The Korean Communications
in Statistics Vol. 10 No. 1, 2003
pp. 239-246

Input Variable Importance in Supervised Learning Models

Myung-Hoe Huh[1] and Yong Goo Lee[2]

Abstract

Statisticians, or data miners, are often requested to assess the importances of input variables in the given supervised learning model. For the purpose, one may rely on separate ad hoc measures depending on modeling types, such as linear regressions, the neural networks or trees. Consequently, the conceptual consistency in input variable importance measures is lacking, so that the measures cannot be directly used in comparing different types of models, which is often done in data mining processes.

In this short communication, we propose a unified approach to the importance measurement of input variables. Our method uses sensitivity analysis which begins by perturbing the values of input variables and monitors the output change. Research scope is limited to the models for continuous output, although it is not difficult to extend the method to supervised learning models for categorical outcomes.

Key Words: Supervised Learning, Input Variable Importance, Linear Regression, Neural Network, Regression Tree, Sensitivity Analysis, Data Mining.

1. Background and Aim

Recently, supervised learning models for data mining applications are actively studied in several leading countries (Hastie, Tibshirani and Friedman, 2001; Ripley, 1996). Also in Korea, there appeared many research papers on topics related to tree methods, neural networks and data mining softwares; Kim (1996), Lee and Moon (1997), Hwang and Kim (1997), Kang, Han and Choi (2000), Lim, Lee and Chung (2001), Song and Yoon (2001), Chung, Jung and Kim (2002), Han, Kang, Lee and Lee (2002) and Lee and Song (2002), to name a few.

Importance of input variables in supervised learning models is one of very needed issues to be solved, but it is conceptually difficult to formulate, as well discussed by Sarle (1998). He classified the notion of input variable importance into "predictive importance" and "causal importance". To quote,

1) Professor, Dept. of Statistics, Korea University. Anam-Dong 5-1, Seoul 136-701, Korea. E-mail: stat420@korea.ac.kr.

2) Professor, Dept. of Applied Statistics, Chung-Ang University. HukSuk-Dong 221, Seoul 156-756, Korea. E-mail: leeyg@cau.ac.kr.

> Predictive importance is concerned with the increase in generalization error when an input is omitted from a network. Causal importance is concerned with situations where you can manipulate the values of the inputs and you want to know how much the outputs change. (Sarle, 1998)

In data mining applications, miners often want to understand the model generated by the statistical and/or data mining software, especially when the generated model is sophisticated. Additionally, at the deployment stage of models, miners often confront quite many records of which one or two input values are not available. Consequently, the model cannot be directly applicable to such records (or cases). Even though data miners may resort to imputation methods for missing data as post hoc treatment, it is strategically more valuable to prevent missing data values for "important" variables. In other words, knowing the data values of specific variables is more important than the others. Hence, we call such data miner's need by *knowledge importance* of input variables to distinguish from Sarle's predictive and causal importance.

For knowledge importance measurement, one may consider monitoring the change of output values when the input values are randomly perturbed. For important variables, small input perturbations result in big change of output values, while perturbations of unimportant input variables yield negligible differences. In such direction of thinking, we propose a sensitivity analysis procedure for input variable importance measurement in Section 2.

Our study is limited to supervised learning models for continuous output such as linear regression models, neural networks and regression trees, although the proposed method can be adapted easily to the models for categorical outcomes. We illustrate our method with a numerical example in Section 3. For partial justification of our method, it is demonstrated on a simulated data set in Section 4. In the final section, we state the differences and commonalities between our method and the one implemented in SPSS Clementine's Neural Net.

2. Input Variable Importances by Sensitivity Analysis

Suppose that we have fit a supervised learning model $\hat{y} = f(x_1, \cdots, x_p)$ with input variables $X_1, \cdots, X_p$ for output variable Y by a random data set:

$$(x_{i1}, \cdots, x_{ip}, y_i), \quad i = 1, \cdots, n$$

of which the values of x_{ij} $(j = 1, \cdots, p)$ are either continuous or categorical while those of y_i are continuous. For the importance measurement of an input variable X_p, for instance, for the given model $f(\cdot)$, we propose the following method.

