Gradient Boost-(ish)

Regularization

A Unique Regression Tree

Approximate Greedy Algorithm

Weighted Quantile Sketch

Sparsity-Aware Split Finding

Parallel Learning

Cache-Aware Access

Blocks for Out-of-Core Computation

In this video, **Part 2**, we'll give an overview of how **XGBoost Trees** are built for **Classification**.

And in **Part 3**, we'll dive into the mathematical details to show you how **Regression** and **Classification** are related and why creating unique trees makes so much sense.

To find the O_{value} that minimizes…

$$\left[\sum_{i=1}^{n} L(y_i, p_i^0 + O_{value})\right] + \gamma + \frac{1}{2}\lambda O_{value}^2$$

1) Take the derivative…

$$\frac{d}{dO_{value}}(g_1 + g_2 + g_3)O_{value} + \frac{1}{2}(h_1 + h_2 + h_3 + \lambda)O_{value}^2$$

2) Set to zero…

$$(g_1 + g_2 + g_3) + (h_1 + h_2 + h_3 + \lambda)O_{value} = 0$$

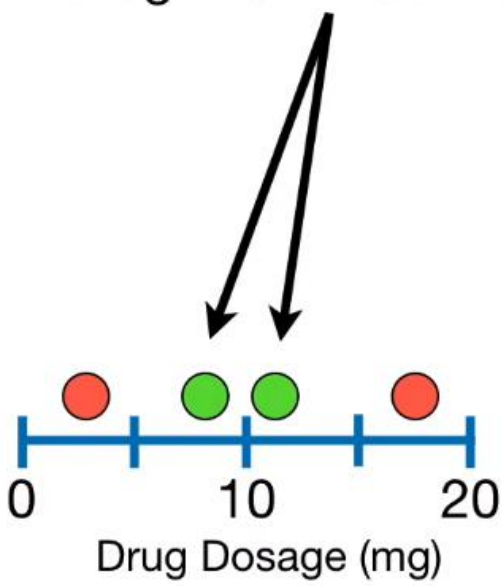
3) Solve for $f_k(x_i)$…

$$O_{value} = \frac{-(g_1 + g_2 + g_3)}{(h_1 + h_2 + h_3 + \lambda)}$$

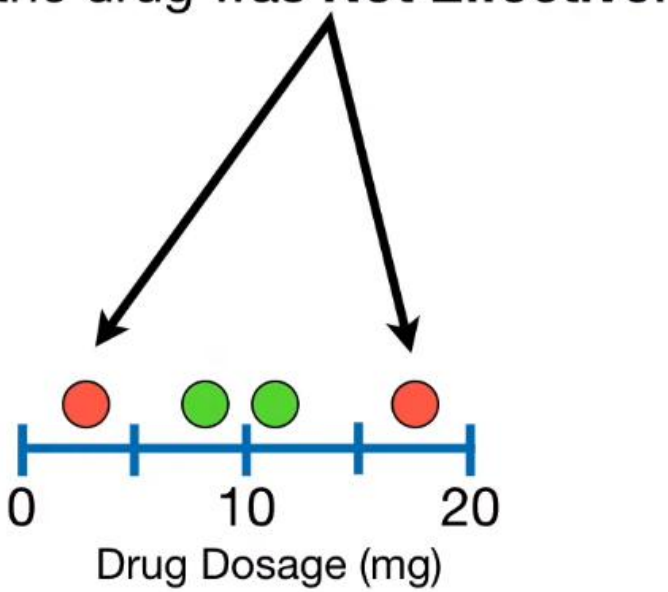
NOTE: XGBoost was designed to be used with large, complicated data sets.

Mass	Age	BP	Color	Movie	Car	Hair	etc.
120	23	102	Brown	T2	Ford	Long	…
150	25	98	Brown	Frozen	Kia	Short	…
165	22	130	Black	Spiderman	Ford	Short	…
123	45	98	Red	T2	Kia	Long	…
156	33`	78	Brown	Frozen 2	Ford	Long	…
…	…	…	…	…	…	…	…

The **Green Dots** indicate that the drug was **Effective**…



…and the **Red Dots** indicate that the drug was **Not Effective**.



Predicted Drug Effectiveness

0.5

The very first step in fitting **XGBoost** to the **Training Data** is to make an initial prediction.

Predicted Drug Effectiveness

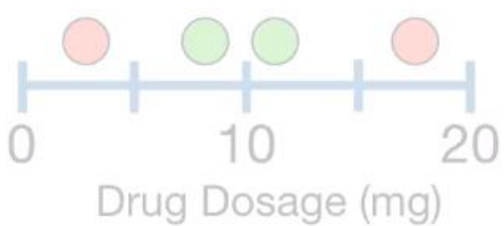
0.5

This prediction can be anything, for example, the **probability** of observing an effective dosage in the **Training Data**, but by default it is **0.5**, regardless of whether you are using **XGBoost** for **Regression** or **Classification**.

Predicted Drug Effectiveness

0.5

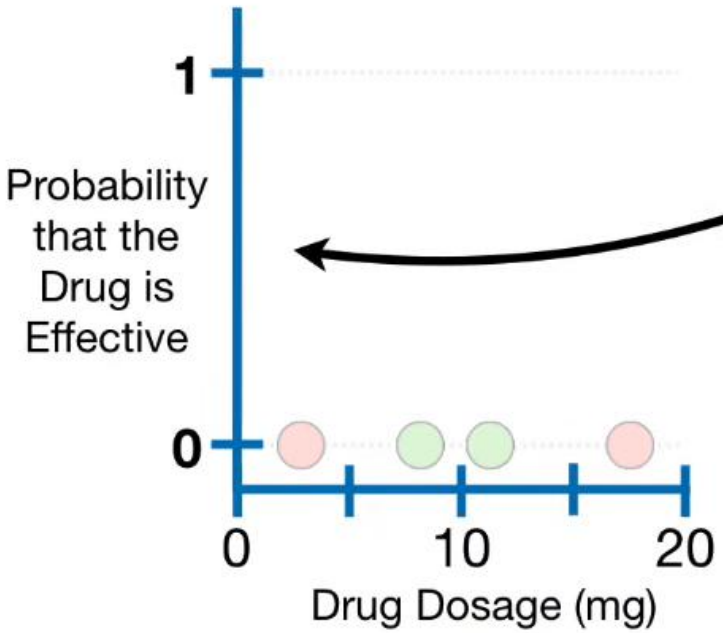
In other words, regardless of the **Dosage**, the default prediction is that there is a **50%** chance the drug is **Effective**.



