Estimating Conditional Event Probabilities with Mixed Regressors: aWeighted Nearest Neighbour Approach

Mahfuza Khatun[1] | *Jahangirnagar University, Dhaka, Bangladesh*
Sikandar Siddiqui[2] | *Deloitte Audit Analytics, Frankfurt, Germany*

Received 29.9.2022, **Accepted** (reviewed) 21.11.2022**, Published** 16.6.2023

Abstract

The *k*-Nearest Neighbour method is a popular nonparametric technique for solving classification and regression problems without having to make potentially restrictive a priori assumptions about the functional form of the statistical relationship under investigation. The purpose of this paper was to demonstrate that the scope of this method can be extended in a way that enables the simultaneous consideration of continuous, ordered discrete, and unordered discrete explanatory variables. An exemplary application to a publicly available dataset demonstrated the feasibility of the proposed approach.

Keywords	***DOI***	***JEL code***
Qualitative choice models, nonparametric estimation	*https://doi.org/10.54694/stat.2022.45*	*C14, C25*

INTRODUCTION

The *k*-nearest neighbour method is a common statistical technique for solving classification problems (see, e.g., Fix and Hodges, 1951). In the following, the symbol Y either represents either a scalar, discrete random variable or a vector of random variables indicating the class to which a given statistical entity belongs. Moreover, the quantity Q represents a sample vector of explanatory variables to be used as inputs for the classification procedure. A sample consisting of N independent statistical entities $i = 1, \ldots, N$, each of which is characterized by a pair of realisations $\{y_i; q_i\}$ of Y and Q is also assumed to be given. Then the *k*-nearest neighbour procedure essentially consists of approximating the function representing the conditional expectation of Y for a given value Q^* of Q by:

- either calculating a locally weighted average of the y_i values of those entities where the associated realisations q_i of Q are among the k closest neighbours of Q^*,

[1] Jahangirnagar University, Savar, 1342 Dhaka, Bangladesh.
[2] Deloitte Audit Analytics GmbH, Europa-Allee 91, 60486 Frankfurt, Germany. Corresponding author: e-mail: siddiqui@web.de, phone: (+49)1718881667.

- or performing a locally weighted linear regression of Y on Q, in which only those entities in the sample are assigned a positive weight if they are among the k closest neighbours of Q^*.

In doing so, each of the k closest neighbours can either be assigned the same weight, or each of them separately can be assigned a weight that declines with increasing distance from Q^*. Both of these approaches have been nicely summarized, for example, by Chen, Härdle and Schulz (2004), and Cleveland and Loader (1995).

Typically, the k-nearest neighbour techniques requires all the conditioning variables included in Q to be continuous. This condition might easily be seen as too restrictive when there are either ordered discrete variables (like, e.g. school grades or credit ratings), or unordered discrete variables (e.g. information on the sectoral or geographical affiliation of a company). Hence, the need may arise to introduce a more general measure of the distance between two statistical entities in which all of these variable types can be included in an intuitively plausible manner. This is the main purpose of this paper. It hence proceeds as follows: Section 1 seeks to clarify how the distance between two statistical entities can be calculated if each of them is characterised by a specific realisation of such a mixture of continuous, ordered discrete and unordered discrete variables. The application of the proposed procedure in the context of binomial or multinomial classification problems is described in Section 2. This is also where a measure of the out-of-sample predictive accuracy and a selection rule for the number of neighbours, k, is given. Section 3 presents an application. Last section concludes.

1 MEASURING THE DISTANCE BETWEEN PAIRS OF ENTITIES

1.1 Starting point

Let there be a sample of N entities (e.g. companies) $i = 1, \ldots, N$, each of which is characterised by a tuple of characteristics $q_i = \{ x_i, v_i, z_i \}$. Here,

- x_i is an entity-specific realisation of a $(K_1 \times 1)$ vector X of continuous variables, (e.g. the total revenue of a firm in a given year, or its total assets at a given point in time, etc.),
- v_i is an entity-specific realisation of a $((K_2 - K_1) \times 1)$ vector V of ordered discrete variables (like, e.g., a firm's credit rating, or an analyst recommendation ["Strong Buy", "Buy", "Hold" or "Sell"]), numbered consecutively in steps of 1, and
- z_i is an entity-specific realisation of a $((K_3 - (K_2 + K_1)) \times 1)$ vector Z of unordered, discrete variables (like, e.g., a firm's ISO country-of-residence identifier or its NAICS industry classification code).

