise identical to  $x_i$ ,  $\tilde{w}_k^{(\gamma_{jl}, \gamma_{jm})}(x_i) = x_{ik}, \forall k \neq j$ , but the  $j$ th element is altered:

$$\tilde{w}_j^{(\gamma_{jl}, \gamma_{jm})}(x_i) = \begin{cases} \gamma_{jm}, & \text{if } x_{ij} = \gamma_{jl} \\ \gamma_{jl}, & \text{if } x_{ij} = \gamma_{jm}. \end{cases} \quad (2.11)$$

The squared differences in the prediction are averaged over the index set of observations, which belong to either of the two classes, and the response can be predicted reliably for both of these classes:  $J^{(\gamma_{jl}, \gamma_{jm})} = \{i \mid \hat{f}(\tilde{w}^{(\gamma_{jl}, \gamma_{jm})}(x_i)) \text{ is reliable AND } (x_{ij} = \gamma_{jl} \text{ OR } x_{ij} = \gamma_{jm})\}$ .

When  $x_{(j)}$  is a binary variable, we can notate

$$\alpha_j = \frac{1}{|J^{(\gamma_{jl}, \gamma_{jm})}|} \sum_{j \in J^{(\gamma_{jl}, \gamma_{jm})}} (\hat{f}(x_i) - \hat{f}(\tilde{w}^{(\gamma_{jl}, \gamma_{jm})}(x_i)))^2. \quad (2.12)$$

If  $x_{(j)}$  has more than two class levels, we can notate as follows: Let  $I_j(x_i)$  be a binary vector of length  $q_j$  with  $q_j - 1$  zeroes. The  $l$ th element  $I_{jl}(x_i)$  is not zero if, and only if, the value of the  $j$ th variable is the  $l$ th class level:  $I_{jl}(x_i) = 1 \iff x_{ij} = \gamma_{jl}$ . Let  $G^j$  be a  $q_j \times q_j$  matrix of distances between the class levels of  $x_{(j)}$

$$G_{ik}^j = 0, \quad \text{when } i = j \\ G_{ik}^j = \frac{1}{|J^{(\gamma_{jl}, \gamma_{jm})}|} \sum_{j \in J^{(\gamma_{jl}, \gamma_{jm})}} [\hat{f}(x_i) - \hat{f}(\tilde{w}^{(\gamma_{jl}, \gamma_{jm})}(x_i))]^2, \quad \text{when } i \neq j. \quad (2.13)$$

The expected squared difference in the response caused by the categorical input variable  $x_{(j)}$  can be expressed as

$$E_x \{f[w_i^{(j)}(x)] - f[w_0^{(j)}(x)]\}^2 = \alpha_j^*(x_0, x_i) = I_j(x_i)^T G^j I_j(x_0) \quad (2.14)$$

and the distance is obtained as the sum of the expected squared differences caused by single input variables.

We notate the  $p_1$  continuous input variables as  $x_{(1)}, \dots, x_{(p_1)}$  and the  $p - p_1$  categorical variables as  $x_{(p_1+1)}, \dots, x_{(p)}$ . We propose to use an approximate optimal distance measure that is the approximated expected squared error loss

$$d(x_0, x_i) = \alpha_0 + \sum_{j=1}^{p_1} \alpha_j (x_{0j} - x_{ij})^2 + \sum_{j=p_1+1}^p \alpha_j^*(x_0, x_i). \quad (2.15)$$

The coefficient  $\alpha_0$  is the error variance estimate  $\hat{\sigma}^2$  of our estimated prediction model  $\hat{f}$ . In other words, we use a modification of squared Euclidean distance to understand categorical variables and with an added constant as the pairwise

distance measure between two single observations. The distance measure is global and monotonic, but not metric. The distance measure  $\sqrt{d(x_0, x_i) - \alpha_0}$  is metric.

### 3. Distance Between a Single Observation and a Data Set

We suggest the distance of a single observation from a set of  $k$  observations,  $S_k$ , to be measured on the basis of the expected squared error when the single observation is predicted on the basis of  $S_k$ . This can be seen as a generalisation of the pairwise optimal distance measure. The true expected loss at  $x_0$  is not known and has to be approximated. We predict  $\mu_0 = E(y_0)$  with a distance-weighted linear combination of the  $y$  values measured in  $S_k$ , which results in measurement of the distance with the harmonic sum of pairwise distances.