Predicted Drug Effectiveness

0.5

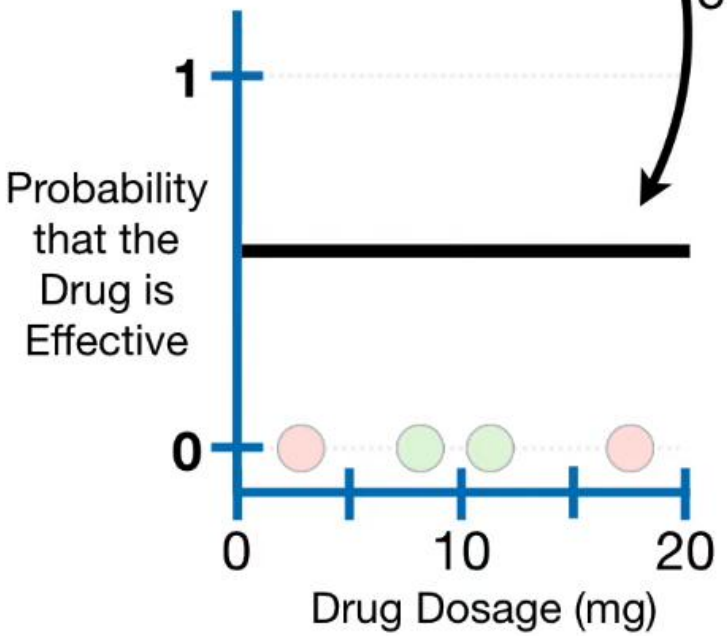
We can illustrate the initial prediction by adding a ***y*-axis** to our graph to represent the **Probability that the Drug is Effective**…



Predicted Drug Effectiveness

0.5

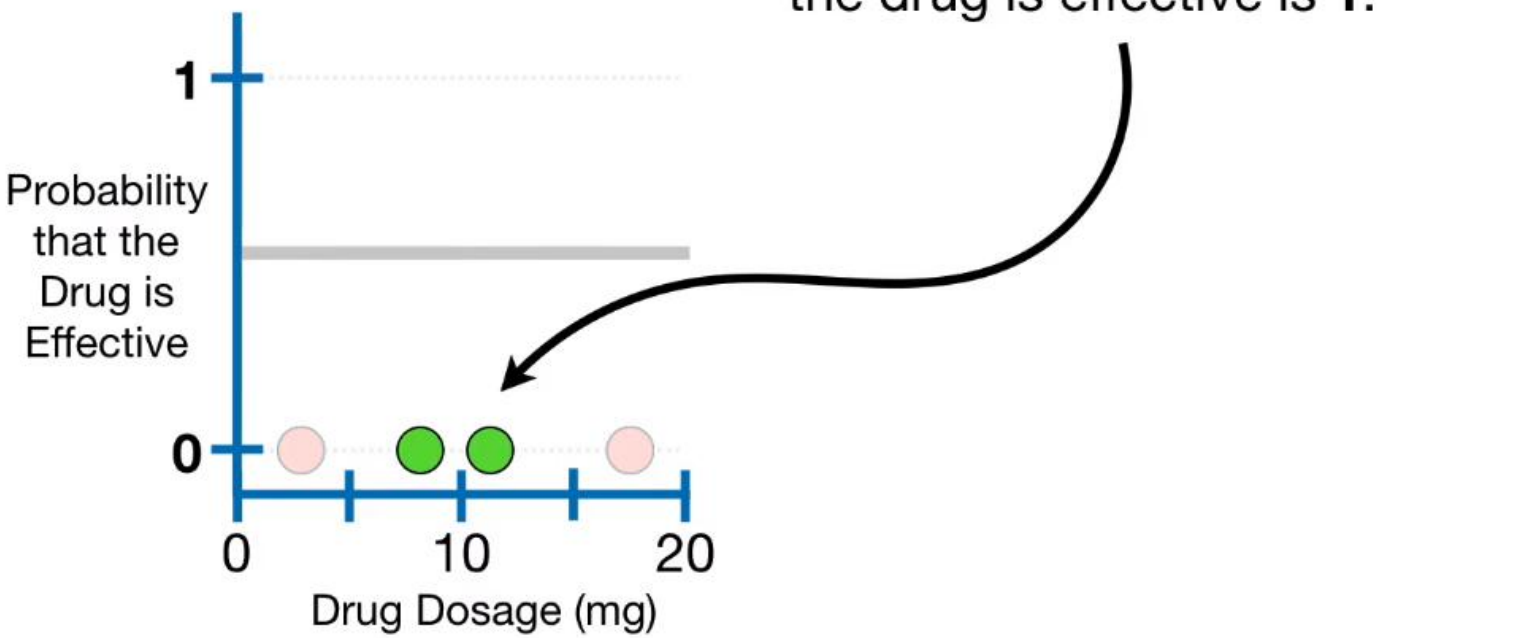
…and drawing a **thick black line** at **0.5** to represent a **50%** chance that the drug is effective.



