Week 01: Boosting

CM50265 Machine Learning 2

Topic 4:
An Example: Predict your ML2 grade

Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark
1	16	10	Yes	85	90
2	4	4	Yes	60	55
3	15	6	Yes	85	84
4	7	8	No	45	64
5	17	10	No	90	78
6	8	7	Yes	68	67

Q: Suppose we have an initial model which outputs a constant value of ML2 marks, what value will it be?

$$\frac{90+55+84+64+78+67}{6} = 73$$

A: The average of ML2 marks (Observed values)

Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark	Residual 1
1	16	10	Yes	85	90	17
2	4	4	Yes	60	55	-18
3	15	6	Yes	85	84	11
4	7	8	No	45	64	-9
5	17	10	No	90	78	5
6	8	7	Yes	68	67	-6

Now calculate the residual for each sample.

Residual = Observed - Predicted

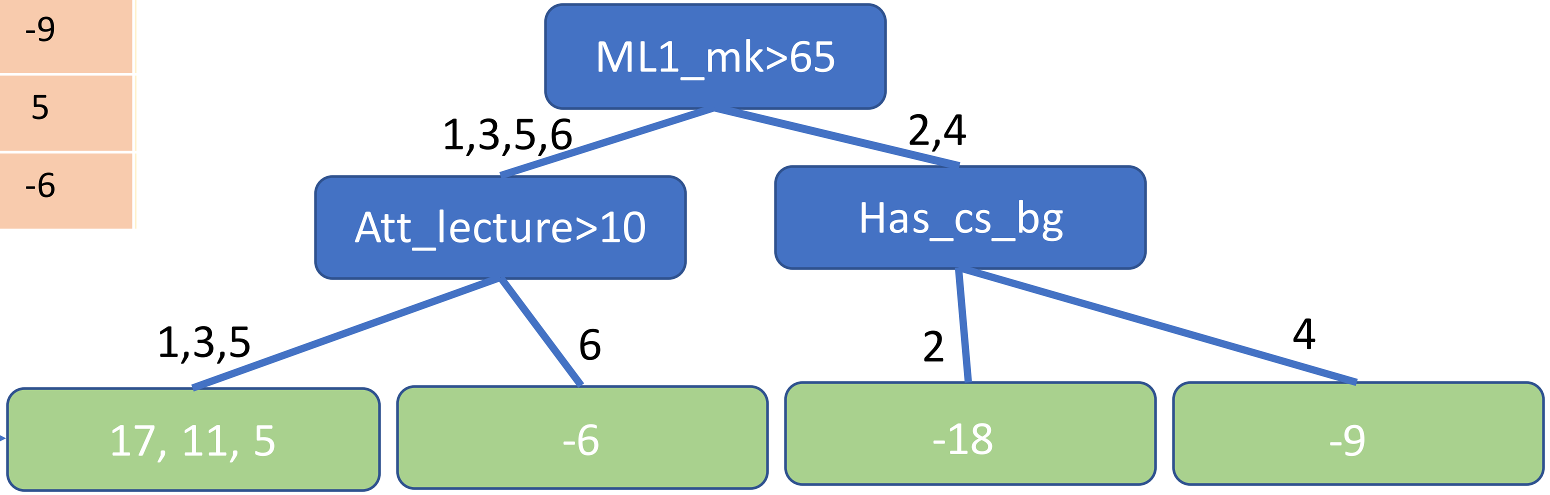
73

Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark	Residual 1
1	16	10	Yes	85	90	17
2	4	4	Yes	60	55	-18
3	15	6	Yes	85	84	11
4	7	8	No	45	64	-9
5	17	10	No	90	78	5
6	8	7	Yes	68	67	-6

Then build a decision tree to fit the residuals.

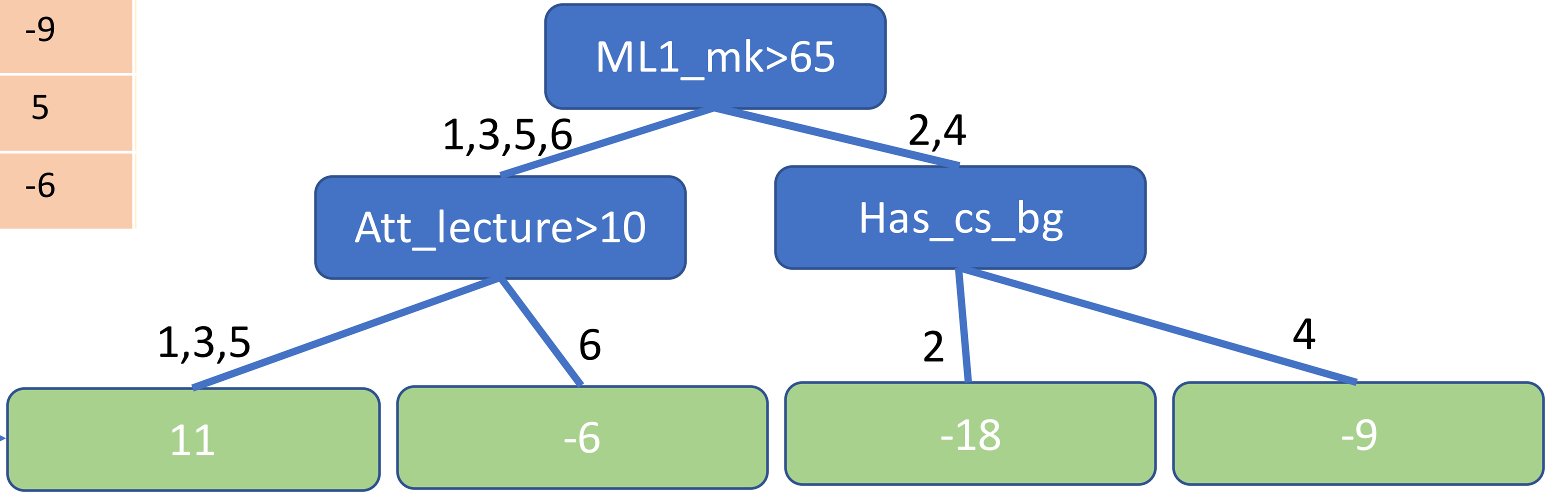
Take average of all the residuals in each leaf



Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark	Residual 1
1	16	10	Yes	85	90	17
2	4	4	Yes	60	55	-18
3	15	6	Yes	85	84	11
4	7	8	No	45	64	-9
5	17	10	No	90	78	5
6	8	7	Yes	68	67	-6

Then build a decision tree to fit the residuals.



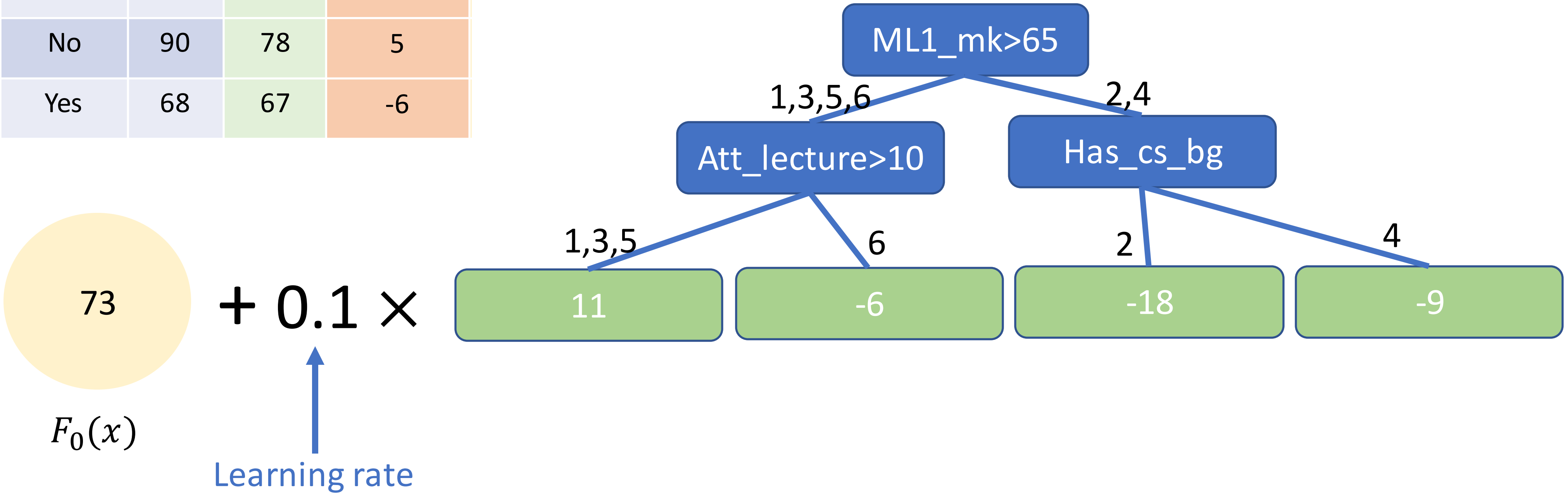
Take average of all the residuals in each leaf

$$\frac{17+11+5}{3}$$

Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark	Residual 1
1	16	10	Yes	85	90	17
2	4	4	Yes	60	55	-18
3	15	6	Yes	85	84	11
4	7	8	No	45	64	-9
5	17	10	No	90	78	5
6	8	7	Yes	68	67	-6

