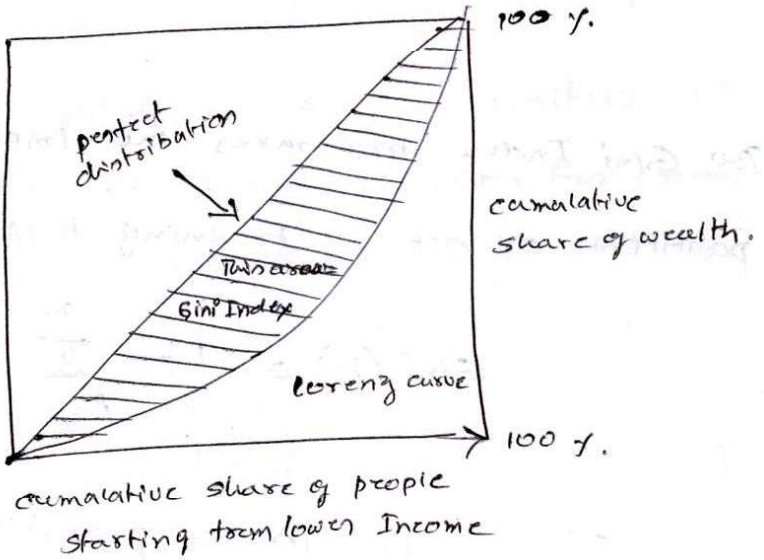
(3) CLASSIFICATION AND REGRESSION TREE (CART):

The GINI Index is used in CART and IBM Intelligent Miner.

An Italian economist, Corrado Gini (1884 - 1965) proposed the Gini Index as a measure of resource inequality in a population. The Index varies from 0 to 1, zero means no inequality and 1 means maximum possible inequality.

The Index is based on the Lorenz curve, which plots cumulative family against the number of families from poorest to richest.

The Lorenz curve is the basis of the Gini Index. The Index is the ratio of the area between the Lorenz curve and the 45° line to the area under the 45° line. The smaller the ~~area~~ ratio, the less is the area between the two curves and the more evenly distributed are the wealth.



For example, the Queensland Government has analyzed crime data using Gini Index to find how different types of crimes are distributed in the state. The results obtained are as follows:

Crime type	Gini Index
prostitution offences	0.65
Liquor "	0.52
Armed Robbery	0.38
Kidnapping	0.33
Motor vehicle theft	0.30
Fraud	0.26
Drug offences	0.22
Stealing from homes	0.17

It is clear that crimes like stealing from homes are fairly evenly distributed in the state since Gini Index for such crimes is small, while crimes like prostitution offences are, perhaps not surprisingly, not evenly distributed through the state since such crimes are more frequent in larger cities than in smaller cities and towns.

The Gini Index measures the impurities of D, a data partition or set of training tuples, as:

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2$$

where, P_i is the probability that a tuple in D belongs to class C_i and is estimated by $|C_{i,D}| / |D|$

The sum is computed over m classes.

Gini Index considers a binary split for each attribute.

Let A be discrete-valued attribute having v distinct values, $\{a_1, a_2, \ldots a_v\}$, occurring in D. To determine best binary split on A, we examine all the possible subsets that can be formed using known values of A.

If A has v possible values, then there are 2^v possible subsets. For example, if income has three possible values, namely {low, medium, high} then the possible subsets are {low, medium, high}, {low, medium}, {low, high}, {medium, high} {low}, {medium}, {high} and {}. We exclude the power set, {low, medium, high} and the empty set from consideration since, conceptually they do not represent a split. Therefore, there are $2^v - 2$ possible ways to form two partitions of the data D, based on a binary split on A.

We compute a weighted sum of the impurities of each resulting partition on a binary split. For example, if a binary split on A partitions D into D_1 and D_2, the gini Index of D given that partitioning is:

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2).$$

For each attribute, each of the possible binary splits is considered.

For continuous-valued attributes, D_1 is the set of tuples in D satisfying $A \leq$ split-point, and D_2 is the set of tuples in D satisfying $A >$ split-point.