Predicted Drug Effectiveness

0.5

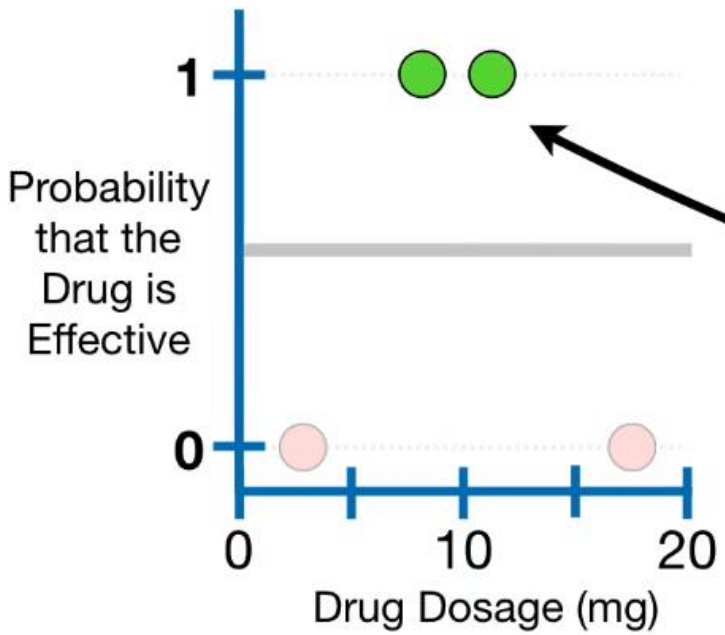
Since these two **Green Dots** represent effective dosages, we will move them to the top of the graph, where the probability that the drug is effective is **1**.



Predicted Drug Effectiveness

0.5

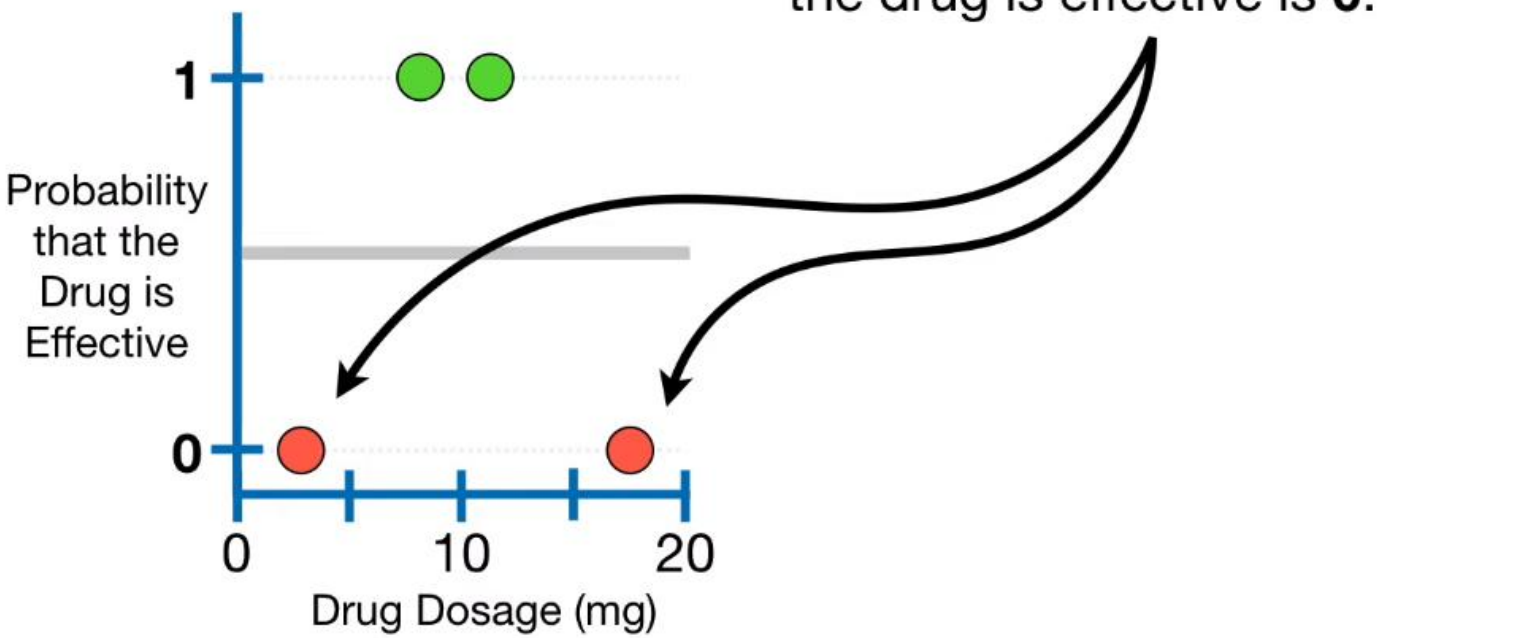
Since these two **Green Dots** represent effective dosages, we will move them to the top of the graph, where the probability that the drug is effective is **1**.



