Assignment and Risk Estimation

This document discusses how a class or a value is assigned to a node and to a case and three methods of risk estimation: the resubstitution method, test sample method and cross validation method. The information is applicable to the tree growing algorithms CART, CHAID, exhaustive CHAID and QUEST. Materials in this document are based on *Classification and Regression Trees* by Breiman, et al (1984). It is assumed that a CART, CHAID, exhaustive CHAID or QUEST tree has been grown successfully using a learning sample.

Notations

Y	The dependent variable, or target variable. It can be either categorical (nominal or ordinal) or continuous. If Y is categorical with *J* classes, its class takes values in C = {1, …, *J*}.
$\hbar = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$	The learning sample where $\mathbf{x}_n$ and y_n are the predictor vector and dependent variable for case *n*.
$\hbar(t)$	The learning samples that fall in node *t*.
f_n	The frequency weight associated with case *n*. Non-integral positive value is rounded to its nearest integer.
w_n	The case weight associated with case *n*.
$\pi(j)$, j = 1, …, J	Prior probability of $Y = j$
$C(i \mid j)$	The cost of miss-classifying a class *j* case as a class *i* case, $C(j \mid j) = 0$.

Assignment

Once the tree is grown, an assignment (also called action or decision) is given to each node based on the **learning** sample. To predict the dependent variable value for an incoming case, we first find in which terminal node it falls, then use the assignment of that terminal node for prediction.

Assignment of a Node

For any node *t*, let d_t be the assignment given to node *t*,

$$d_t = \begin{cases} j^*(t) & Y \text{ is categorical} \\ \bar{y}(t) & Y \text{ is continuous} \end{cases},$$

$$j^*(t) = \arg\min_i \sum\nolimits_j C(i \mid j)\, p(j \mid t),$$

$$\bar{y}(t) = \frac{1}{N_w(t)} \sum_{n \in \hbar(t)} w_n f_n y_n ,$$

where

$$p(j \mid t) = \frac{p(j,t)}{\sum\limits_j p(j,t)}, \quad p(j,t) = \pi(j) \frac{N_{w,j}(t)}{N_{w,j}},$$

$$N_w = \sum_{n \in \hbar} w_n f_n \ , \quad N_{w,j} = \sum_{n \in \hbar} w_n f_n I(y_n = j),$$

$$N_w(t) = \sum_{n \in \hbar(t)} w_n f_n \ , \quad N_{w,j}(t) = \sum_{n \in \hbar(t)} w_n f_n I(y_n = j).$$

If there is more than one class j that achieves the minimum, choose $j^*(t)$ to be the smallest such j for which $N_{f,j}(t) = \sum_{n \in \hbar(t)} f_n I(y_n = j)$ is greater than 0, or the absolute smallest if $N_{f,j}(t)$ is zero for all of them.

For CHAID and exhaustive CHAID, use $\pi(j) = N_{w,j} \big/ N_w$ in the equation.

Assignment of a case

For a case with predictor vector **x**, the assignment or prediction $\mathrm{d}_{\mathrm{T}}(\mathbf{x})$ for this case by the tree T is

$$d_T(\mathbf{x}) = \begin{cases} j^*\big(t(\mathbf{x})\big) & Y \text{ is categorical} \\ \bar{y}(t(\mathbf{x})) & Y \text{ is continuous} \end{cases},$$

where $t(\mathbf{x})$ is the terminal node the case falls in.

Risk estimation

Note that case weight is not involved in risk estimation, though it is involved in tree growing process and class assignment.

Loss Function

A loss function *L*(*y*, *a*) is a real-valued function in which *y* is the actual value of *Y* and *a* is the assignment taken. Throughout this document, the following types of loss functions are used.

$$L(y,a)=\begin{cases} C(a\mid y) & Y \text{ is categorical} \\ (y-a)^2 & Y \text{ is continuous} \end{cases}.$$

Risk Estimation of a tree T

Suppose that a tree *T* is grown and assignments have been given to each node. Let $\tilde{T}$ denote the set of terminal nodes of the tree. Let *D* be the data set used to calculate the risk. Dropping all cases in *D* to *T*, let *D(t*) denote the set of cases that fall in node *t*. The risk of the tree based on data *D* is estimated by

$$R(T\mid D)=\begin{cases} \sum_j \pi(j)\bar{L}_j & Y \text{ categorical} \\ \bar{L} & Y \text{ continuous} \end{cases} = \begin{cases} \bar{L} & Y \text{ categorical, M1} \\ \sum_j \pi(j)\bar{L}_j & Y \text{ categorical, M2} \\ \bar{L} & Y \text{ continuous} \end{cases},$$