Let  $S_k = \{(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)\}$  be a set of observations whose distances from  $x_0$  are  $d_1, d_2, \dots, d_k$  are optimal:  $d_i = d(x_0, x_i) = E(\mu_0 - y_i)^2$  (Eq. (2.2)). Let us now estimate  $\mu_0$  with a weighted linear combination  $\hat{y}_0 = \omega_1 y_1 + \omega_2 y_2 + \dots + \omega_k y_k$  of the measured responses. Under the symmetry assumption  $E(\mu_0 - y_i) = 0$ , the minimum variance unbiased estimator gives weights proportional to the inverses of the variances and sums the weights to unity

$$\omega_j = \frac{1}{d_j} / \sum_{i=1}^k \frac{1}{d_i} \quad (3.1)$$

(Goldberg et al., 2005). We use this distance-weighted estimator

$$\hat{y}_0 = \left( \sum_{i=1}^k \frac{1}{d_i} y_i \right) / \sum_{i=1}^k \frac{1}{d_i} \quad (3.2)$$

to predict  $y_0$  based on  $S_k$ . We keep the estimator Eq. (3.2) as a natural basis for the interpretation of our distance measure because the approach does not make any assumption about the form of the regression function. The expected squared loss of our estimator is the harmonic sum of pairwise distances  $d_i$  plus a bias term

$$\begin{aligned} E(\hat{y}_0 - \mu_0)^2 &= E\left(\sum_{i=1}^k \omega_i y_i - \mu_0\right)^2 \\ &= E\left(\sum_{i=1}^k \omega_i (y_i - \mu_0)\right)^2 \\ &= \sum_{i=1}^k \omega_i^2 E(y_i - \mu_0)^2 + \sum_{j=1}^k \sum_{i \neq j} E(y_i - \mu_0) E(y_j - \mu_0) \omega_i \omega_j \\ &= \left(\frac{1}{\sum_{i=1}^k \frac{1}{d_i}}\right)^2 \left[ \sum_{i=1}^k d_i / d_i^2 + \sum_{j=1}^k \sum_{i \neq j} \frac{E(y_i - \mu_0) E(y_j - \mu_0)}{d_i d_j} \right] \\ &= \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} + \left(\frac{1}{\sum_{i=1}^k \frac{1}{d_i}}\right)^2 \sum_{j=1}^k \sum_{i \neq j} \frac{E(y_i - \mu_0) E(y_j - \mu_0)}{d_i d_j}. \end{aligned} \quad (3.3)$$

We also take the expectations Eqs. (3.3), (3.6), (3.10), and (3.11) over  $x_i$ , which means that  $x_i$  are assumed to be random points satisfying the condition  $d(x_0, x_i) = d_i$ .

From  $d_i = E(\mu_0 - y_i)^2$  it follows that  $|E(y_i - \mu_0)| = \sqrt{d_i - \sigma^2}$ . The absolute value of the bias term

$$\left( \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} \right)^2 \sum_{j=1}^k \sum_{i \neq j} \frac{E(y_i - \mu_0)E(y_j - \mu_0)}{d_i d_j} \quad (3.4)$$

increases along with the pairwise distances,  $d_i$ . If the assumption  $E_{y, x_i | d_i}(y_i - \mu_0) = 0 \forall i$  holds, the bias term would be zero, and the expected squared error would be the harmonic sum of the pairwise distances. However, that is not a realistic assumption. Some query points  $x_0$  may lie in a ‘symmetric’ position, where the assumption holds. But some query points may lie at the bottom of a valley, where the expectation  $E(y_i - \mu_0)$  is positive for all possible neighbors  $x_i$ , or on the top of a hill, where  $E(y_i - \mu_0)$  is negative.

The following lemma gives an upper bound for the bias term when  $x_0$  lies at the bottom of the valley.

**Lemma 3.1.** *If  $d_i = E(y_i - \mu_0)^2$ ,  $\text{Var}(y_i) = \sigma^2$ , and  $E(y_i - \mu_0) = \sqrt{d_i - \sigma^2}$ , then*

$$\left( \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} \right)^2 \sum_{j=1}^k \sum_{i \neq j} \frac{E(y_i - \mu_0)E(y_j - \mu_0)}{d_i d_j} \leq \frac{k-1}{\sum_{j=1}^k \frac{1}{d_j}} - \sigma^2 \frac{k-1}{k}. \quad (3.5)$$

*Proof.* Let  $\bar{d}$  denote the average inverse distance  $\bar{d} = \frac{1}{k} \sum_{i=1}^k \frac{1}{d_i}$ . The bias term can be written as

$$\begin{aligned} \sum_{j=1}^k \sum_{i \neq j} \frac{E(y_i - \mu_0)E(y_j - \mu_0)}{d_i d_j} &= \sum_{j=1}^k \sum_{i \neq j} \frac{1}{d_j} \sqrt{\frac{1}{d_j} - \sigma^2} \frac{1}{d_i} \sqrt{\frac{1}{d_i} - \sigma^2} \\ &= \sum_{j=1}^k \sum_{i \neq j} q\left(\frac{1}{d_j}, \frac{1}{d_i}\right). \end{aligned} \quad (3.6)$$