Predicted Drug Effectiveness

0.5

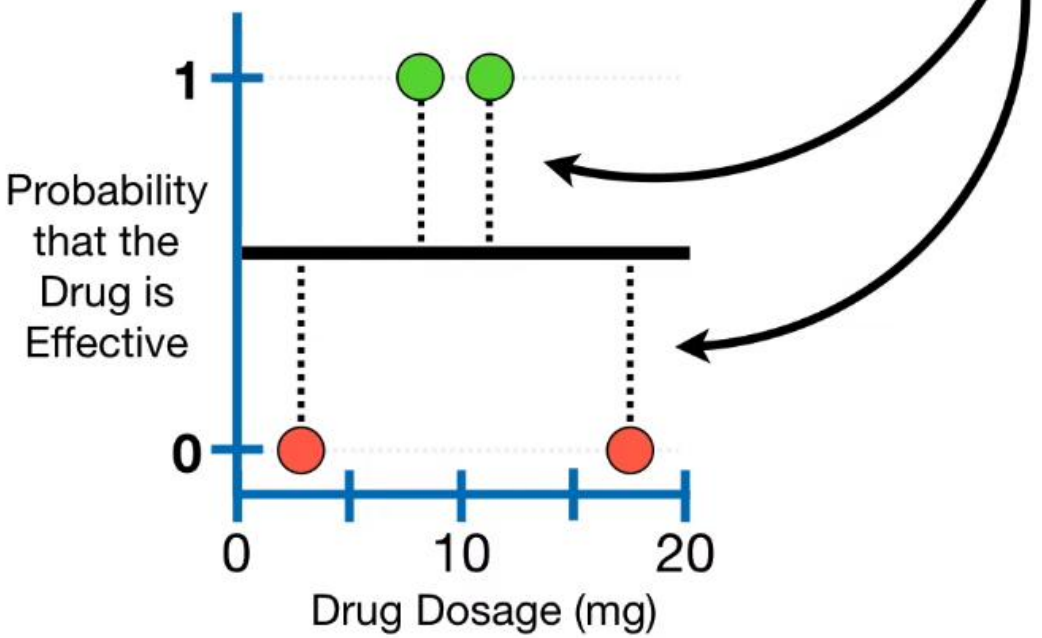
These two **Red Dots** represent ineffective dosages, so we will leave them at the bottom of the graph, where the probability that the drug is effective is **0**.



Predicted Drug Effectiveness

0.5

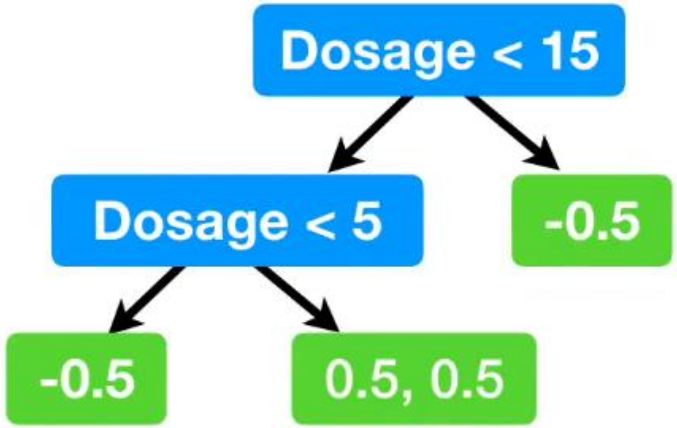
The **Residuals**, the differences between the **Observed** and **Predicted** values, show us how good the initial prediction is.

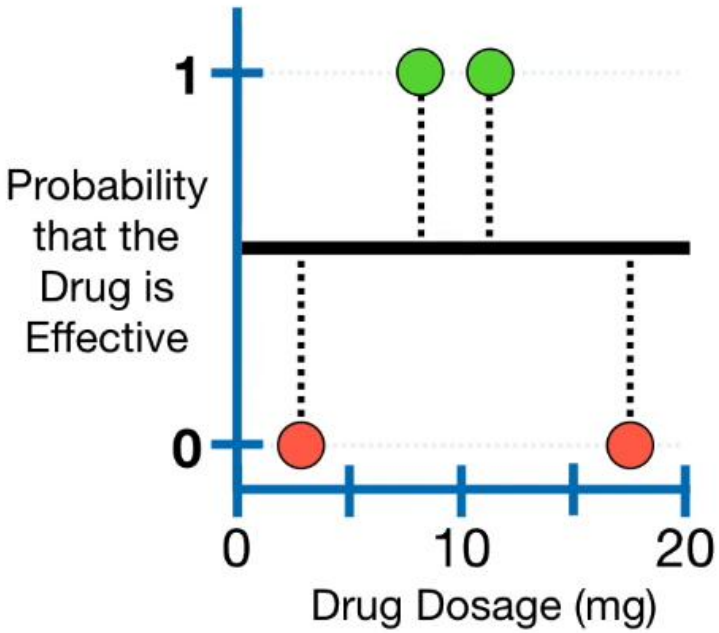


Predicted Drug Effectiveness

0.5

Now, just like we did for **Regression**, we fit an **XGBoost Tree** to the **Residuals**…