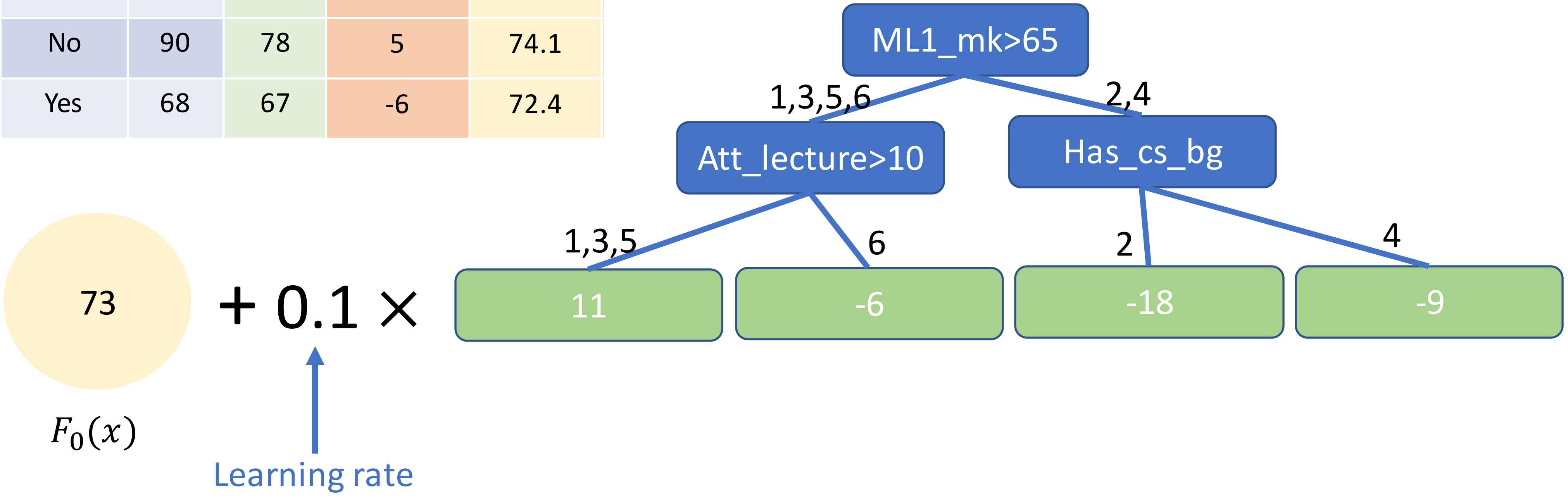
The new prediction: $F_1(x)$



Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark	Residual 1	Predict 1
1	16	10	Yes	85	90	17	74.1
2	4	4	Yes	60	55	-18	71.2
3	15	6	Yes	85	84	11	74.1
4	7	8	No	45	64	-9	72.1
5	17	10	No	90	78	5	74.1
6	8	7	Yes	68	67	-6	72.4

The new prediction: $F_1(x)$

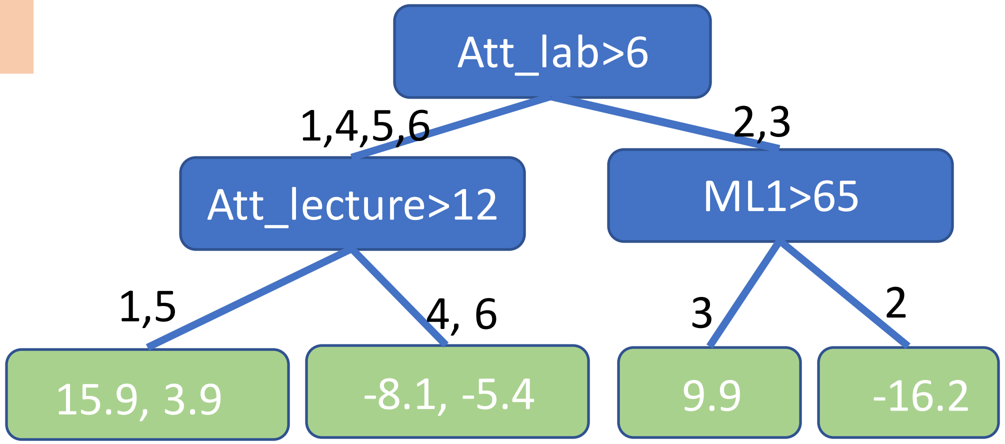


Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark	Residual 1	Predict 1	Residual 2
1	16	10	Yes	85	90	17	74.1	15.9
2	4	4	Yes	60	55	-18	71.2	-16.2
3	15	6	Yes	85	84	11	74.1	9.9
4	7	8	No	45	64	-9	72.1	-8.1
5	17	10	No	90	78	5	74.1	3.9
6	8	7	Yes	68	67	-6	72.4	-5.4

Now calculate the new residual for each sample.
Then build another decision tree to fit its residual:

Take average of all the residuals in each leaf

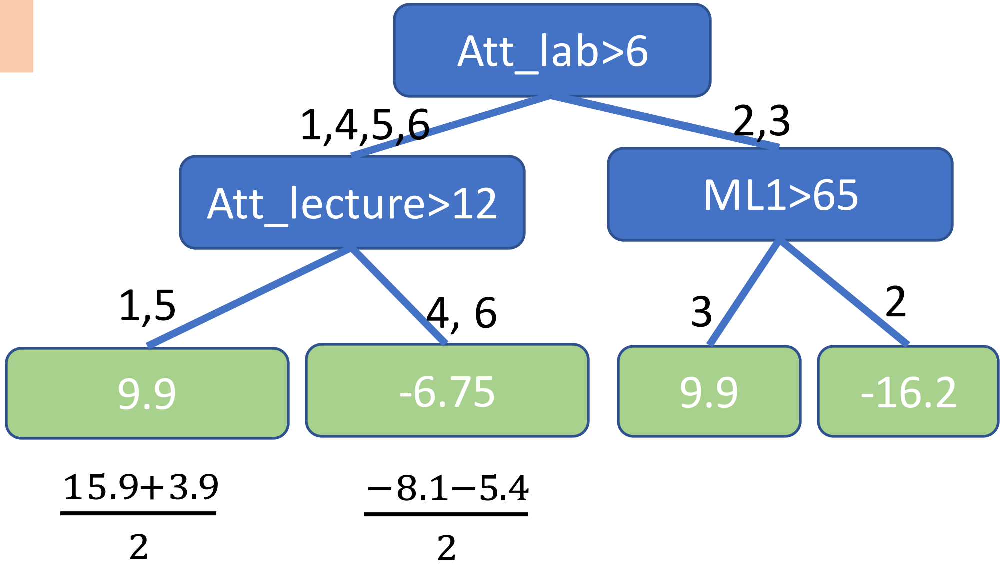


Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark	Residual 1	Predict 1	Residual 2
1	16	10	Yes	85	90	17	74.1	15.9
2	4	4	Yes	60	55	-18	71.2	-16.2
3	15	6	Yes	85	84	11	74.1	9.9
4	7	8	No	45	64	-9	72.1	-8.1
5	17	10	No	90	78	5	74.1	3.9
6	8	7	Yes	68	67	-6	72.4	-5.4

Now calculate the new residual for each sample. Then build another decision tree to fit its residual:

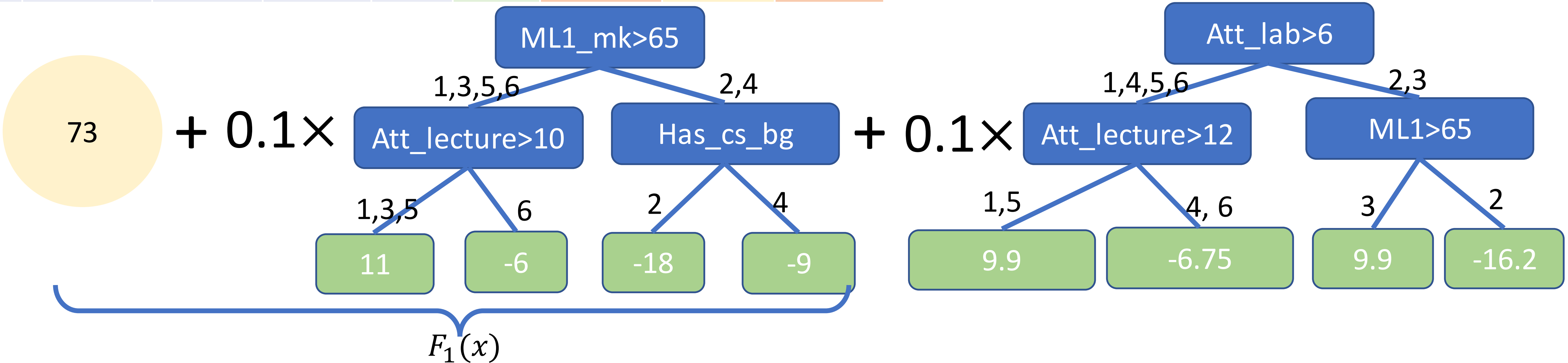
Take average of all the residuals in each leaf



Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark	Residual 1	Predict 1	Residual 2
1	16	10	Yes	85	90	17	74.1	15.9
2	4	4	Yes	60	55	-18	71.2	-16.2
3	15	6	Yes	85	84	11	74.1	9.9
4	7	8	No	45	64	-9	72.1	-8.1
5	17	10	No	90	78	5	74.1	3.9
6	8	7	Yes	68	67	-6	72.4	-5.4