The function  $q(x, z) = x\sqrt{1/x - \sigma^2} z\sqrt{1/z - \sigma^2}$  is concave; it can be proved by showing that the matrix of its second derivatives is negative semidefinite. From the concavity of  $q$ , it follows that there exists an affine function  $s(x, z) = \kappa_0 + \kappa_1 x + \kappa_2 z$  for which  $q(\bar{x}, \bar{z}) = s(\bar{x}, \bar{z})$  and  $q(x, z) \leq s(x, z)$  (Kallenberg, 1997, p. 26). Now  $\frac{1}{k(k-1)} \sum_{j=1}^k \sum_{i \neq j} \frac{1}{d_j} = \frac{1}{k(k-1)} \sum_{j=1}^k \sum_{i \neq j} \frac{1}{d_i} = \bar{d}$ . Thus, we can apply the multivariate Jensen inequality  $\sum_{i=1}^k q(x_i, z_i) \leq \sum_{i=1}^k s(x_i, z_i) = k(\kappa_0 + \kappa_1 \frac{1}{k} \sum_{i=1}^k x_i + \kappa_2 \frac{1}{k} \sum_{i=1}^k z_i) = ks(\bar{x}, \bar{z}) = kq(\bar{x}, \bar{z})$  to Eq. (3.6) giving

$$\sum_{j=1}^k \sum_{i \neq j} \frac{1}{d_j} \sqrt{\frac{1}{d_j} - \sigma^2} \frac{1}{d_i} \sqrt{\frac{1}{d_i} - \sigma^2} \leq k(k-1) \bar{d}^2 \left( \frac{1}{\bar{d}} - \sigma^2 \right) \quad (3.7)$$

and after simplifying

$$\left( \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} \right)^2 k(k-1) \bar{d}^2 \left( \frac{1}{\bar{d}} - \sigma^2 \right) = \frac{k-1}{\sum_{j=1}^k \frac{1}{d_j}} - \sigma^2 \frac{k-1}{k} \quad (3.8)$$

we get the result.  $\square$

The equality holds exactly when all the neighbours are equally distant  $1/d_i = \bar{d} \forall i$ . If  $x_0$  lies on the top of a hill, the only difference is that  $E(y_i - \mu_0) = -\sqrt{d_i - \sigma^2}$ . When all the neighbors are roughly equally distant and  $x_0$  lies at the bottom of a valley or on the top of a hill, the bias term can be approximated as a linear function of the harmonic sum

$$\frac{1}{\sum_{i=1}^k \frac{1}{d_i}}, \quad (3.9)$$

the number of neighbors,  $k$ , and error variance,  $\sigma^2$ ,

$$\left( \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} \right)^2 \sum_{j=1}^k \sum_{i \neq j} \frac{E(y_i - \mu_0)E(y_j - \mu_0)}{d_i d_j} \approx \frac{k-1}{\sum_{i=1}^k \frac{1}{d_i}} - \sigma^2 \frac{k-1}{k}. \quad (3.10)$$

At all query points  $x_0$ , the true bias can be expressed in relation to the maximum bias with  $E_{y,x_j|x_0,d_i}(y_i - \mu_0) = c(x_0)\sqrt{d_i - \sigma^2}$ . When  $x_0$  lies in a symmetric position,  $c(x_0) = 0$ : at the bottom of the valley  $c(x_0) = 1$ ; on the top of the hill  $c(x_0) = -1$ . When we assume that  $c(x_0)$  does not depend on the distance  $d_i$  and denote  $E_{x_0}c(x_0)^2 = \delta^2$ , the expected squared prediction error can be approximated with

$$\begin{aligned} E_{y,x|d_1 \dots d_k}(\mu_0 - \hat{y}_0)^2 &= E_{x_0} \left( \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} \right)^2 \sum_{j=1}^k \sum_{i \neq j} \frac{c(x_0)^2 \sqrt{d_i - \sigma^2} \sqrt{d_j - \sigma^2}}{d_i d_j} + \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} \\ &\approx \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} + \delta^2(k-1) \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} - \sigma^2 \delta^2 \frac{k-1}{k}. \end{aligned} \quad (3.11)$$

This is a linear transformation of the harmonic sum when  $k$  is kept fixed. Thus, the harmonic sum  $1/\sum_{i=1}^k \frac{1}{d_i}$  can be used as a measure of the uncertainty about  $\mu_0$  when  $y_1, \dots, y_k$  and  $d_1, \dots, d_k$  are given. Because the true distances  $d_i = E(\mu_0 - y_i)^2$  are not known, they are replaced by the pairwise distances  $d_i = d(x_0, x_i)$  of Eq. (2.15), and in the approximations Eq. (3.11), the unknown  $\sigma^2$  is replaced by  $\alpha_0$ .