The reduction in impurity that would be incurred by a binary split on a discrete or continuous valued attribute A is:

$$\Delta \text{Gini}(A) = \text{Gini}(D) - \text{Gini}_A(D).$$

The attribute that maximizes the reduction in impurity (or equivalently, has the minimum Gini index) is selected as the splitting attribute.

Example:

Nine tuples belong to class buys-computer = YES and five tuples belong to the class buys-computer = NO.

A root node N is created for the tuples in D. To compute the impurities of D:

$$\text{Gini}(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459.$$

To find the splitting criterion for the tuples in D, we need to compute the gini index for each attribute.

The possible splitting subsets for the attribute "Income"

(i) {low, medium}, {high}
(ii) {low, high}, {medium}
(iii) {medium, high}, {low}

TREE PRUNING :

Gender	car ownership	Travel cost ($/1/km)	Income level	Transportation mode
			Attributes ←	Classes
Male	0	cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	standard	Medium	Train

Compute Gini Index:

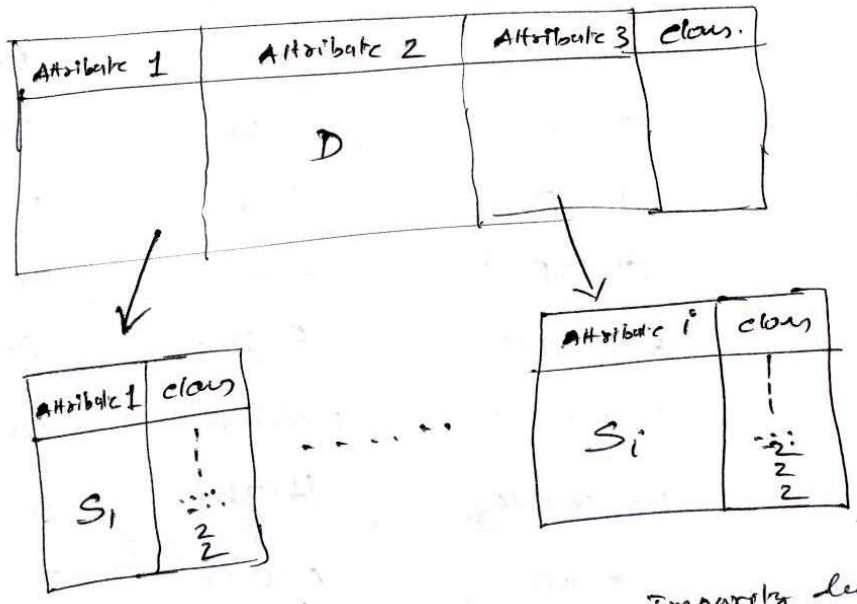
~~Gini (D) = 1 − $\left(\frac{4}{10}\right)^2$~~

classes of Transportation consist of three groups of Bus, car and Train. In this case, we have 4 buses, 3 cars and 3 trains. The total data is 10 rows.

∴ ~~Gini Index~~ Gini (D) = $1-\left(\left(\frac{4}{10}\right)^2+\left(\frac{3}{10}\right)^2+\left(\frac{3}{10}\right)^2\right)$

= 0.660

From Table D and for each associated subset S_i, we compute degree of impurity.



To compute the degree of Impurity, we must distinguish whether it is come from the parent table D or it came from a subset table S_i with attribute i.

If the table in a parent table D, we simply compute the number of records of each class. We can compute the degree of impurity based on the Transportation mode class. In this case we have 4 Buses, 3 cars and 3 trains.

$$\therefore \text{Gini}(D) = 1 - \left(\left(\frac{4}{10}\right)^2 - \left(\frac{3}{10}\right)^2 - \left(\frac{3}{10}\right)^2 \right)$$

$$= 0.660$$

FIRST ITERATION:

① Attribute "Travel cost" :

The attribute travel cost per km has three values: cheap, standard and expensive. Now we sort the table S_p = [Travel cost/Km, Transportation mode] based on the values of Travel cost per km. Then we separate each value of the travel cost and compute the degree of impurity using Gini Index:

Travel cost ($)/Km	Transportation mode
cheap	Bus
cheap	Bus
cheap	Bus
cheap	Bus
cheap	Train
Expensive	car
Expensive	car
Expensive	car
Standard	Train
Standard	Train

Travel cost ($)/Km	classes
cheap	Bus
cheap	Bus
cheap	Bus
cheap	Bus
cheap	Train

4 Buses, 1 Train.

Travel cost ($)/Km	classes
Expensive	car
Expensive	car
Expensive	car

3 cars

Travel cost ($)/Km	classes.
Standard	Train
Standard	Train

2 Trains.

$$\text{Gini Index} = \frac{5}{10}\left(1-\left(\frac{4}{5}\right)^2-\left(\frac{1}{5}\right)^2\right) + \frac{3}{10}\left(1-\left(\frac{3}{3}\right)^2\right) + \frac{2}{10}\left(1-\left(\frac{2}{2}\right)^2\right)$$

$$= 0.16 + 0.0 + 0.0 = 0.16$$

Gain of Travel cost/km based on Gini Index =

$$0.660 - 0.16 = 0.500$$

We try to compute Gain for other three attributes of Gender, car ownership and Income level.

② Attribute 'Gender':

Gender	Transportation mode
Male	Bus
Male	Bus
Female	Train
Female	Bus
Male	Bus
Male	Train
Female	Train
Female	car
Male	car
Female	car

Gender	Transportation mode
~~Male~~ Female	~~Bus~~
~~Male~~ Female	car
~~Male~~ Female	car
~~Male~~ Female	Train
~~Male~~ Female	Train

1 Bus, 2 car, 2 Train = 5

Gender	Transportation mode
~~Male~~	Bus
Male	Bus
~~Male~~	Bus
Male	car
Male	Train

3 Bus, 1 car, 1 Train = 5

Gini Index = $\frac{5}{10}\left(1-\left(\frac{1}{5}\right)^2-\left(\frac{2}{5}\right)^2-\left(\frac{2}{5}\right)^2\right)+\frac{5}{10}\left(1-\left(\frac{3}{5}\right)^2-\left(\frac{1}{5}\right)^2-\left(\frac{1}{5}\right)^2\right)$

$= 0.32 + 0.28 = 0.6$

Gain of Gender based on Gini Index =

$= 0.660 - 0.6 = 0.060$

(3) Attribute 'car ownership':

car ownership	Transportation
0	Bus
1	Bus
1	Train
0	Bus
1	Bus
0	Train
1	Train
1	car
2	car
2	car

car ownership	classes
0	Bus
0	Bus
0	Train

2 Bus, 1 Train = 3 records

carownership	classes
1	Bus
1	Bus
1	car
1	Train
1	Train

2 Bus, 1 car, 2 Train = 5 records

car ownership	classes
2	car
2	car

2 cars =

$$\text{Gini Index} = \frac{3}{10}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right) + \frac{5}{10}\left(1-\left(\frac{2}{5}\right)^2-\left(\frac{1}{5}\right)^2-\left(\frac{2}{5}\right)^2\right) + \frac{2}{10}\left(1-\left(\frac{2}{2}\right)^2\right)$$

$$= 0.133 + 0.32 + 0.0$$

$$= 0.453$$

Gain of car ownership on Gini Index =

$$= 0.660 - 0.453 = 0.207$$

④ Attribute 'Income level':

Income level	Transportation mode
Low	Bus
medium	Bus
Medium	Train
Low	Bus
Medium	Bus
Medium	Train
Medium	Train
High	car
Medium	car
High	car

Income level	classes
High	car
High	car

2 cars

Income level	classes
Low	Bus
Low	Bus

2 Bus

Income level	classes
Medium	Bus
Medium	Bus
Medium	car
Medium	Train
Medium	Train
Medium	Train