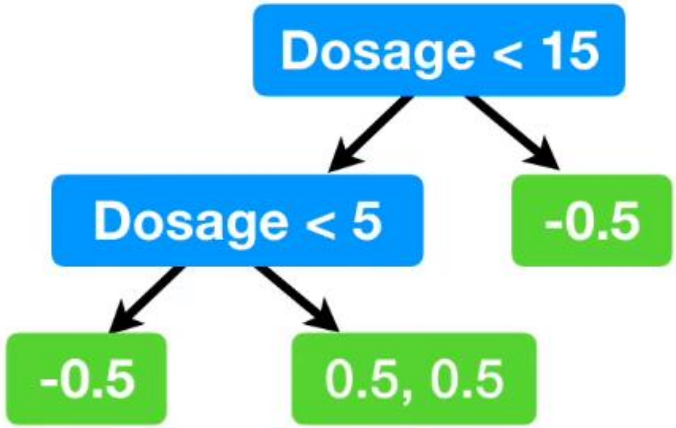


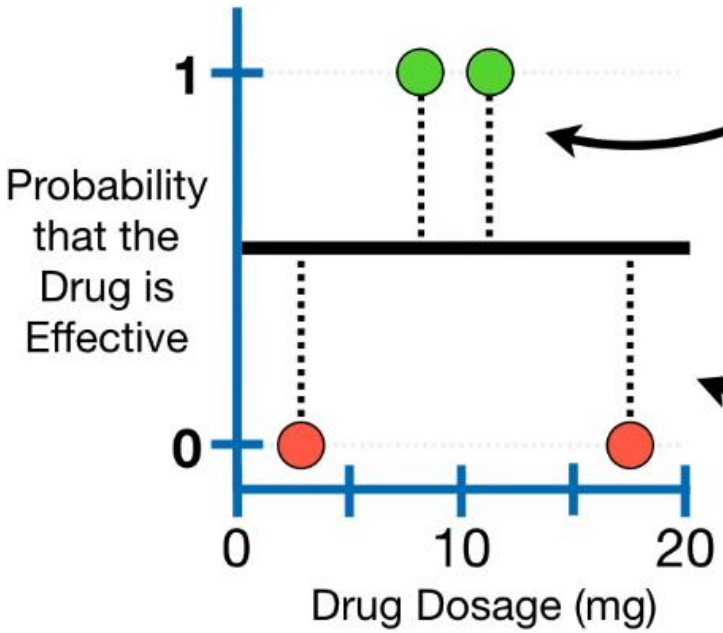


Predicted Drug Effectiveness

0.5

Now, just like we did for **Regression**, we fit an **XGBoost Tree** to the **Residuals**…

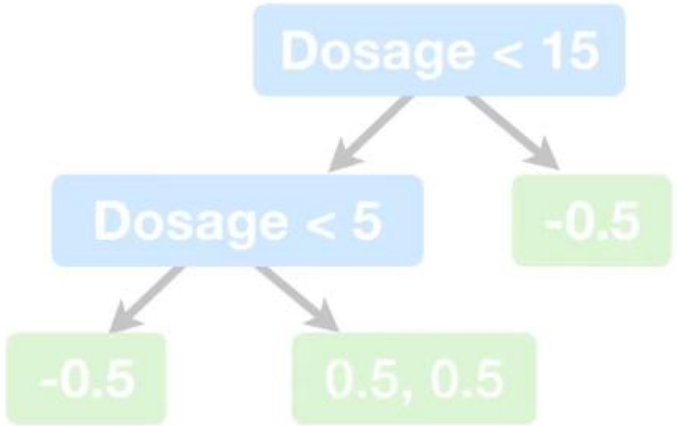


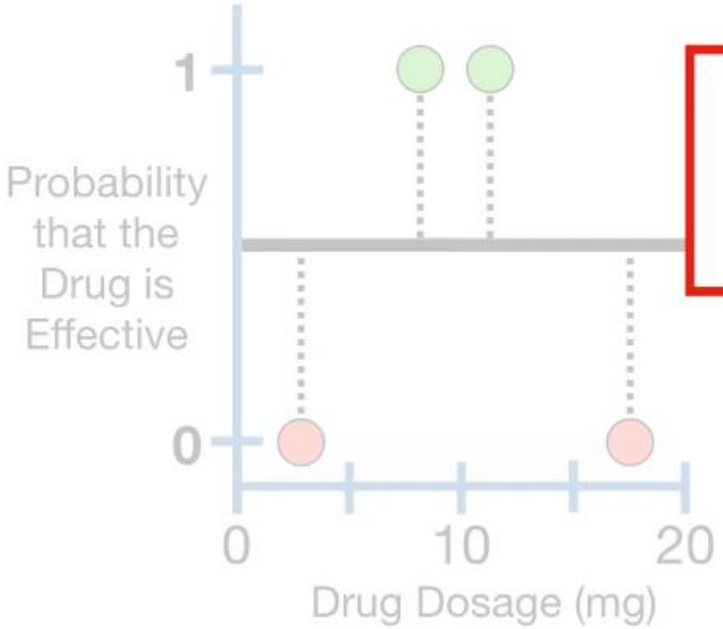


Predicted Drug Effectiveness

0.5

…however, since we are using **XGBoost** for **Classification**, we have a new formula for the **Similarity Scores**.



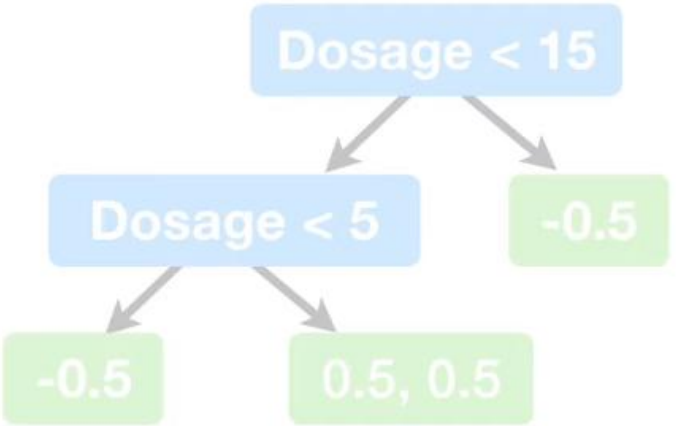


$$\frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

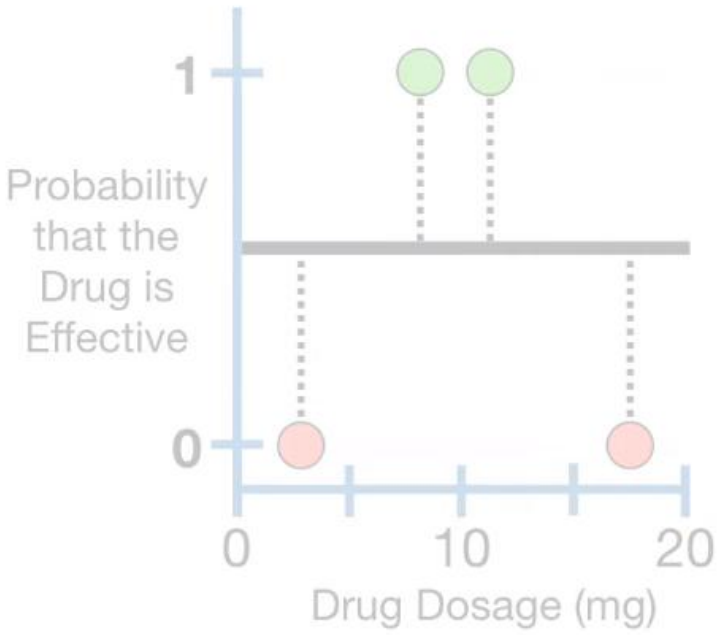
Predicted Drug Effectiveness

0.5

NOTE: Even though the numerator looks fancy, it's just the **Sum of the Residuals, Squared**.