We propose the distance between a single observation  $x_0$  and a set of observations  $S_k$  to be measured with the harmonic sum of pairwise distances  $d_i = d(x_i, x_0)$ . When the pairwise distances correspond to the expected squared error  $d_i \approx E(\mu_0 - y_i)^2$ , our distance measure  $d(x_0, S_k)$  approximates an increasing linear function of the expected squared prediction error  $E(\mu_0 - \hat{y}_0(S_k))^2$ . We suggest the distance between  $x_0$  and  $S_k$  to be measured with

$$\begin{aligned} d(x_0, S_k) &= \frac{1}{\sum_{i=1}^k \frac{1}{d_i}} \\ d_i &= \alpha_0 + \sum_{j=1}^{p_1} \alpha_j (x_{0j} - x_{ij})^2 + \sum_{j=p_1+1}^p \alpha_j^* (x_0, x_i). \end{aligned} \quad (3.12)$$

#### 4. Measuring the Distance to a Training Data Set

Our method could be used directly to measure the distance between a single observation and the training data set by letting  $S_k = T$ . However, when the training

data set is large, it makes more sense to use only the  $k$  nearest observations. In the  $k$ -nearest-neighbor method, typically 5–100 neighbors are used to obtain the most accurate prediction. Thus, the observations lying far away from  $x_0$  should not have an effect on the distance measure, because they do not affect the prediction. Let  $d^{(k)}$  be the  $k$ th smallest distance  $d(x_0, x_i)$ . Our suggestion for the distance between the training data set and a single observation is

$$\begin{aligned} d(x_0, T) &= d(x_0, S_k), \\ S_k &= \{(x_i, y_i) \in T \mid d(x_0, x_i) \leq d^{(k)}\}. \end{aligned} \quad (4.1)$$

Our distance measure is problem-dependent. If we have the same inputs and several responses, the distance measure has to be defined separately for each response. The distance measure adapts itself to the estimated regression function. If a variable has only a small effect on the response, it will also have little effect on the distance. The distance measure is invariant for linear transformations and approximately invariant for order-preserving transformations of the inputs. The distance measure also has a reasonable interpretation as the approximate measure of the expected loss function, which is an informative and novel way to measure the uncertainty about a new observation. The distance measure uses the squared error loss function, but can also be used for non Gaussian responses. If  $\mu_0$  were estimated with the unweighted  $k$ -nearest-neighbor method, the result would be the sum of single distances, precisely as proposed by Angiulli and Pittuzi (2005). The scale of the distance measure depends on the scale of the response.

After the distance measure has been initialized by defining the  $\alpha$ -coefficients, the major computational task is to find the  $k$ -nearest training data observations. The computation of a single distance to the training data set requires about  $N(p+2) + k^2$  operations. With the algorithm of Friedman et al. (1979), the nearest neighbors can be found in a time proportional to  $\log N$ . The organization of a data set for the algorithm requires about  $pN \log N$  operations (Friedman et al., 1979). Initialization of the distance measure consists of fitting a prediction model and defining the  $\alpha$ -coefficient for each explanatory variable.

If there were an observation  $(x_a, y_a) \in T$  whose pairwise distance from the prediction point  $x_0$  were zero,  $d(x_0, x_a) = 0$ , the distance to the training data would be zero  $d(x_0, T) = 0$ . In practice, this cannot happen because the minimum pairwise distance  $\min d(x_0, x_a) = \alpha_0 > 0$  is achieved when  $x_a$  is a replicate of  $x_0$ . If we have  $k$  replications of  $x_0$  in the training data, then  $d(x_0, T) = \alpha_0/k \approx \sigma^2/k$ . The standardized distance measure

$$\frac{d(x_0, S_k) - \alpha_0/k}{\alpha_0} \quad (4.2)$$

is invariant to the scale of the response. The standardized distance is non negative and zero if, and only if, there are  $k$  replications of  $x_0$  in the training data set. If all the observations are far away from  $x_0$ , they give little information for prediction, and the distance measure obtains high values.

Our distance measure is strongly dependent on the size of the neighborhood. The distance measure decreases as a function of  $k$ . In a neighborhood of optimal size, the distance measure best reflects the uncertainty of the model. The size of the neighborhood,  $k$ , should be so large that the observations giving significant predictive

information would be included in the distance, but their effect on the distance would not be obscured by distant observations. The optimal  $k$  is surely problem-dependent. We suggest the use of  $k = 30$ , because that seemed to work best in our simulation studies. Also, it seems intuitively reasonable that the distance to the training data can be defined on the basis of the distances to the 30 nearest neighbors.