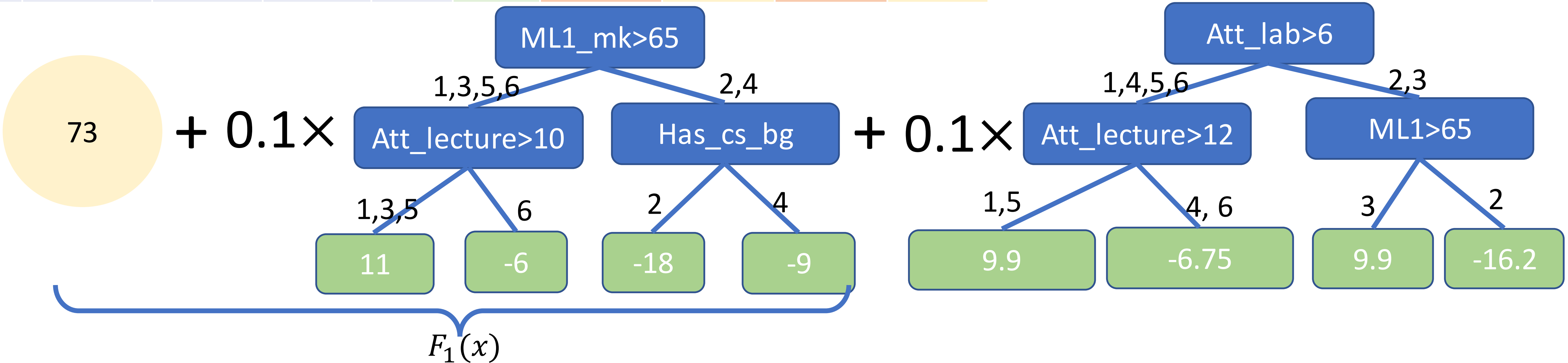
New prediction: $F_2(x)$



Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	CS Bg	ML1 Mark	ML2 Mark	Residual 1	Predict 1	Residual 2	Predict 2
1	16	10	Yes	85	90	17	74.1	15.9	75.09
2	4	4	Yes	60	55	-18	71.2	-16.2	69.58
3	15	6	Yes	85	84	11	74.1	9.9	73.11
4	7	8	No	45	64	-9	72.1	-8.1	71.425
5	17	10	No	90	78	5	74.1	3.9	75.09
6	8	7	Yes	68	67	-6	72.4	-5.4	71.725

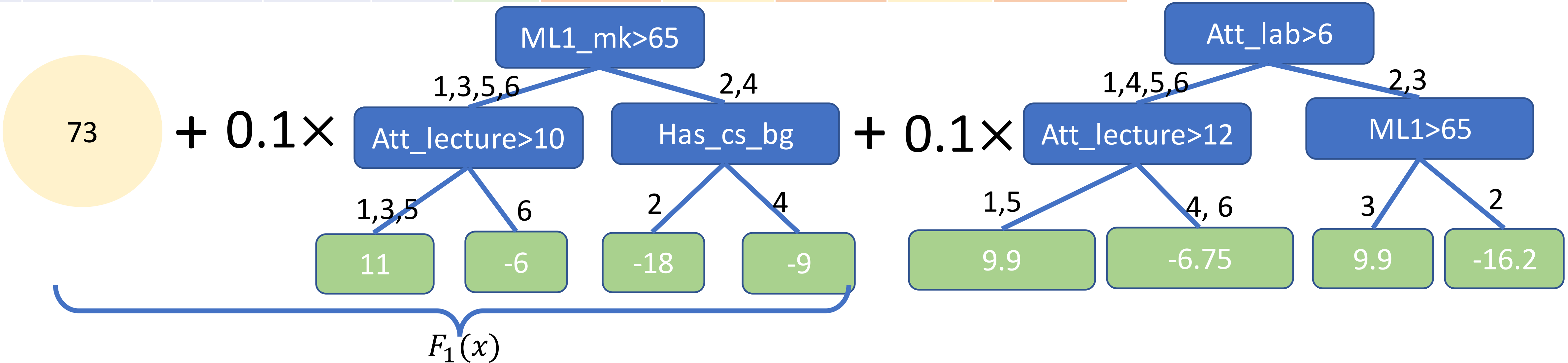
New prediction: $F_2(x)$



Training a Gradient Boosting model

	Attendance of lectures	Attendance of labs	Prior course	ML1 Mark	ML2 Mark	Residual 1	Predict 1	Residual 2	Predict 2	Residual 3
1	16	10	Yes	85	90	17	74.1	15.9	75.09	14.91
2	4	4	Yes	60	55	-18	71.2	-16.2	69.58	-14.58
3	15	6	Yes	85	84	11	74.1	9.9	73.11	10.89
4	7	8	No	45	64	-9	72.1	-8.1	71.425	-7.425
5	17	10	No	90	78	5	74.1	3.9	75.09	2.91
6	8	7	Yes	68	67	-6	72.4	-5.4	71.725	-4.725

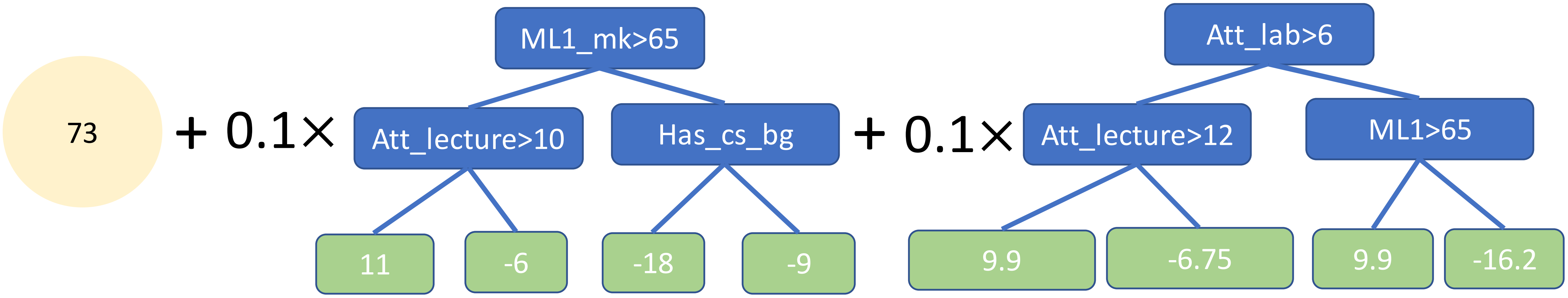
New prediction: $F_2(x)$



Now let's predict a new sample

	Attendance of lectures	Attendance of labs	Prior course	ML1 Mark	ML2 Mark
	14	5	Yes	80	?