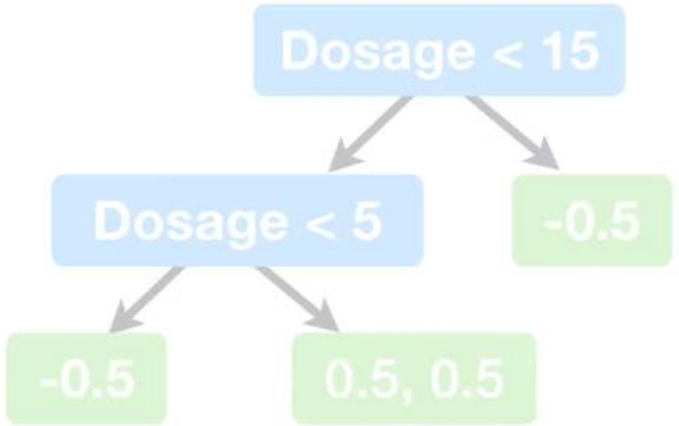
$$\frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

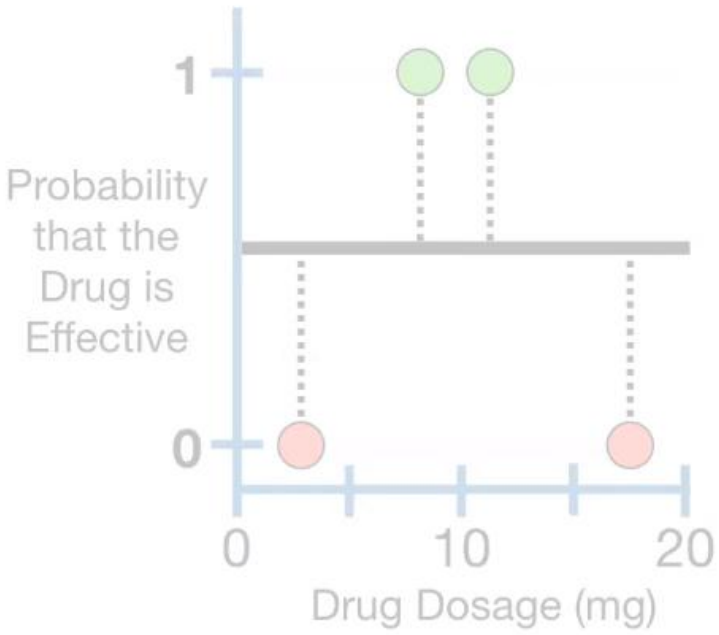


Predicted Drug Effectiveness

0.5

In other words, the numerator for **Classification** is the same as the numerator for **Regression**.



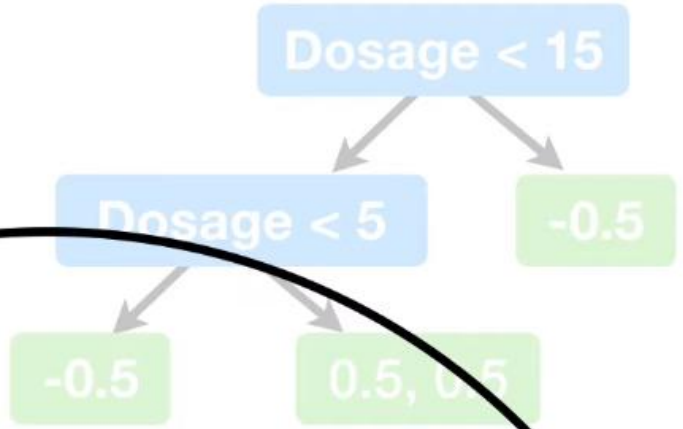


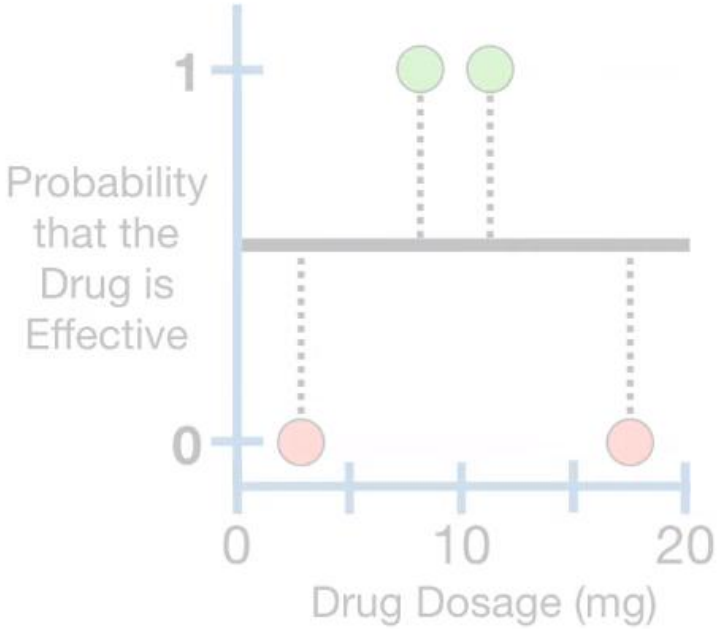
$$\frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