## 5. Performance in Simulated Data Sets

The proposed distance measure reflects the expected squared error loss function

$$d(x_0, S_k) \approx c_1 + c_2 E(\mu_0 - \hat{y}_0(S_k))^2. \quad (5.1)$$

---

```

data set:  $X_T$  = initialise 10000  $\times$  16 matrix
clusterCounts = (3, 6, 30, 40, 50, 100, 200)
clusterSizes = (1000, 300, 100, 30, 10, 3, 1)
counter = 0
for  $i = 1$  to 7 do
  for  $j = 1$  to clusterCounts[ $i$ ] do
    enter = NID(8)
    rDev = select from [0.03, 0.05, 0.08, 0.1, 0.13, 0.15, 0.17, 0.18, 0.2, 0.22, 0.23, 0.25, 0.27, 0.3]
    to clusterSizes[ $i$ ] do
      ++
      center, 1 : 8] = clusterCenter + inClusterDev * NID(8)

  next j
next i
for  $i = 9$  to 16 do
  ownDev = select randomly from [0.69, 0.63, 0.57, 0.51, 0.45, 0.39, 0.33, 0.28, 0.25, 0.22, 0.2]
  while  $RSquared < 1 - ownDev^2$  do
    correlatingVariable = randomly select one unselected integer 1-8
    corrCoeF = random instance from uniform(0, 1.2)
     $RSquared = corrCoeF^2$ 
    if  $RSquared > ownDev$  then
      coefficient =  $(1 - ownDev^2 - RSquared + corrCoeF^2)^{1/2}$ 
       $X_T[, i] = corrCoeF * X_T[, correlatingVariable]$ 
    else  $X_T[, i] = corrCoeF * X_T[, correlatingVariable]$ 
    end if
  loop
next i

the expected response: trueY = initialise vector of size 10000
for  $i = 1$  to 24 do
  monotone = randomise with probabilities  $P(1) = 0.7, P(0) = 0.3$ 
  nOfAttendingVariables = randomise,  $P(1) = 0.5, P(2) = 0.25, P(3) = 0.125, P(4) = 0.125$ 
  randomise  $b = \exp(0.5NID(1))$ 
  randomise  $a = 0.25NID(1)$ 
  ( $v_1, v_2, v_3, v_4$ ) = random sample of integers between 1 and 16
  randomise  $(\beta_1, \beta_2, \beta_3, \beta_4) = \pm 0.6 + 0.25NID(4)$ 
  if nOfAttendingVariables = 1 then  $\beta_2 = 0, \beta_3 = 0, \beta_4 = 0$ 
  else if nOfAttendingVariables = 2 then  $\beta_3 = 0, \beta_4 = 0$ 
  else if nOfAttendingVariables = 3 then  $\beta_4 = 0$ 
  end if
   $\nu = \pm 20 / \left[ \left( 0.2 + \sqrt{\frac{2}{\pi} \sum_{j=0}^4 \beta_j^2} \right)^b + b(b-1) \left( 0.2 + \sqrt{\frac{2}{\pi} \sum_{j=0}^4 \beta_j^2} \right)^{b-2} \sum_{j=0}^4 \beta_j^2 \right]$ 
  if monotone = 1 then
    trueY +=  $\nu \text{sign}(a + \beta_1 X_T[, v_1] + \beta_2 X_T[, v_2] + \beta_3 X_T[, v_3] + \beta_4 X_T[, v_4]) \text{abs}(a + \beta_1 X_T[, v_1] + \beta_2 X_T[, v_2] + \beta_3 X_T[, v_3] + \beta_4 X_T[, v_4])^b$ 
  else
    trueY +=  $\nu \text{abs}(a + \beta_1 X_T[, v_1] + \beta_2 X_T[, v_2] + \beta_3 X_T[, v_3] + \beta_4 X_T[, v_4])^b$ 
  end if
next i
observed response: measuredY = trueY + 10 * NID(10000)
NID( $k$ ) generates a vector of length  $k$  from standard normal distribution.
 $\pm$  means that the sign is generated randomly.

```

---

**Figure 1.** Algorithm used for the generation of simulated data sets.

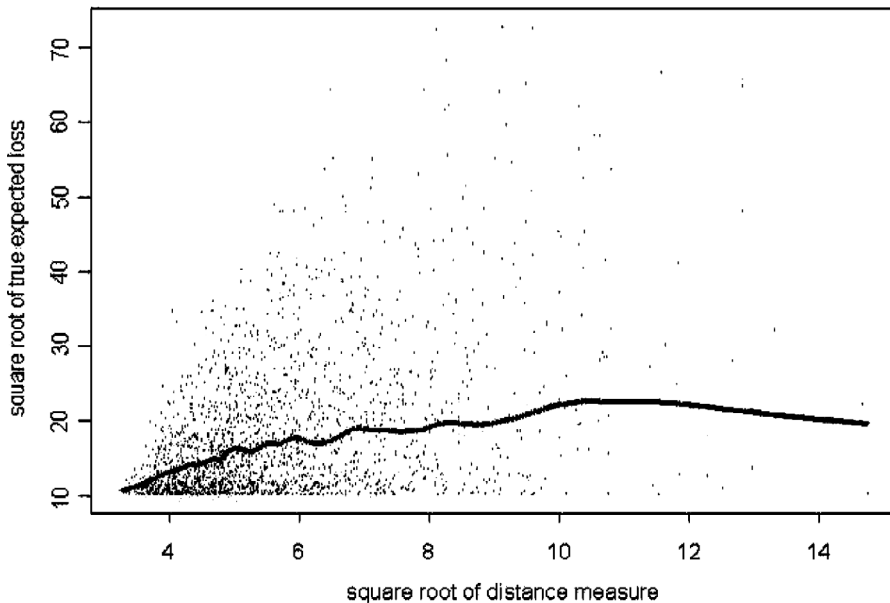
We evaluated the correlation between the distance measure and the true expected loss using simulated data sets. The simulated data sets were intended to represent a range of data sets which could arise from an industrial production process. The observations occurred in clusters of different sizes, and the input variables were correlated. The true regression function was a sum of 24 randomized effects