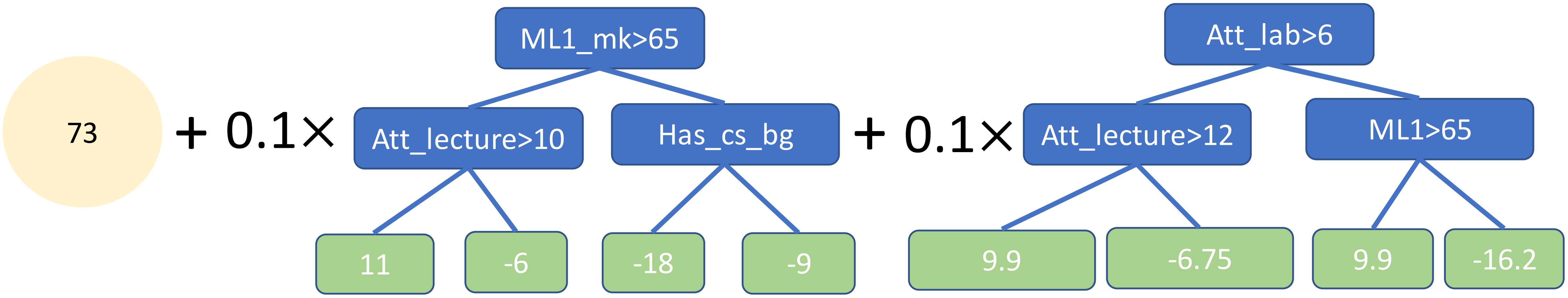
Final model: $F_2(x)$



Now let's predict a new sample

	Attendance of lectures	Attendance of labs	Prior course	ML1 Mark	ML2 Mark
	14	5	Yes	80	75.09

Final model: $F_2(x)$



Topic 5:
Gradient Boosting

Step 1 Initialize model $F_0(x)$ with a constant value for prediction.

$F_0(x) =$ The average of the observed values

WHY?

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

MSE loss function for regression:

$$J = \frac{1}{2}\sum_{i=1}^{n}(y_i - \gamma)^2$$

$$\frac{\partial J}{\partial \gamma} = 0$$

$$\frac{\partial J}{\partial \gamma} = \frac{2}{2}\sum_{i=1}^{n}(y_i - \gamma) = 0$$

$$\Rightarrow \quad \gamma = \frac{1}{n}\sum_{i=1}^{n} y_i$$

2. for $m = 1$ *to* M:

Step 2-1 Calculate the residual for each sample (Observed – Predicted)

WHY (Observed – Predicted)?

Negative gradient

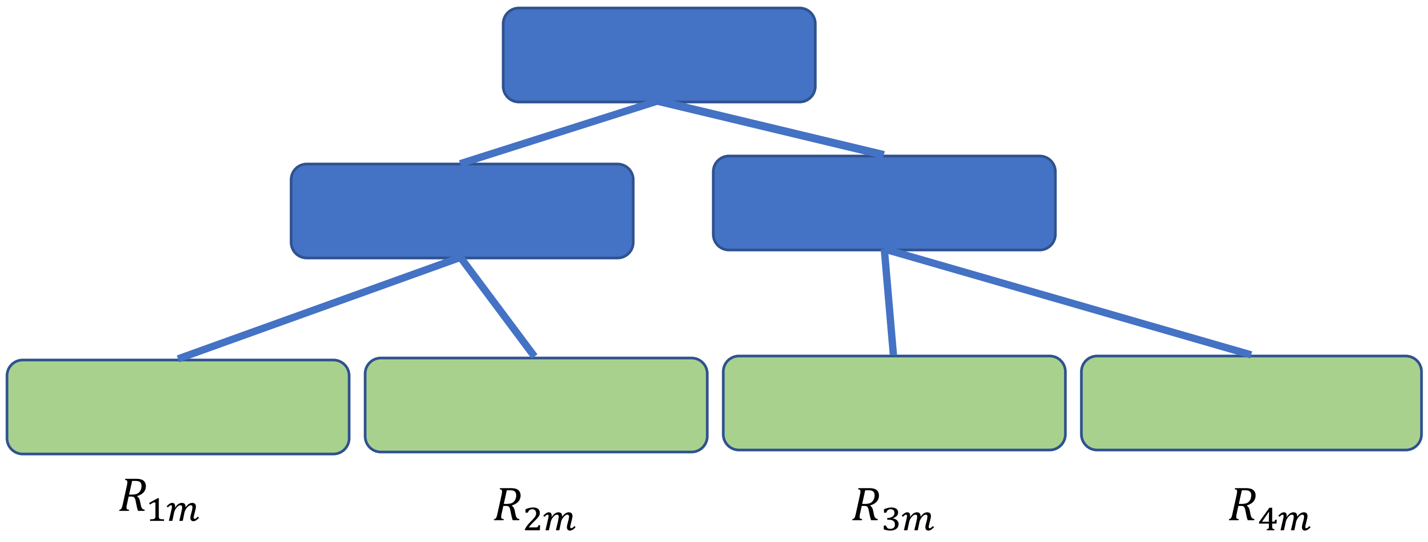
$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