2 Bus, 1 car, 3 Train = 6

$$\text{Gini Index} = \frac{2}{10}\left(1-\left(\frac{2}{2}\right)^2\right) + \frac{2}{10}\left(1-\left(\frac{2}{2}\right)^2\right) + \frac{6}{10}\left(1-\left(\frac{2}{6}\right)^2-\left(\frac{1}{6}\right)^2-\left(\frac{3}{6}\right)^2\right)$$

$$= 0.0 + 0.0 + 0.366$$

$$= 0.366$$

Gain of Income level based on Gini Index =

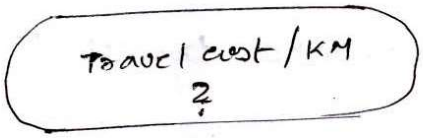
$$0.660 - 0.366 = 0.294$$

Results of First Iteration:

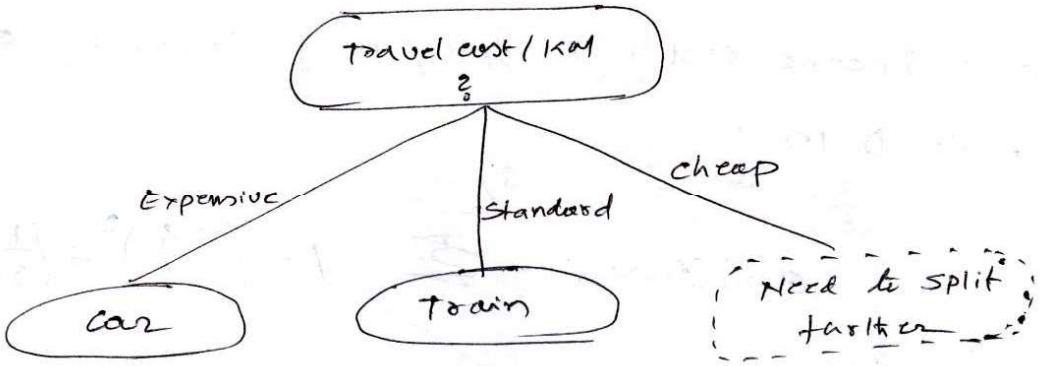
Gain	Gender	car ownership	Travel cost/KM	Income
Gini Index	0.060	0.207	0.500	0.29

Once we get the Gain for all attributes, then we find optimum attribute that produce the maximum Gain. It

This case travel cost produces maximum gain. We put this optimum attribute into the node of our decision tree. As it is the first node, it is the root node of the decision tree.



Expensive travel cost/KM is associated only with pure class of Car while standard travel cost/KM is only related to pure class of Train. pure class is always assigned into leaf node of a decision tree. The decision tree after first iteration is as follows.



Once the optimum attribute is obtained, we can split the data table according to that optimum attribute. We split the data table based on the value of travel cost per KM.

SECOND ITERATION:

In the second iteration, since Expensive and Standard travel cost/KM have been associated with pure class, we do not need more data any longer. Our data for the second iteration D after removing attribute travel cost/KM is:

DATA:

Gender	car ownership	Income level	Transportation mode
Female	0	low	Bus
Male	0	low	Bus
Male	1	medium	Bus
Male	1	medium	Bus
Female	1	medium	Train

4 Buses, 1 Train = 5 records

Now we have three attributes: Gender, car ownership and Income level. The degree of Impurity of the data table D is:

$$\text{Gini Index} = 1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right)$$

$$= 0.32$$

We repeat the procedure of computing degree of impurity and Gain for the three attributes.

Subsets of Second Iteration:

Gender	classes
Female	Bus
Female	Train

1 Bus, 1 Train

carownership	Classes
0	Bus
0	Bus

2 Bus

Income level	classes
Low	Bus
Low	Bus

2 Bus

Gender	classes.
Male	Bus
Male	Bus
Male	Bus

3 Bus

car ownership	classes
1	Bus
1	Bus
1	Train

2 Bus, 1 Train

Income level	classes.
Medium	Bus
Medium	Bus
Medium	Train

2 Bus, 1 Train

Gini Index =

Gain =

Gini Index = $\frac{2}{5}\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right)+\frac{3}{5}\left(1-\left(\frac{3}{3}\right)^2\right)$