$$f(x) = \sum_{j=1}^{24} v_j |a_j + \beta_{1j}x_{(1)} + \beta_{2j}x_{(2)} + \cdots + \beta_{16j}x_{(16)}|^{b_j} s_j. \quad (5.2)$$

The parameters  $v_j, a_j, \beta_{ij}, b_j > 0$ , and  $s_j \in \{1, \text{sign}(a_j + \beta_{1j}x_{(1)} + \beta_{2j}x_{(2)} + \cdots + \beta_{16j}x_{(16)})\}$  were generated randomly for each effect,  $j$ . Only 1, 2, 3, or 4 of  $\beta_{ij}$  differs from 0  $\forall j = 1, \dots, 24$ , so that at most 4th order interactions were allowed. One simulated data set consisted of 10,000 observations and 16 input variables. The algorithm used for data generation is given in Fig. 1.

We simulated 20 data sets. We split all the simulated data sets randomly into a learning data set and a validation data set. Out of the 2,000 observations in the validation data, we calculated the distances to the learning data set using the proposed method. For each data set, we fitted an additive model with univariate thin plate regression splines as basis functions to define the  $\alpha$ -coefficients of our distance measure. We defined the true pairwise distance as the true expectation  $E(\mu_0 - y_i)^2$  and the true distance to the training data as the true expected squared prediction error

$$E \left( \mu_0 - \frac{\sum_{i=1}^k y_i / d(x_i, x_0)}{\sum_{i=1}^k 1/d(x_i, x_0)} \right)^2. \quad (5.3)$$



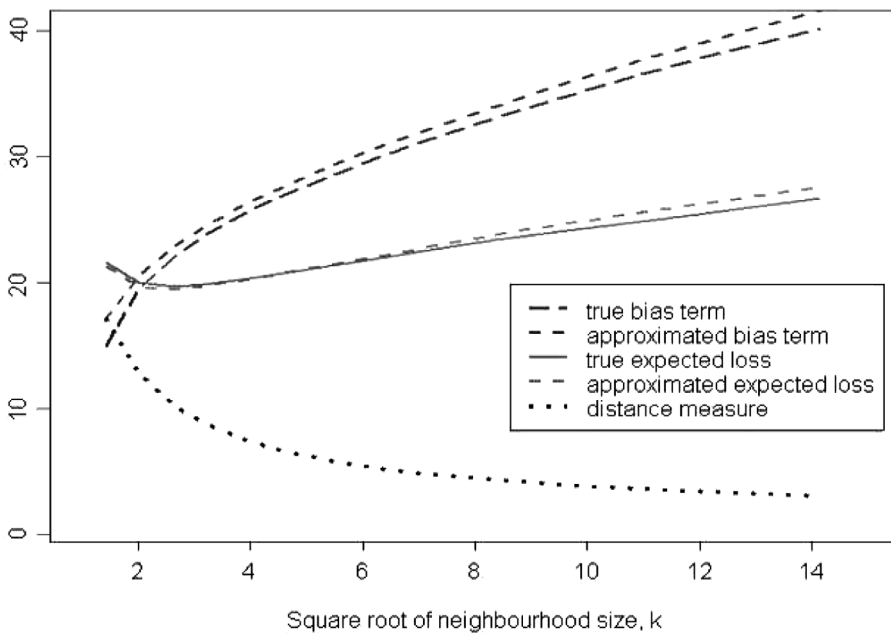
**Figure 2.** True expected loss plotted against the distance measure in a simulated data set. The black line is the smoothed average of the expected loss.

We examined the accuracy of our pairwise distance measure  $d(x_0, x_i)$  in approximating the expected loss  $E(\mu_j - y_i)^2$  based on the correlations between the pairwise distance measure  $d(x_i, x_j)$  (Eq. (2.15)) and its theoretical reciprocal  $(\mu_i - \mu_j)^2 + \sigma^2$ . In the simulated data sets, the correlation varied between 0.23 and 0.54, the average correlation being 0.34. In every simulation the average expected loss was about 21% smaller than the average pairwise distance measure.