Predicted Drug Effectiveness

0.5

And just like for **Regression**, the denominator contains **λ (lambda)**, the **Regularization Parameter**…

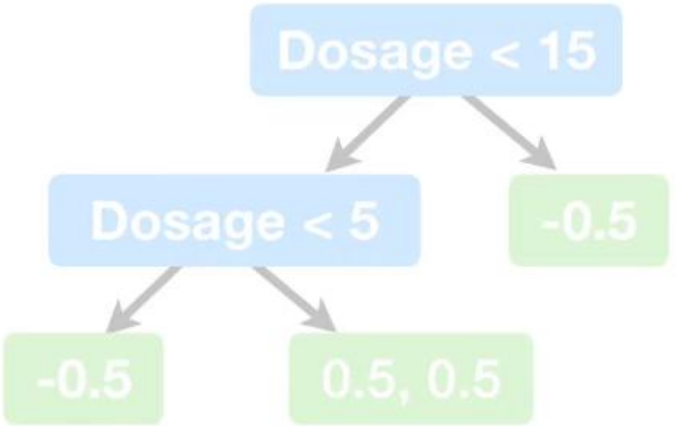




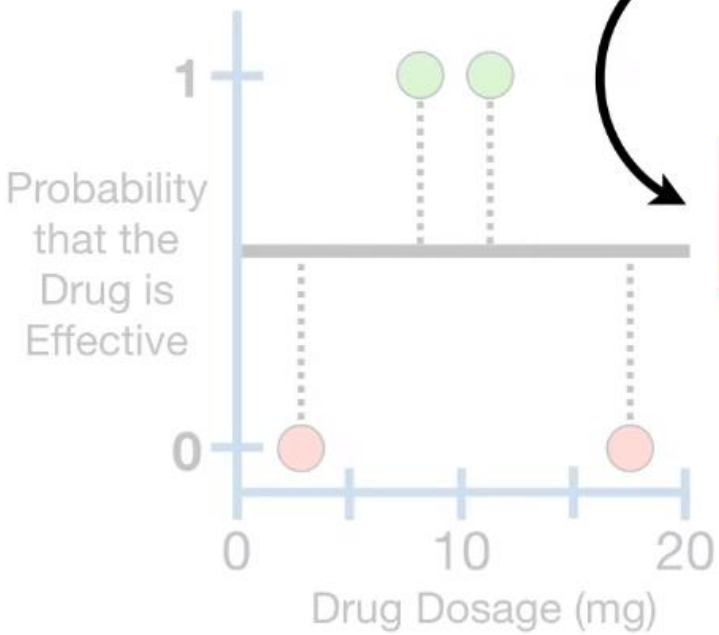
$$\frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

Predicted Drug Effectiveness

0.5



…however, the rest of the denominator is different.

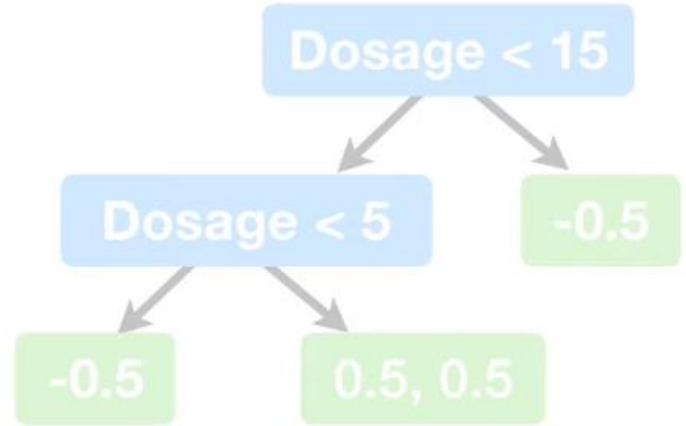


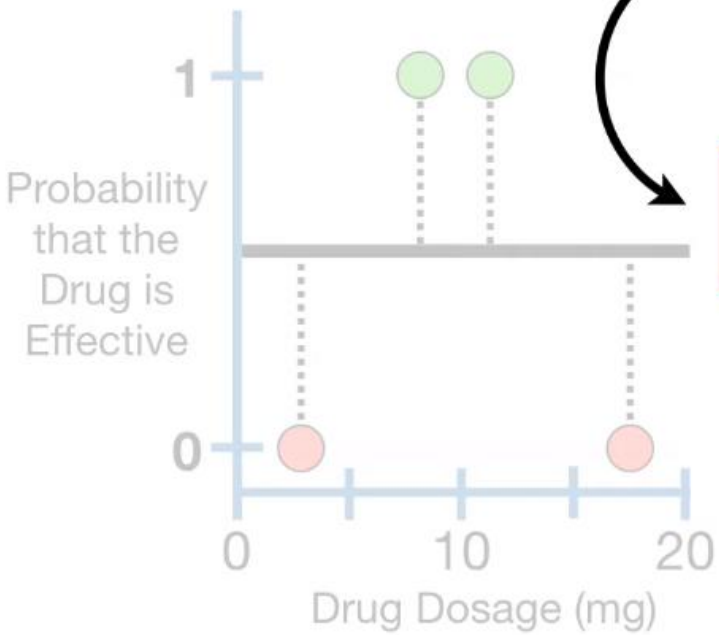
$$\frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

The good news is that we already saw something just like this in regular, unextreme **Gradient Boost**.