$$L(y_i - F(x_i)) = \frac{1}{2}(y_i - F(x_i))^2$$

$$r_{im} = \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}$$

$$r_{im} = y_i - F_{m-1}(x_i)$$

Step 2-2 Build a decision tree $h_m(x)$ to fit the residuals and create terminal nodes $R_{jm}, j = 1, \dots, J_m$



Step 2-3 The output value γ_{jm} of j-th leaf is the average of all the residuals in that leaf.

WHY average of all the residuals?

$$\gamma_{jm} = \underset{\gamma}{argmin} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$

$$= \underset{\gamma}{argmin} \sum_{x_i \in R_{jm}} (y_i - F_{m-1}(x_i) - \gamma)^2$$

$$\frac{\partial}{\partial \gamma} \sum_{x_i \in R_{jm}} (y_i - F_{m-1}(x_i) - \gamma)^2 = 0$$

$$-2 \sum_{x_i \in R_{jm}} (y_i - F_{m-1}(x_i) - \gamma) = 0$$

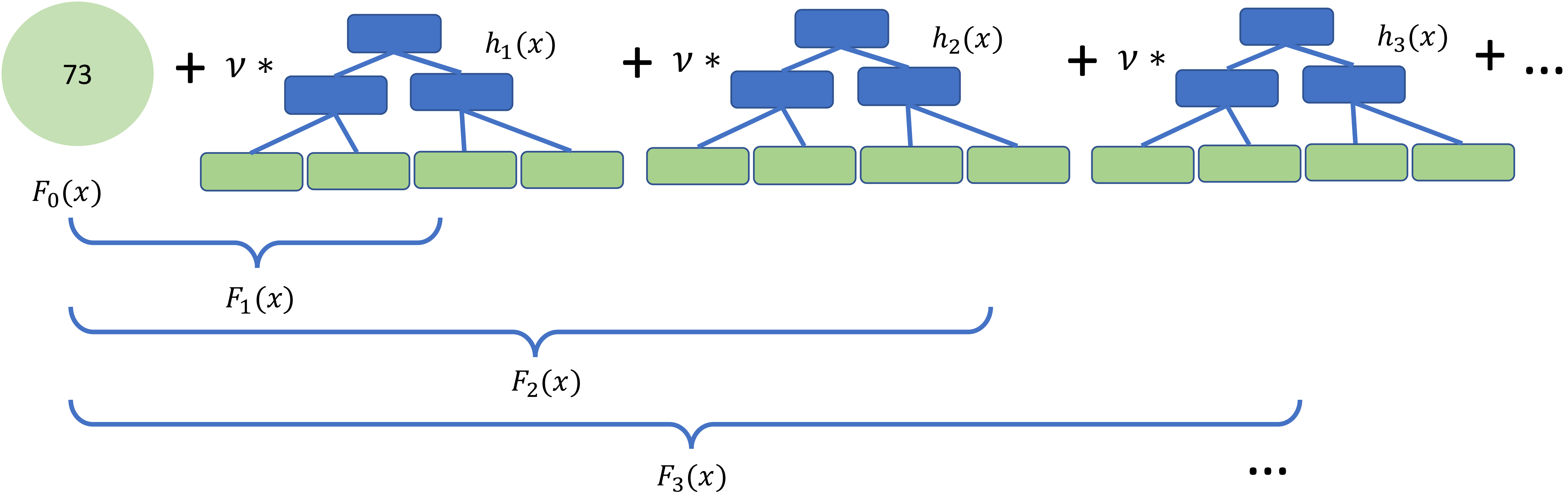
$$n_j \gamma = \sum_{x_i \in R_{jm}} (y_i - F_{m-1}(x_i))$$

$$\gamma = \frac{1}{n_j} \sum_{x_i \in R_{jm}} r_{im}$$

The number of samples in j-th leaf

Step 2-4 Update the model by adding the current tree with a factor.

$$F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} 1(x \in R_{jm})$$



Gradient Boosting Algorithm

1. Initialize model with a constant value:

$$F_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, \gamma)$$

2. for $m = 1\ to\ M$:

2-1. Compute residuals $r_{im} = -\left[\frac{\partial L(y_i,\ F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad for\ i = 1,\ldots,n$

Negative gradient

2-2. Train regression tree with features x against r and create terminal node reasions R_{jm} for $j = 1,\ldots,J_m$