We used neighborhood size  $k = 30$  in the distance measure. The correlation between the distance measure Eq. (3.12) and the true expected squared error  $E_Y(\mu_0 - \hat{y}_0)^2$  ( $\hat{y}_0$  is defined in Eq. (3.2)) varied between 0.22 and 0.51, the average correlation being 0.42. Our distance measure  $d(x_0, T)$  acceptably reflects its theoretical reciprocal, (Eq. 5.3), the expected squared error loss when  $x_0$  is predicted on the basis of  $T$  using the distance-weighted  $k$ -nearest-neighbor method (Fig. 2).

The deviation between the true expected squared error and our distance measure is mainly the consequence of the difficulty in approximating pairwise expected loss based only on  $x$ . If the true pairwise expected losses were known and used as the pairwise distances,  $d_i = E(y_i - \mu_0)^2$ , the approximation would work much better: The correlation between the true expected loss  $E(y_0 - \hat{y}_0)^2$  and the harmonic sum of the true pairwise distances  $1/\sum_{i=1}^k \frac{1}{d_i}$  was typically about 0.93 and over 0.83 in all the simulated data sets for  $k = 30$ .

Simulation studies showed that the approximation shown in Eq. (3.10) holds well in practice: In all of the simulated data sets, the correlation between the harmonic sum Eq. (3.9) and the bias term Eq. (3.4) was over 0.99 when pairwise distances Eq. (2.15) depending only on  $x$  were used, and over 0.92 when the true



**Figure 3.** Averages of the bias term Eq. (3.4), true expected loss Eq. (3.3), their approximations Eq. (3.10) and Eq. (3.11) and the distance measure Eq. (3.12) as functions of  $k$ .

distances  $d_i = E(\mu_0 - y_i)^2$  were used and  $k = 30$ . Also, the average of approximation was near the average of the true bias term (Fig. 3).

The approximation Eq. (3.11) explained about 38% of the variance of the true expected loss. The values of  $\delta^2$  giving the best approximation varied between the simulated data sets and also depending on  $k$ . For  $k = 30$ , the average estimated  $\delta^2$  was 0.51 and the estimated values of  $\delta^2$  varied between 0.37 and 0.66. In Fig. 3, the average of the approximation is compared with the average of the true expected loss. Note that all of the quantities presented with lines are on a square root scale.

The size of the neighborhood had a relatively small effect on the results, and all alternatives between  $k = 5$  and  $k = 500$  gave satisfactory correlations, and the best size of the neighborhood varied greatly between the simulation runs. On average, a neighborhood size around  $k = 30$  worked best. No larger neighborhood than  $k = 30$  was needed for the  $k$ -nearest-neighbor prediction in the simulated data sets.

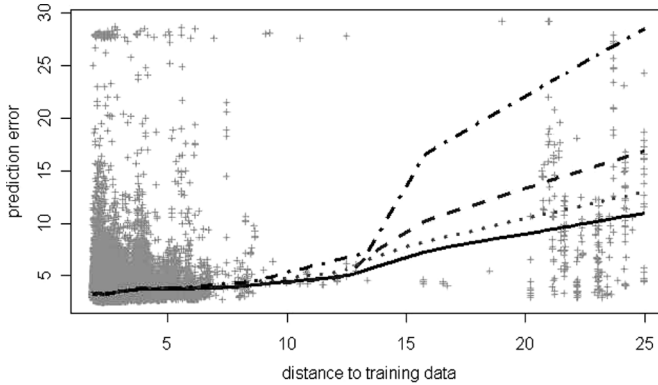
We compared our distance measure to the sum of pairwise distances of Angiulli and Pittuzi (2005). Using  $k = 30$ , our distance measure was slightly better in 90% of the simulation runs, and the average difference in correlation was 0.035. We also examined the effect of the method on defining  $\alpha$ -coefficients for the distance measure. The average correlation between the pairwise distances based on a fitted additive model and on the true response surface was 0.92. The correlations between distances calculated on the basis of two different learning methods were around 0.95, which means that the model fitting had only a small effect on the results. Also, changes in the error variance estimate  $\alpha_0$  did not decrease the correlations significantly. The method of Mahamud and Hebert (2003) for specifying  $\alpha$ -coefficients gave poor results: The average pairwise correlation was much smaller.

## 6. Applications

We applied the distance measure to real industrial process data. We used a training data set having 90,000 observations, 26 continuous input variables, and 6 binary input variables without any computational problems. In the test data set containing 60,000 observations, the average prediction error increased along with the distance from the training data (Fig. 4). The correlations between the measured loss and the distance measure were between 0.25 and 0.5, depending on the response variable and the prediction model.

There are several potential applications for the proposed measure between a single query point or observation and a training data set. The distance measure can be used in assessing the reliability of prediction, in novelty detection, data selection, outlier detection, and model selection.