1.2 Pairwise distance based on continuous characteristics

The distance between two distinct entities i and j with respect to the values taken by X can be measured by Gower's (1971) measure:

$$d_x(i,j) = \sum_{s=1}^{K_1} \frac{\left| x_{i,s} - x_{j,s} \right|}{r(X_s)}, \tag{1}$$

where $x_{i,s}$ denotes the s'th element of the $(K_1 \times 1)$ vector x_i, and $r(X_s)$ is the range (sample maximum minus sample minimum) of the values of X_s observed.

In principle, the above Formula (1) is only one out of several distance measures that could be applied to measure the degree of dissimilarity prevailing between realisations of continuous random vectors; see, e.g., Weller-Fahy, Borghetti, and Sodemann (2015) for a theoretically substantiated overview. The motivation for choosing this particular measure is that it facilitates the aggregation of distance measures for the different variable types involved, as will be demonstrated below.

1.3 Pairwise distance based on ordered discrete characteristics

Likewise, the distance between i and j with respect to the values taken by the components of V can be measured by:

$$d_v(i,j) = \sum_{s=K_1+1}^{K_1+K_2} \frac{\left| v_{i,s} - v_{j,s} \right|}{m_s}, \tag{2}$$

where m_s is the number of possible realisations of the s-th ordered discrete variable.

A rationale for this way of proceeding can be given as follows: Assume that that variable V_s is a rating grade, measured in equally spaced steps of unit length from 1 (= best possible outcome) to m_s (= worst possible outcome). Then, the actual realisation taken by V_s can be assumed to be dependent on the value taken by a latent (=unobservable) variable as follows:

$$v_{i,s} = j \text{ if } v^*_{i,s} \in \left]\tfrac{j-1}{m_s}, \tfrac{j}{m_s}\right] \text{ with } j \in \{1, 2, ..., m_s\}. \tag{3}$$

The distance between the mid-points of two neighbouring intervals inside the range of is V^*_s is $1/m_s$.

1.4 Pairwise distance based on unordered discrete characteristics

Finally, the distance between i and j with respect to the values taken by the components of Z can be measured by their separateness, or lack of overlap (see, e.g., Stanfill and Waltz, 1986), based on the Hamming Distance (Hamming, 1950):

$$d_z(i,j) = \sum_{s=K_1+K_2+1}^{K_1+K_2+K_3} \frac{I\left(z_{i,s} \neq z_{j,s}\right)}{m_s}. \tag{4}$$

In Formula (4), $I(\cdot)$ denotes an indicator function that takes the value 1 if the condition in brackets holds true, and is set to zero otherwise, and m_s is the number of distinct possible realisations of the s-th unordered discrete variable.

The rationale underlying this way of proceeding can be explained as follows: the m_s possible, yet mutually exclusive, realisations of a single, discrete unordered variable of the Z_s can be recoded as a vector of m_s auxiliary, binary variables A_1 to $A_{m(s)}$. This technique of representing unordered categorical data, which is referred to as one-hot encoding, is explained in an exemplary manner in Table 1 (for the special case of $m_s = 4$).

Table 1 Representing unordered categorical data via one-hot encoding

	Values taken by the auxiliary variables A_1 to $A_{m(s)}$			
Variables taken by original variable Z	A_1	A_2	A_3	A_4
'Tinker'	1	0	0	0
'Tailor'	0	1	0	0
'Soldier'	0	0	1	0
'Spy'	0	0	0	1

Source: Own construction

The distance prevailing between two specific realisations $z_{i,s}$ and $z_{j,s}$ of Z_s with respect to the values $a_{i,g}$ and $a_{j,g}$ taken by the p-th of the m_s auxiliary variables can thus be measured using the same latent variable approach as sketched in (3), which comes down to:

$$Distance\,between\,z_{i,s}\ \ and\,z_{j,s}\ with\,regard\,to\,A_g = \frac{I\left(a_{i,p} \neq a_{j,p}\right)}{2}. \tag{5}$$