where M1 represents empirical prior situation, and M2 non-empirical prior, and

$$\bar{L}=\frac{1}{N_f}\sum_{n\in D} f_n L(y_n, d_T(\boldsymbol{x}_n)),\quad \bar{L}_j=\frac{1}{N_{f,j}}\sum_{n\in D} f_n L(y_n, d_T(\boldsymbol{x}_n)) I(y_n=j),$$

$$N_f=\sum_{n\in D} f_n\ ,\ \ N_{f,j}=\sum_{n\in D} f_n\, I(y_n=j).$$

Assuming that $L(y_n, d_T(\boldsymbol{x}_n))$ are independent of each other, then the variance of *R(T*) is estimated by

$$\mathrm{Var}(R(T))=\begin{cases} \sum_j \pi(j)^2 \frac{s_j^2}{N_{f,j}} & Y \text{ categorical, M2} \\ \frac{s^2}{N_f} & Y \text{ con, or, } Y \text{ cat and M1} \end{cases},$$

where

$$
\begin{aligned}
s_j^2 &= \frac{1}{N_{f,j}} \sum_{n \in D} f_n \left(L(y_n, d_T(\boldsymbol{x}_n)) - \bar{L}_j \right)^2 I(y_n = j) \\
&= \frac{1}{N_{f,j}} \sum_{n \in D} f_n L^2(y_n, d_T(\boldsymbol{x}_n)) I(y_n = j) - \bar{L}_j^{\,2}
\end{aligned},
$$

$$
s^2 = \frac{1}{N_f} \sum_{n \in D} f_n \left(L(y_n, d_T(\boldsymbol{x}_n)) - \bar{L} \right)^2 = \frac{1}{N_f} \sum_{n \in D} f_n L^2(y_n, d_T(\boldsymbol{x}_n)) - \bar{L}^2 .
$$

Putting everything together we get

$$
R(T \mid D) = \begin{cases}
\dfrac{1}{N_f} \displaystyle\sum_{t \in \tilde{T}} \sum_j C(j^*(t) \mid j) N_{f,j}(t) & Y \text{ categorical, M1} \\
\displaystyle\sum_j \dfrac{\pi(j)}{N_{f,j}} \sum_{t \in \tilde{T}} C(j^*(t) \mid j) N_{f,j}(t) & Y \text{ categorical, M2} \\
\dfrac{1}{N_f} \displaystyle\sum_{t \in \tilde{T}} \sum_{n \in D(t)} f_n (y_n - \bar{y}(t))^2 & Y \text{ continuous}
\end{cases},
$$

$$
\begin{aligned}
&\operatorname{Var}(R(T \mid D)) \\
&= \begin{cases}
\dfrac{1}{(N_f)^2} \left\{ \displaystyle\sum_j \sum_{t \in \tilde{T}} N_{f,j}(t) C(j^*(t) \mid j)^2 - N_f R(T \mid D)^2 \right\} & Y \text{ cat, M1} \\
\displaystyle\sum_j \left(\dfrac{\pi(j)}{N_{f,j}} \right)^2 \left[\sum_{t \in \tilde{T}} N_{f,j}(t) C(j^*(t) \mid j)^2 - \dfrac{\left\{ \sum_{t \in \tilde{T}} N_{f,j}(t) C(j^*(t) \mid j) \right\}^2}{N_{f,j}} \right] & Y \text{ cat, M2} \\
\dfrac{1}{N_f^{\,2}} \left\{ \displaystyle\sum_{t \in \tilde{T}} \sum_{n \in D(t)} f_n (y_n - \bar{y}(t))^4 - N_f R(T \mid D)^2 \right\} & Y \text{ con}
\end{cases},
\end{aligned}
$$

where

$$
N_{f,j}(t) = \sum_{n \in D(t)} f_n I(y_n = j) .
$$

The estimated standard error of *R(T|D)* is given by $\operatorname{se}(R(T \mid D)) = \sqrt{\operatorname{var}(R(T \mid D))}$.