Step 1: For each i $(= 1, \cdots, n)$, compute

$\hat{y}_i^* = f(x_{i1}, \cdots, x_{i,p-1}, x_{ip}^*)$ and $\hat{y}_i^{**} = f(x_{i1}, \cdots, x_{i,p-1}, x_{ip}^{**})$

where x_{ip}^* and x_{ip}^{**} are two randomly selected data values among $x_{1p}, \cdots, x_{np}$.

Step 2: Derive $d_i = \hat{y}_i^* - \hat{y}_i^{**}$ $(i = 1, \cdots, n)$ for the sensitiveness of X_p in the ith case on the output.

Step 3: Take the average of $|d_1|, \cdots, |d_n|$ and write

$$D_p = \frac{1}{n}(|d_1| + \cdots + |d_n|).$$

Step 4: For comparison yard stick, consider the average absolute difference of two random selections among observed output values: Write

$$D_y = \frac{1}{N}\{|y_1^* - y_1^{**}| + \cdots + |y_N^* - y_N^{**}|\},$$

where $y_1^*, y_1^{**}, \cdots, y_N^*, y_N^{**}$ are randomly selected values among $y_1, \cdots, y_n$. For numerical stability, it is recommended to take N much larger than n.

Step 5: Variable importance of X_p is defined by

$$Imp(X_p) = D_p / D_y.$$

Similarly, $Imp(X_1), \cdots, Imp(X_{p-1})$ of $X_1, \cdots, X_{p-1}$ can be obtained.

Therefore, importance $Imp(X_j)$ is the relative ratio of average absolute perturbation due to random substitution of X_j values while other input variable values are kept intact.

We note a primary property of the proposed method: $Imp(X_1), \cdots, Imp(X_p)$ are random quantities. The measures vary every time one executes the assessment task. To avoid the instability due to randomness, it is necessary to repeat the same procedure a number of times and take the average, especially for small data sets. For large data sets, however, such brute force approach is not practical due to extensiveness in computing. But it is also not needed since large data set size generally stabilizes the numerical outcomes. In the next section, we gives a numerical illustration of the proposed method.

3. A Numerical Example

This example concerns the fraud or false claim in government grant applications for arable development from farm owners (SPSS, 2002). Output variable is the amount of grant applied for (=Y), and input variables are declared farm income (=X1), farm size (=X2), land quality on integer scale (=X3), main crop [maze/wheat/potatoes/rapeseed] (=X4), amount of rainfall (=X5), and geographic region [midlands/north/southwest /southeast] (=X6). Four input variables, X1, X2, X3 and X5, can be considered as continuous, while two other input variables, X4 and X6, are categorical. Output variable Y is continuous. The number of input variables is six (= p).

The number of records is 121 (= n).

We consider three types of supervised learning models: General Linear Model, Neural Network and Regression Tree.

General Linear Model:

$$Y = b0 + b1 \cdot X1 + b2 \cdot X2 + b3 \cdot X3 + b4.1 \cdot X4.1 + b4.2 \cdot X4.2 + b4.3 \cdot X4.3 + b4.4 \cdot X4.4 + b5 \cdot X5 + b6.1 \cdot X6.1 + b6.2 \cdot X6.2 + b6.3 \cdot X6.3 + b6.4 \cdot X6.4 + e,$$

where X4.1, X4.2, X4.3 and X4.4 are dummy codes for X4, and, similarly, X6.1, X6.2, X6.3 and X6.4 are dummy codes for X6. To prevent redundancy, it is assumed that b4.4 = 0 and b6.4 = 0. The model is fitted using SPSS 10.0.

Neural Network:

Neural network with one hidden layer of five nodes was fitted using SPSS Clementine 7.1 (under default setting). Logistic function was used for activating nodes. The generated model is exported to calculate importance measures.

Regression Tree:

C&R Tree (Breiman, Friedman, Olshen and Stone, 1984) with tree depth 5 was fitted using SPSS Clementine 7.1 (under default setting). The generated model is exported in text codes.