The overall distance between $z_{i,s}$ and $z_{j,s}$ can then be calculated as the average of the distances (5), calculated across all values 1, …, m_s of the auxiliary variable index g:

$$\begin{aligned} &Distance\,between\,z_{i,s}\ \ and\,z_{j,s} \\ &= \left(\frac{1}{m_s}\right) \cdot \sum_{g=1}^{m_s} Distance\,between\,z_{i,s}\ \ and\,z_{j,s}\ with\,regard\,to\,A_g, \\ &= \left(\frac{1}{m_s}\right) \cdot \sum_{g=1}^{m_s} \frac{I\left(a_{i,g} \neq a_{j,g}\right)}{2}, \\ &= \left(\frac{1}{m_s}\right) I\left(z_{i,s} \neq z_{j,s}\right). \end{aligned} \tag{6}$$

The above normalisation will cause the distance between two different realisations of an unordered discrete variable (e.g. country of residence) to take the same value as the distance between two distinct yet immediately adjacent realisations of an ordered discrete variable having the same number of possible realisations. Using m_s in the denominator of Formula (4) will cause the average perceived distance between pairs of observations with different values of Z_s to shrink as the number of different realisations of this variable increases. This is well in line with intuition because *ceteris paribus*, the more fine-grained a classification scheme with discrete, unordered categories becomes, the greater the average pairwise similarity between two entities assigned to different classes will tend to become.

1.5 Overall pairwise distance between two entities

The overall distance score between two entities i and j can then be calculated by summing up the distance measures for the different types of variables given in (1), (2), and (4):

$$d(i,j) := d_x(i,j) + d_v(i,j) + d_z(i,j)\,. \tag{7}$$

2 ESTIMATION PROCEDURE AND GOODNESS-OF-FIT ASSESSMENT

2.1 Assigning estimated conditional probabilities to possible outcomes

The following setting applies to a situation where we have a discrete dependent variable Y with a finite number of possible values $g = \{1, 2, \ldots G\}$ (which may, but need not, be ordered). Then, if we have an observed or hypothetical entity characterized by the tuple $q^* = \{ x^*, v^*, z^* \}$, we can estimate the conditional probability $\Pr(Y = g \mid Q = q^*)$ using the k-nearest neighbour principle by applying the following sequence of steps:

(i) For each entity i in the sample, calculate the distance score with respect to the entity under consideration as:

$$d\left(i, *\right) = \sum_{s=1}^{K_1} \frac{\left|x_{i,s} - x_s^*\right|}{r\left(x_s\right)} + \sum_{s=K_1+1}^{K_1+K_2} \frac{\left|v_{i,s} - v_s^*\right|}{m_s} + \sum_{s=K_1+K_2+1}^{K_1+K_2+K_3} \frac{I\left(z_{i,s} \neq z_s^*\right)}{m_s}. \tag{8}$$

(ii) Sort the object-specific dissimilarity scores obtained in (i) in ascending order. Let $\tilde{d}_k(*)$ denote the k-th score in this ascending sequence, and $\tilde{d}_{k+1}(*)$ denote the smallest value of $d(i,*)$ that exceeds $\tilde{d}_k(*)$.

(iii) Assign to each object in the sample a weighing factor $\tilde{w}(i,*)$ that is calculated as follows:

$$\tilde{w}(i,*) := \begin{cases} 1-\left(\dfrac{d(i,*)}{\tilde{d}_{k+1}(*)}\right) \text{ if } d(i,*) < \tilde{d}_{k+1}(*) \\ 0 \text{ otherwise} \end{cases}. \tag{9}$$