$\frac{2}{5}\left(1-\left(\frac{2}{2}\right)^2\right)+\frac{3}{5}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right)=$

$\frac{2}{5}\left(1-\left(\frac{2}{2}\right)^2+\frac{3}{5}\left(1-\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right)=.$

1

0.2

① Gini Index (D) (Gender) $=\frac{2}{5}\left(1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right)+\frac{3}{5}\left(1-\frac{3}{3}\right)^2=0.2$

Gain of Gender based on Gini Index = $0.32-0.2=0.120$

② Gini Index (D) carownership $=\frac{2}{5}\left(1-\left(\frac{2}{2}\right)^2\right)+\frac{3}{5}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right)=0.266$

Gain of carownership based on Gini Index: $0.32-0.266=0.054$

$$\text{Gini Index}_{\text{Income level}}(D) = \frac{2}{5}\left(1-\left(\frac{2}{2}\right)^2\right) + \frac{3}{5}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right) = 0.266$$

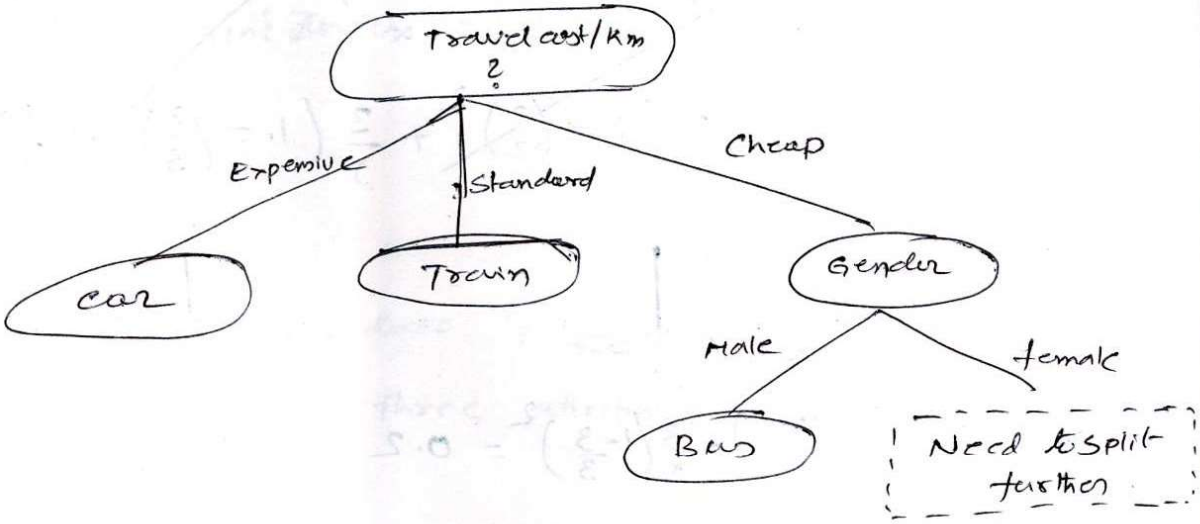
Gain of Income level based on Gini Index = 0.32 − 0.266 = 0.054

Results after 2nd Iteration:

Gain	Gender	car ownership	Income level
Gini Indx	0.120	0.054	0.054

The maximum gain is obtained for the optimum attribute Gender. The data table is split according to that optimum attribute. Male Gender is only associated with pure class Bus, while female still need further split of attribute.

The decision tree is now updated as follows:



add Gender which has two values of male and female the pure class is related to leaf node, thus male gender has leaf node of Bus. For female Gender, we ne to split further the attributes in the next Iteration.