We obtained Table 1 for input variable importance measures by replicating the assessment procedures of Section 2 ten times and taking the average. The average absolute difference D_y of Y was computed from $N = 10\,n$ pairs. In this paper, we define model goodness-of-fit by

$$\text{GoFit} = 1 - \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{\sum_{i=1}^{n} | y_i - \text{med}(y_1, \cdots, y_n) |}.$$

We note several facts in Table 1. First, in General Linear Model and Regression Tree, Farm Income (=X1) is dominantly important for determining the Grant (=Y). Other input variables exert smaller influences on the output. Second, in Neural Network, Farm Income (=X1), Farm Size (=X2), Land Quality (=X3) and Rainfall (=X5) are relatively more important compared to Main Crop (=X4) and Region (=X6). Among four major inputs, Farm Income (=X1) is most important. For important inputs such as Farm Income in this example, special attention should be paid not to miss correct data values in data collection or reporting stage. False declaration of income should be checked.

Table 1. Input Variable Importances in Grant Models

	Linear Model	Neural Network	Regression Tree
X1: Income	0.89	0.52	0.96
X2: Size	0.04	0.27	0.02
X3: Quality	0.05	0.24	0.01
X4: Crop	0.02	0.01	0.01
X5: Rain	0.06	0.29	0.00
X6: Region	0.02	0.06	0.01
D_y (in 10^3)	111	111	111
GoFit	0.85	0.80	0.91

4. Empirical Demonstration with A Simulated Data Set

To investigate how the proposed method works, we will apply it to a simulated data set from a known structure. Generate $X_1, \cdots, X_4$ and e independently from N(0,1) and get Y from

$$Y = 4X_1 + 3X_2 + 2X_3 + X_4 + e.$$

1,000 records are generated in such a way.

We consider three types of supervised learning models: Linear Regression Model, Neural Network and Regression Tree.

Linear Regression Model:

$$Y = b0 + b1 \cdot X1 + b2 \cdot X2 + b3 \cdot X3 + b4 \cdot X4 + e.$$

Neural Network:

Neural network with one hidden layer of two nodes was fitted using SPSS Clementine 7.1 (under default setting). Logistic function was used for activating hidden and output nodes.

Regression Tree:

C&R Tree with tree depth 5 was fitted using SPSS Clementine 7.1 (under default setting).

We obtained Table 2 for input variable importance measures by replicating the assessment procedures of Section 2 ten times and taking the average. The average absolute difference D_y of Y was computed from $N = 10n$ pairs.

Table 2. Input Variable Importances in Models for Simulated Data Set

	Linear Regression	Neural Network	Regression Tree
X_1	0.72	0.71	0.70
X_2	0.54	0.53	0.47
X_3	0.37	0.36	0.22
X_4	0.19	0.18	0.01
D_y	6.18	6.18	6.17
GoFit	0.82	0.81	0.60

As expected, the importance measures in Linear Regression are close to 4 : 3 : 2 : 1 in relative magnitude. Impressively, we note that same pattern appears in Neural Network. By contrast, in Regression Tree which takes the form of a step function, the importance measures appear proportional to 4 : 2.69 : 1.26 : 0.06, shrinking more rapidly than they should be. This may be the reason why Regression Tree has relatively low GoFit in this case.

5. Concluding Remarks

Our approach to importance measurement of input variables was motivated by a SPSS Clementine's implementation of Neural Net Modeling. In Version 7.1, input variable importances in neural network model are defined as follows (Watkins, 1997). In Steps 1, 2, and 3, we consider the case for X_p, to simplify mathematical expressions.

Step 1: For each i $(=1,\cdots,n)$, compute

$$\hat{y}_i^{\dagger} = f(x_{i1},\cdots,x_{i,p-1},x_{ip}^{\dagger})$$

for all possible values $x_{ip}^{\dagger}$ of X_p. For continuous input, $\hat{y}_i^{\dagger}$ is evaluated at five equally spaced values of the input variable (=0.0, 0.25, 0.50, 0.75, 1.0, if it is normed to be in the interval [0,1]).

Step 2: Find the maximum of $f(x_{i1},\cdots,x_{i,p-1},x_{ip}^{\dagger}) - f(x_{i1},\cdots,x_{i,p-1},x_{ip}^{\dagger\dagger})$ for each i, which is always non-negative, and denote the quantity by $d_{i,p}^{\max}$. Here, $x_{ip}^{\dagger}$ and $x_{ip}^{\dagger\dagger}$ are two possible values selected from $x_{1p},\cdots,x_{np}$.