(iv) Transform the outcomes of step (iii) into a normalized set of weighting factors $w(i,*)$ that sum up to 1:

$$w(i,*) = \frac{\tilde{w}(i,*)}{\sum_{j=1}^{N}\tilde{w}(j,*)}. \tag{10}$$

(v) Let , $\check{p}(g \mid q^*) := \max\left\{\min\left\{x'\hat{\beta}; 1\right\}; 0\right\}$, (11)

where $\hat{\beta}^* := argmin_{\tilde{\beta}} \sum_{i=1}^{N} w(i,*)\left[I(y_i = g) - x_i'\tilde{\beta}\right]^2$ denotes the locally linear least squares estimated of the conditional probability of ($Y = g$) given ($q = q^*$). If the dependent variable Y has only two possible values (i.e. $G = 2$, and $g \in \{1; 2\}$), the concluding estimates of the related conditional probabilities can be set to $\widehat{Pr}(Y = 1 \mid q = q^*) = \check{p}(1 \mid q^*)$ and $\widehat{Pr}(Y = 2 \mid q = q^*) = 1 - \check{p}(1 \mid q^*)$, respectively.

(vi) Whenever the dependent variable Y has more than two possible values (i.e. $G > 2$), it cannot be taken for granted that the estimated conditional probabilities $\check{p}(g \mid q^*)$ from step (v) sum up to unity. In this case, a normalisation needs to be applied to these provisional estimates, which amounts to setting the concluding estimate of $Pr(Y = g \mid q = q^*)$ to:

$$\widehat{Pr}(Y = g \mid q = q^*) = \frac{\check{p}(g \mid q^*)}{\sum_{c=1}^{G}\check{p}(c \mid q^*)}. \tag{12}$$

2.2 Out-of-sample goodness of fit assessment and choice of the number of neighbours

In order to assess the predictive reliability of the proposed method, and to choose a favourable value of number of neighbours, k, to be used, the leave-one-out log likelihood function can be used. It can be calculated as follows:

(1) For each observation in the sample, and for each of the possible values of g, calculate the quantity $\hat{p}_{-i}(g \mid q_i)$, i.e. the estimated conditional probability:

$$\widehat{Pr}(Y = g \mid q = q_i),$$

that is obtained by applying the above procedure to a sample from whichthe observation with index i has deliberately been omitted.

(2) Calculate the leave-one-out log likelihood function as:

$$\sum_{i=1}^{N}\sum_{g=1}^{G} I\left(y_i = g\right) \cdot \ln(\hat{p}_{-i}\left(g \mid q_i\right)).$$

The optimum value of k could then be equated with the one that maximizes the quantity given under point (2) above.

3 APPLICATION

3.1 Data

The dataset chosen for our sample application is the Automobile Data Set from 1985 Ward's Automotive Yearbook, donated by Jeffrey C. Schlimmer[3] and freely available online,[4] where also more detailed information on the informational content of the variables of interest can be found. It contains $N = 205$ rows and 26 columns, the first of which is a relative risk score compared to equally priced vehicles. Theoretically, it ranges from (–3 = very safe) to (3 = very risky) in steps of 1, but the best score found in the sample was (–2). From the remaining 25 columns, the 20 summarized in Table 1 were used as explanatory variables. Ten entities were removed from the dataset used for estimation due to missing values.

Table 2 1985 Automobile Data Set: variables in use

Name	Description
Risk score	Dependent variable, ranging from –3 (safest) to 3 (most risky) in steps of 1
Make	Brand name of manufacturer
Fuel-type	Diesel or gas
Aspiration	Standard or turbo
Num-of-doors	Two or four
Drive-wheels	4wd, front-wheel or rear-while drive
Engine-location	Front or rear
Wheel-base	Continuous
Length	Continuous
Width	Continuous
Height	Continuous
Curb-weight	Continuous
Num-of-cylinders	Integer
Engine-size	Continuous
Bore	Continuous
Stroke	Continuous
Compression-ratio	Continuous
Horsepower	Continuous
Peak-rpm	Continuous
City-mpg	Continuous
Highway-mpg	Continuous

Source: Automobile Data Set from 1985 Ward's Automotive Yearbook, donated by Jeffrey C. Schlimmer, available at: <*https://archive.ics.uci.edu/ml/datasets/automobile*>

[3] Jeffrey.Schlimmer@a.gp.cs.cmu.edu.