THIRD ITERATION :

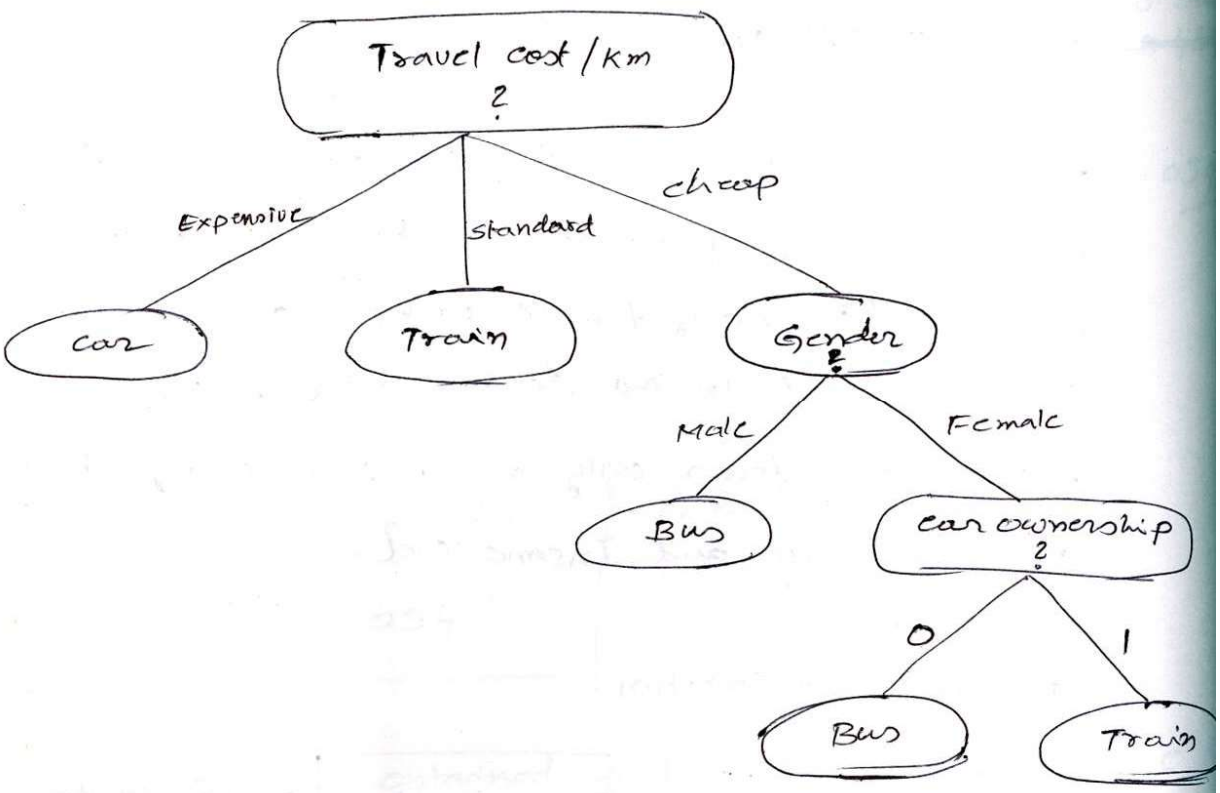
Data table of the third iteration comes only from part of the data table of the second iteration with male gender removed (thus only female part). Since attribute Gender has been used in the decision table, we can remove the attribute and focus only on the remaining two attributes: car ownership and Income level.

Data for third Iteration.

car ownership	Incomelevel	Transportation mode
0	Low	Bus
1	medium	Train

If we observe the data table of the third iteration, it consists only two rows. Each row has distinct values. If we use attribute car ownership, we will get pure class for each ~~of its~~ of its value. Similarly, attribute income level will also give pure class for each value. Therefore, we can use either one of the two attributes.

Suppose, we select attribute car ownership, we can update our decision tree into the final version.



Now we have grown the final full decision tree based on the data D.

NOISY DATA:

Frequently, training data contains "noise" i.e example wh are misclassified, or where one or more of the attribute values is wrong.

In such cases, we end up with or part of our decision tree w considers may 100 examples, of which 99 are in class C_1 other is apparently in class C_2 (because it is misclassified)

If there are any unused attributes, we might be abl to use them to elaborate the tree to care of them, b the subtree we would be building would in fact be wrong would likely misclassify real data.

Thus, if there is noise in the training data, it may be wis "prune" the decision tree to remove nodes which, statistically seem likely to arise from noise in the training data