When prediction using novel input values is needed, there rises the question of whether the model gives a reliable prediction or not. If the query point has enough training data instances nearby, the prediction can be kept reliable. If the query point is far away from the training data instances, the model will give a poor prediction with high probability. The information about model accuracy is especially important on the boundaries of the training data set. The distance between the query point and the training data set gives information about the uncertainty of the prediction. The prediction accuracy of the model for test data observations distant from the training data is related to the interpolation ability of the model. The relationship



**Figure 4.** Prediction error plotted against distance from the training data set in the real data set. The lines are the smoothed medians for four different prediction models.

between the distance measure and the accuracy of the model on the boundaries of the data set has been studied by Juutilainen et al. (2005).

The distance measure could be used for model selection to find a model performing better at the boundaries of the data. The distances from the observations in the validation data set,  $V$ , to the training data set,  $T$ , can be calculated and utilised in model selection. We suggest that a model selection criterion giving more weight to validation data observations distant from the training data set might result in a model with better interpolation capability at the boundaries of the training data. We tested the proposed model selection approach in the 20 simulated data sets, each divided between training ( $n = 6,000$ ), validation ( $n = 2,000$ ), and test ( $n = 2,000$ ). We used a distance weighted validation error, giving more weight to the 25% of validation data observations with the largest distance to the training data:

$$\sum_{i \in V} (1 + I[d(x_i, T) > \mathcal{D}_{75}] + 2I[d(x_i, T) > \mathcal{D}_{95}])(y_i - \hat{y}_i)^2, \tag{6.1}$$

where  $I(d > \mathcal{D}_a) = 1$  if  $d > ath$  percentile of  $\{d(x_i, T) \mid x_i \in V\}$ , otherwise  $I(d > \mathcal{D}_a) = 0$ . We selected the model from the framework of quadratic regression models using a forward stepwise algorithm. The model selection criterion Eq. (6.1) was compared with the usual validation error  $\sum_{i \in V} (y_i - \hat{y}_i)^2$ . The novel model selection

**Table 1**

Average squared prediction error in the 20 test data sets, each split in two based on the distance to training data

Model selection criterion	Distance to training data			
	small (90% of data)		large (10% of data)	
	mean	median	mean	median
Distance weighted validation error	137	125	429	337
Average validation error	139	117	444	356

criterion performed better in predicting the observations distant from the training data, but the difference was not statistically significant: The  $p$ -value against the equality of means ( $t$ -test) was  $p = 0.17$ , and  $p = 0.20$  against the equality of medians (Wilcoxon signed rank) (Table 1).

The distance measure could be applied to processing of data. Outliers, measurement errors, and novelties can be found as observations whose distance to the data set is exceptionally large. The distance measure could be applied to restrict the excessively large number of data: The observations distant from the data are important for modeling, but plenty of observations with a small distance to the data can be discarded without losing much information.

## 7. Conclusion

We proposed a novel distance measure for the distance between a data set and a single observation. The proposal is constructed using an approximation of the optimal distance measure of the  $k$ -nearest-neighbor regression. The distance measure reflects the expected squared error loss when the single observation is predicted on the basis of the data set using the distance-weighted  $k$ -nearest-neighbor method. The proposed approach to measuring the distance by the expected loss provides clear advantages compared to the existing approaches: The location and exceptionality of observations are interpreted on the basis of their significance in predicting the response variable, and the proposed distance measure provides information about model uncertainty at query points without any assumption about the distribution of inputs.

## References

- Angiulli, F., Pittuzi, C. (2005). Outlier mining in large high-dimensional data sets. *IEEE Trans. Knowledge Data Eng.* 17:203–215.
- Friedman, J. H., Bentley, J. L., Finkel, R. A. (1977) An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Mathemat. Software* 3:209–226.
- Goldberg, L. R., Kercheval, A. N., Lee, K. (2005)  $t$ -Statistics for weighted means in credit risk modelling. *J. Risk Fin.* 6:349–365.
- Juutilainen, I., Röning, J., Laurinen, P. (2005) A study on the differences in the interpolation capabilities of models. Proc. 2005 IEEE Mid-Summer Workshop on Soft Computing in Industrial Applic. (SMCia/05), pp. 202–207.
- Kallenberg, O. (1997). *Foundations of Modern Probability*. New York: Springer-Verlag.
- Kaufman, L., Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.
- Mahamud, S., Hebert, M. (2003). Minimum risk distance measure for object recognition. Proc. Ninth IEEE Int. Conf. Comput. Vision (ICCV), pp. 242–248.
- Markou, M., Singh, S. (2003). Novelty detection: a review. *Signal Process.* 83:2481–2521.
- Wettschereck, D., Aha, D. W., Mohri, T. (1997). Review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Rev.* 11:273–314.