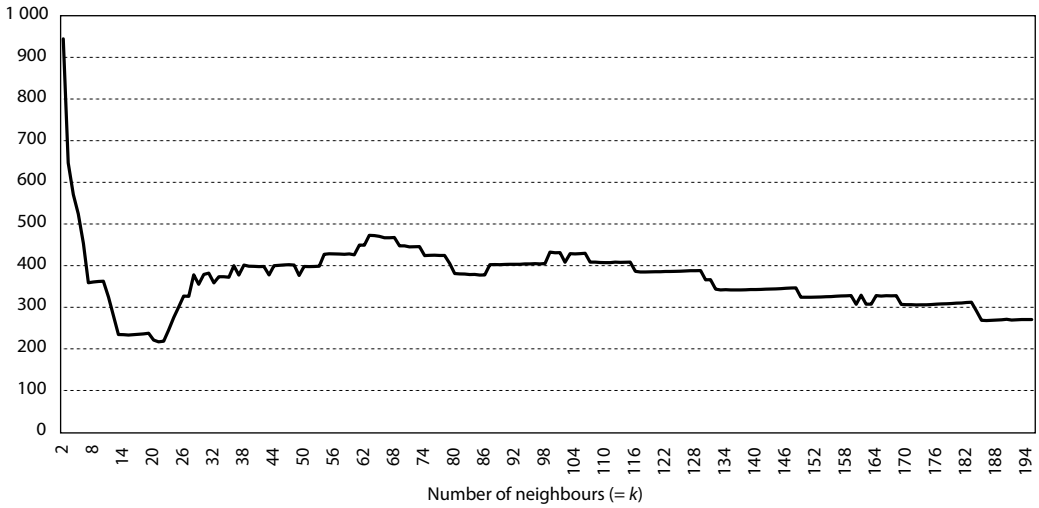
[4] <*https://archive.ics.uci.edu/ml/datasets/automobile*>.

In this example, only continuous, integer and unordered discrete explanatory variables have been used, but it should nevertheless be suitable for demonstrating the general usefulness of the proposed way of proceeding.

3.2 Results obtained with the for the *k*-Nearest Neighbour procedure

In order to avoid overfitting, the leave-one out approach sketched in Section 2.2 was employed. A related grid search procedure, the results of which are given in Figure 1, yielded an optimal value of *k* equal to 21.

Figure 1 Negative log-likelihood function for leave-one-out estimates for different values of *k*



Source: Own calculation

Assuming that the predicted value of the dependent variable is always the one to which the model has assigned the highest conditional probability, the comparison of actual and predicted values results from applying the *k*-Nearest Neighbour Procedure with $k = 21$ to the entire dataset (Table 3).

Table 3 *k*-Nearest Neighbour estimation results: actual vs. predicted outcomes

		Predicted						
	Indicator	**–2**	**–1**	**0**	**1**	**2**	**3**	**Σ**
Actual	–2	0.0000%	1.5385%	0.0000%	0.0000%	0.0000%	0.0000%	**1.5385%**
	–1	0.0000%	7.6923%	3.0769%	0.5128%	0.0000%	0.0000%	**11.2821%**
	0	0.0000%	0.5128%	31.2821%	1.0256%	0.0000%	0.0000%	**32.8205%**
	1	0.0000%	0.5128%	1.5385%	23.5897%	0.5128%	0.5128%	**26.6667%**
	2	0.0000%	0.0000%	1.0256%	3.0769%	11.7949%	0.0000%	**15.8974%**
	3	0.0000%	0.0000%	0.0000%	1.0256%	0.5128%	10.2564%	**11.7949%**
	Σ	**0.0000%**	**10.2564%**	**36.9231%**	**29.2308%**	**12.8205%**	**10.7692%**	
% correctly predicted = 84.6154%								

Source: Own calculation

The geometric mean of model-generated ex-post probability estimates for the actually observed values of Y, which is calculated as

$$\prod_{i=1}^{N}\prod_{g=1}^{G}\widehat{Pr}(Y=g \mid Q=q_i)^{I(y_i=g)}$$

equals 56.31%.

3.3 Comparison with an Ordered Probit Model

In order to assess the predictive performance of the proposed approach, we compared the nonparametric model advocated here to the more commonly used, parametric Ordered Probit model (see, e.g. Greene, 2003, chapter 21.8). Results obtained for the Ordered Probit are given in Table 4.