Step 3: Take the average of $d_{1,p}^{\max},\cdots,d_{n,p}^{\max}$ and write

$$D_p^{\max} = \frac{1}{n}\left(d_{1,p}^{\max} + \cdots + d_{n,p}^{\max}\right).$$

Table 3. Input Variable Importances in Models for Simulated Data Set

	Neural Net [Table 2]	Clementine's Neural Net	Linear Reg [Table 2]	Beta Coef's in Linear Reg
X_1	0.71	0.59	0.72	0.72
X_2	0.53	0.47	0.54	0.54
X_3	0.36	0.36	0.37	0.37
X_4	0.18	0.17	0.19	0.18

Similarly, $D_j^{\max}$ can be defined for the input X_j ($j = 1, \cdots, p-1$).

Step 4: For comparison yard stick, consider the range of output values: Write

$$D_y^{\max} = \max\ y_i - \min\ y_i .$$

Step 5: Importance of the input X_j ($j = 1, \cdots, p-1$) is defined by

$$Imp^{\max}(X_j) = D_j^{\max} / D_y^{\max} .$$

See Table 3 for Clementine calculations of input variable importances in the neural network model (with one hidden layer of two nodes) for the simulated data set of Section 4. Note that the importance measures given by Clementine are not proportional to 4:3:2:1, betraying our expectation.

The difference between the proposed method and Clementine 7.1 implementation is that in the former one gathers the absolute difference of output values from two random perturbed inputs, while in the latter one takes the maximum difference of output values among all possible perturbations of the input variable in consideration. Hence SPSS Clementine's approach is a bit more mathematical rather than statistical, and, above all, it needs a lot more computations. Furthermore, in data mining applications, the extreme statistics can be volatile, so that Clementine implementation may not be resistant to outlying records.

In linear regression model, it is not difficult to show that our approach is equivalent to beta's, which is the regression coefficient for standardized inputs and output. This is numerically demonstrated at the last two columns of Table 3.

One of the contribution of this short paper is that input variable importances are defined for all types of supervised learning models including regression trees in a single context. As consequence, one can easily catch the differences and the commonalities between different types of models.

References

[1] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth, CA: Belmont.

[2] Chung, Y.S., Jung, J.Y. and Kim, C.S. (2002). "Bayesian analysis for neural network models," *Korean Communications in Statistics.* Vol, 9, 155-166.

[3] Han, S.T., Kang, H.C., Lee, S.K. and Lee, D.K. (2002) "A comparison on the efficiency of data mining softwares," *Korean Journal of Applied Statistics.* Vol. 15, 190-201 (Written in Korean).

[4] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer, New York.

[5] Hwang, C.H. and Kim, D.H. (1997). "Bootstrap model selection criterion for determining the number of hidden units in neural network model," *Korean Communications in Statistics.* Vol, 4, 827-832.

[6] Lee, T.R. and Moon, H.S. (1997). "Tree-structured classification for high risk dental caries," *Journal of Data Science and Classification* (Korean Classification Society). Vol. 1, 69-84.

[7] Lee, Y.M. and Song, M.S. (2002). "A study on unbiased methods in constructing classification trees," *Korean Communications in Statistics.* Vol. 9, 809-824.

[8] Kang, H.C., Han, S.T. and Choi, J.H. (2000). "Interpretation of data mining prediction model using decision tree," *Korean Communications in Statistics.* Vol, 7, 937-943.

[9] Kim, S.H. (1996). "Model selection for tree-structured regression," *Journal of Korean Statistical Society.* Vol. 25, 1-24.

[10] Lim, Y.B., Lee, S.Y. and Chung, J.H. (2001). "A combined multiple regression trees predictor for screening large chemical databases," *Korean Journal of Applied Statistics.* Vol. 14, 91-101 (Written in Korean).

[11] Ripley, R.D. (1996). *Pattern Recognition and Neural Network.* University Press, Cambridge.

[12] Sarle, W.S. (1998). How to measure importance of inputs? Unpublished White Paper, SAS Inc., NC: Cary.

[13] Song, M.S. and Yoon, Y.J. (2001). "A study on variable selection bias in data mining softwares," *Korean Journal of Applied Statistics.* Vol. 14,. 475-486 (Written in Korean).

[14] SPSS Inc. (2002). *Clementine 7.0 User's Guide,* SPSS Inc., Chicago. p.525.

[15] Watkins, D. (1997). Clementine's neural networks technical overview. Unpublished White Paper, SPSS Inc., Chicago.

[Received February 2003, Accepted April 2003]