K-nearest neighbors as pricing tool in insurance

Raul Kangro and Kalev Pärna

University of Tartu, Estonia, email: raul.kangro@ut.ee, kalev.parna@ut.ee

Keywords: curse of dimensionality, distance measures, feature selection, k-nearest neighbors, local regression, premium calculation, supervised learning.

The method of k-nearest neighbors (k-NN) is recognized as a simple but powerful toolkit in statistical learning [1], [2]. It can be used both in discrete and continuous decision making known as classification and regression, respectively. In the latter case the k-NN is aimed at estimation of conditional expectation $y(\mathbf{x}) := E(Y|X = \mathbf{x})$ of an output Y given the value of an input vector $\mathbf{x} = (x_1, \ldots, x_m)$. In accordance with supervised learning set-up, a training set is given consisting of n pairs $(\mathbf{x_i}, y_i)$ and the problem is to estimate $y(\mathbf{x})$ for a new input $\mathbf{x}$. This is exactly the situation in insurance where the pure premium $y(\mathbf{x})$ for a new client (policy) $\mathbf{x}$ is to be found as conditional mean of loss. Typically the data do not contain any other record with the same $\mathbf{x}$, thus the other data points have to be used in order to estimate $y(\mathbf{x})$. Using the k-NN methodology, one first finds a neighborhood $U_{\mathbf{x}}$ consisting of k samples which are nearest to $\mathbf{x}$ w.r.t a given distance measure d. Secondly, the (weighted) average of Y is calculated over the neighborhood $U_{\mathbf{x}}$ as an estimate of $y(\mathbf{x})$:

$$\hat{y}(\mathbf{x}) := \frac{1}{\sum_{i \in U_{\mathbf{x}}} \alpha_i} \sum_{i \in U_{\mathbf{x}}} \alpha_i \cdot Y_i,$$

where the weights α_i are chosen so that the nearer neighbors contribute more to the average than the more distant ones. We use the distance between the instances $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ in the form

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{m} w_j \cdot d_j(x_{ij}, x_{i'j}),$$

where w_j is the weight of the feature j and $d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$ (and a zero-one type variable for categorical features).

We address the following key issues related to k-NN method: feature weighting (w_j), distance weighting (α_i), determining the optimum value of the smoothing parameter k. We propose a three-step multiplicative procedure to define w_j which consists of 1) normalization (eliminating the scale effect), 2) accounting for statistical dependence between the feature j and Y, 3) feature selection to obtain a subset of features that performs best. All our optimization procedures are based on cross-validation techniques. The so-called 'curse of dimensionality' is effectively handled by our feature selection process which optimizes the dimension of input.

Finally, comparisons with other methods for estimation of the regression function $y(x)$ (CART, generalized linear regression, use of model distributions) are drawn, which demonstrate high competetiveness of the k-NN method. The conclusions are based on the analysis of a real data set.

References

[1] Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

[2] Mitchell, T.M. (2001). *Machine-Learning.* McGraw-Hill.