Table 4 Ordered Probit Model: actual vs. predicted outcomes

		Predicted						
	Indicator	**–2**	**–1**	**0**	**1**	**2**	**3**	**Σ**
Actual	–2	0.5128%	1.0256%	0.0000%	0.0000%	0.0000%	0.0000%	**1.5385%**
	–1	0.5128%	7.6923%	3.0769%	0.0000%	0.0000%	0.0000%	**11.2821%**
	0	0.0000%	0.5128%	29.2308%	3.0769%	0.0000%	0.0000%	**32.8205%**
	1	0.0000%	0.0000%	4.1026%	20.0000%	1.0256%	1.5385%	**26.6667%**
	2	0.0000%	0.0000%	0.5128%	5.6410%	6.1539%	3.5897%	**15.8974%**
	3	0.0000%	0.0000%	0.0000%	1.5385%	1.5385%	8.7180%	**11.7949%**
	Σ	**1.0256%**	**9.2308%**	**36.9231%**	**30.2564%**	**8.7180%**	**13.8462%**	
% correctly predicted = 72.3077%								

Source: Own calculation

In this case, we obtain a geometric mean of model-generated ex-post probability estimates for the actually observed values of Y that amounts to 50.41%.

CONCLUSION

The key advantage that nonparametric classification methods have over parametric models is that they do not require any assumptions about the form of the function that translates the values of the explanatory variables into conditional probabilities of the possible outcomes. Against this background, the purpose of this paper was to demonstrate that the scope of the standard k-Nearest Neighbour method can be extended in a way that enables the simultaneous consideration of continuous, ordered discrete, and unordered discrete explanatory variables. An exemplary application to a publicly available dataset demonstrated the feasibility of the proposed approach.

ACKNOWLEDGMENT

We are indebted to two anonymous referees and this journal's managing editor, whose helpful comments have greatly improved this paper.

The project underlying this publication was funded by the German Federal Ministry of Economic Affairs and Climate Action under project funding reference number 01MK21002G.

References

ALTMAN, N. S. (1991). *An Introduction to Kernel and Nearest Neighbor Nonparametric Regression*. Cornell University, Ithaca, NY14853: Biometrics Unit.

CHEN, Y., HÄRDLE, W. K., SCHULZ, R. (2004). *Prognose mit nichtparametrischen Verfahren*. Humboldt-Universität Berlin: Center for Applied Statistics and Economics.

CLEVELAND, W. S., LOADER, C. (1995). Smoothing by local regression: Principles and methods. In: HÄRDLE, W. K., SCHIMEK, M. G. (eds.) *Statistical Theory and Computational Aspects of Smoothing*, Proceedings of the COMPSTAT '94 Satellite Meeting held in Semmering, Austria, 27–28 August: 10–49.

FIX, E., HODGES, J. L. (1951). *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*. Randolph Field, Texas: USAF School of Aviation Medicine.

GOWER, J. C. (1971). A general coefficient of similarity and some of its properties [online]. *Biometrics,* 27: 623–637. <https://doi.org/10.2307/2528823>.

GREENE, W. H. (2005). *Econometric Analysis*. Upper Saddle River (New Jersey): Prentice Hall.

HAMMING, R. W. (1950). Error detecting and error correcting codes [online]. *Bell System Technical Journal*, 29(2): 147–160. <https://doi.org/10.1002/j.1538-7305.1950.tb00463.x>.

STANFILL, C., WALTZ, D. (1986). Toward memory-based reasoning [online]. *Communications of the ACM*, 29(12): 1213–1228. <https://doi.org/10.1145/7902.7906>.

STONE, C. J. (1977). Consistent Nonparametric Regression [online]. *Annals of Statistics,* 5(4): 595–645. <https://doi.org/10.1214/aos/1176343886>.

WELLER-FAHY, D., BORGHETTI, B. J., SODEMANN, A. A. (2015). Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection. *IEEE Communication Surveys & Tutorials*, 17 (1): 70–91. <https://doi.org/10.1109/ COMST.2014.2336610>.