Risk estimation of a tree is often written as $R(T \mid D) = \sum_{t \in \tilde{T}} R(t \mid D)$ with $R(t \mid D)$ being the contribution from node *t* to the tree risk such that

$$R(t \mid D) = \begin{cases} \dfrac{1}{N_f} \sum_j N_{f,j}(t) C(j^*(t) \mid j) & Y \text{ categorical, M1} \\ \sum_j \dfrac{\pi(j) N_{f,j}(t)}{N_{f,j}} C(j^*(t) \mid j) & Y \text{ categorical, M2} \\ \dfrac{1}{N_f} \sum_{n \in D(t)} f_n (y_n - \bar{y}(t))^2 & Y \text{ continuous} \end{cases}.$$

Resubstitution Estimate of the Risk of a tree T

The resubstitution risk estimation method uses the same set of data (learning sample $\hbar$) that is used to grow the tree *T* to calculate its risk, i.e.

$$\begin{aligned} &R(t) = R(t \mid \hbar) \\ &R(T) = R(T \mid \hbar) = \sum_{t \in \tilde{T}} R(t) \quad . \\ &\mathrm{Var}(R(T)) = \mathrm{Var}(R(T \mid \hbar)) \end{aligned}$$

Test Sample Estimate of the Risk

The idea of test sample risk estimation is that the whole data set is divided into 2 mutually exclusive subsets $\hbar$ and $\hbar'$. $\hbar$ is used as a learning sample to grow a tree *T* and $\hbar'$ is used as a test sample to check the accuracy of the tree. The test sample estimate is

$$\begin{aligned} &R^{ts}(T) = R(T \mid \hbar') \\ &\mathrm{Var}(R^{ts}(T)) = \mathrm{Var}(R(T \mid \hbar')) \end{aligned}.$$

Cross Validation Estimate of the Risk of a tree T

Cross validation estimation is provided only when a tree is grown using the *automatic tree growing process*. Let *T* be a tree which has been grown using *all* data from the *whole data set* $\hbar^0$. Let $V \geq 2$ be a positive integer.

1. Divide $\hbar^0$ into *V mutually exclusive* subsets $\hbar'_v$, *v* = 1, …, *V*. Let $\hbar_v$ be $\hbar^0$ - $\hbar'_v$, *v* = 1, …, *V*.
2. For each *v*, consider $\hbar_v$ as a learning sample and grow a tree T_v on $\hbar_v$ by using the *same* set of user specified stopping rules which was applied to grow *T*.
3. After T_v is grown and assignment $j^*_v(t)$ or $\bar{y}_v(t)$ for node *t* of T_v is done, consider $\hbar'_v$ as a test sample and calculate its test sample risk estimate $R^{ts}(T_v)$.

4. Repeat above for each v = 1, …, V. The weighted average of these test sample risk estimates is used as the V-fold cross validation risk estimate of T.

The V-fold cross validation estimate, $R^{cv}(T)$, of the risk of a tree T and its variance are estimated by

$$R^{CV}(T) = \begin{cases} \sum_j \pi(j) \frac{1}{N^0_{f,j}} \sum_v N'_{v,f,j} R^{ts}(T_v \mid j) & \text{Y categorical, M2} \\ \frac{1}{N^0_f} \sum_v N'_{v,f} R^{ts}(T_v) & Y \text{ con, or, Y cat and M1} \end{cases},$$

$$\begin{aligned} &\text{Var}(R^{CV}(T)) \\ &= \begin{cases} \frac{1}{(N^0_f)^2} \left\{ \sum_v \sum_j \sum_{t \in \tilde{T}_v} N'_{v,f,j}(t) C(j^*_v(t) \mid j)^2 - N^0_f R^{cv}(T)^2 \right\} & Y \text{ cat, M1} \\ \sum_j \left(\frac{\pi(j)}{N^0_{f,j}} \right)^2 \left[\sum_v \sum_{t \in \tilde{T}_v} N'_{v,f,j}(t) C(j^*_v(t) \mid j)^2 - \frac{\left\{ \sum_v N'_{v,f,j} R^{ts}(T_v \mid Y = j) \right\}^2}{N^0_{f,j}} \right] & Y \text{ cat, M2} \\ \frac{1}{(N^0_f)^2} \left\{ \sum_v \sum_{t \in \tilde{T}_v} \sum_{n \in \hbar'_v(t)} f_n (y_n - \bar{y}_v(t))^4 - N^0_f R^{cv}(T)^2 \right\} & Y \text{ con} \end{cases}, \end{aligned}$$

where

$$R^{ts}(T_v \mid j) = \frac{1}{N'_{v,f,j}} \sum_{t \in \tilde{T}_v} N'_{v,f,j}(t) C(j^*_v(t) \mid j),$$

$$N^0_f = \sum_{n \in \hbar^0} f_n \ , \ \ N^0_{f,j} = \sum_{n \in \hbar^0} f_n I(y_n = j),$$

$$N'_{v,f} = \sum_{n \in \hbar'_v} f_n \ , \ \ N'_{v,f,j} = \sum_{n \in \hbar'_v} f_n I(y_n = j), \ \ N'_{v,f,j}(t) = \sum_{n \in \hbar'_v(t)} f_n I(y_n = j).$$

References

Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J., 1984. *Classification and Regression Trees* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.