2-3. Compute $\gamma_{jm} = \underset{\gamma}{argmin} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma) \quad for\ j = 1,\ldots,J_m$

2-4. Update the model:

$$F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} 1(x \in R_{jm})$$

Gradient boosting variants

- Through this approach, one could consider a variety of loss functions.
  - MSE loss, MAE loss
  - Cross-entropy loss

$$L = -\left(y_i \cdot log(p) + (1 - y_i) \cdot log(1 - p)\right)$$

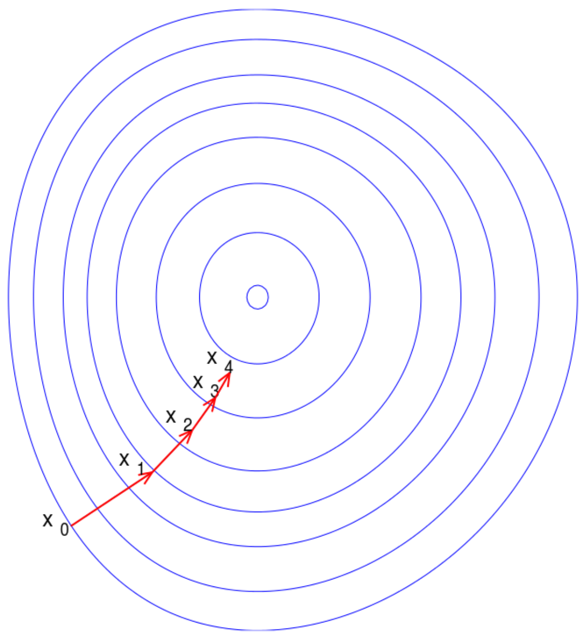
Regression problem

Classification problem

How is Gradient boosting related to Gradient Descent?

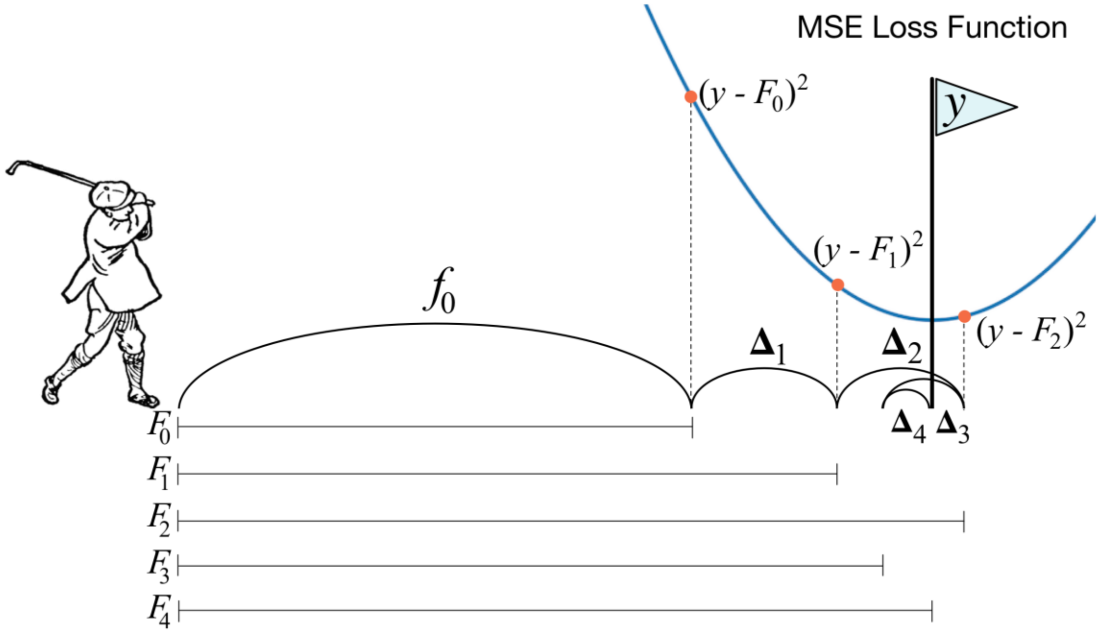
In Gradient Descent, if we want to minimise a function $f(x)$, the fastest way is to proceed in the direction of **negative of the gradient** of function

$$x_t = x_t - \eta \nabla f(x)$$



In Gradient boosting, we aim to minimize some loss function between the observed values and the prediction. We update the prediction by fitting the residual – **negative of gradient**.

$$F_m(x) = F_{m-1}(x) + v h_m(x)$$



Img from: https://en.wikipedia.org/wiki/Gradient_descent

Img from: https://explained.ai/gradient-boosting/index.html

Q: What is the similarity and different between Adaboost and Gradient boost?

References

- Statquest videos that explain the Gradient boosting: https://www.youtube.com/watch?v=3CC4N4z3GJc
- Tomonori Masui's blog: All you need to know about gradient boosting algorithm. https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502
- Terence Parr and Jeremy Howard, How to explain gradient boosting. https://explained.ai/gradient-boosting/index.html