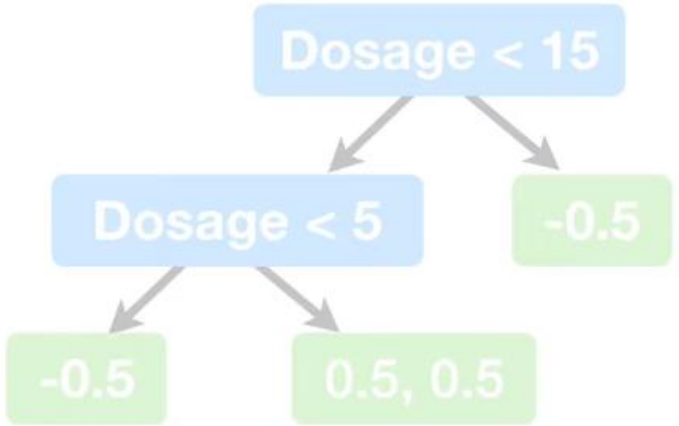




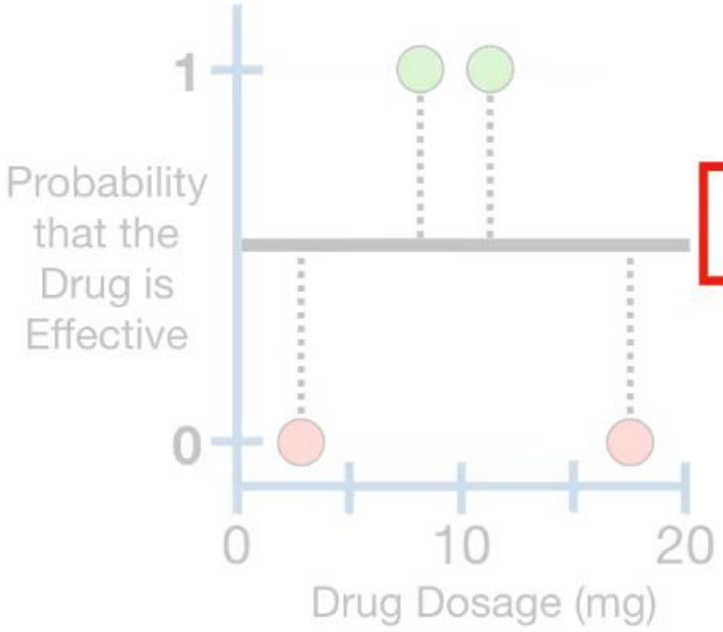
$$\frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5



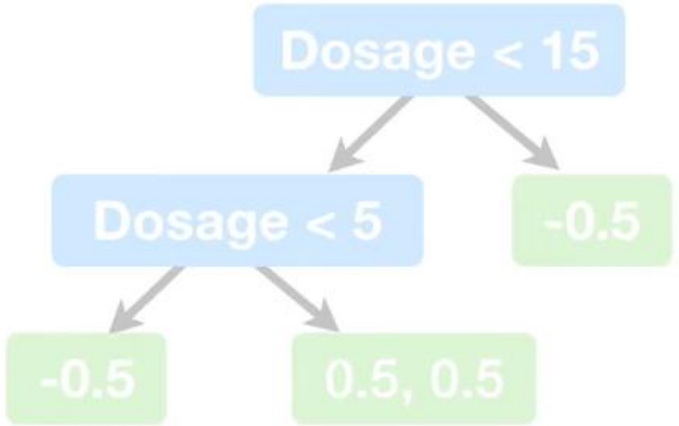
It's just the sum, for each observation…



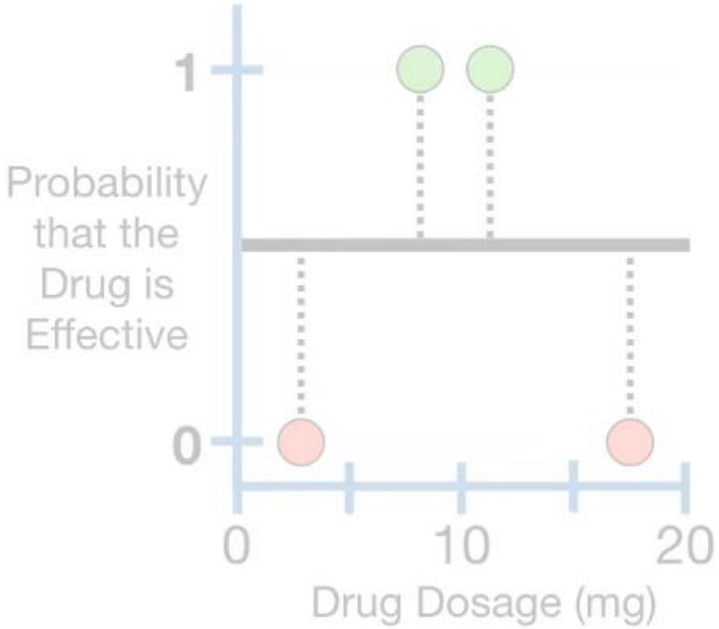
$$\frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5



…of the previously predicted probability…

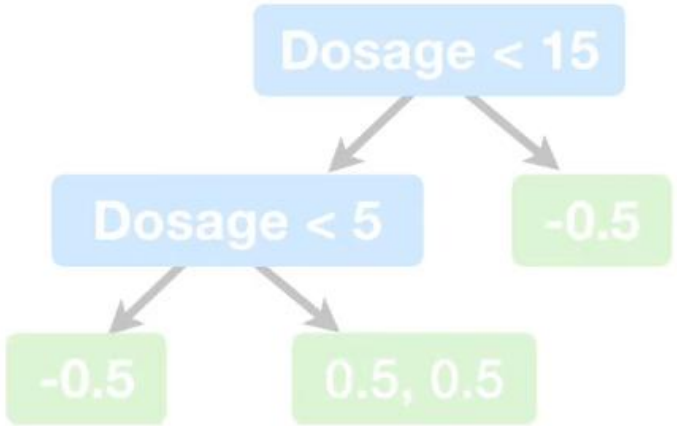


$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda$$

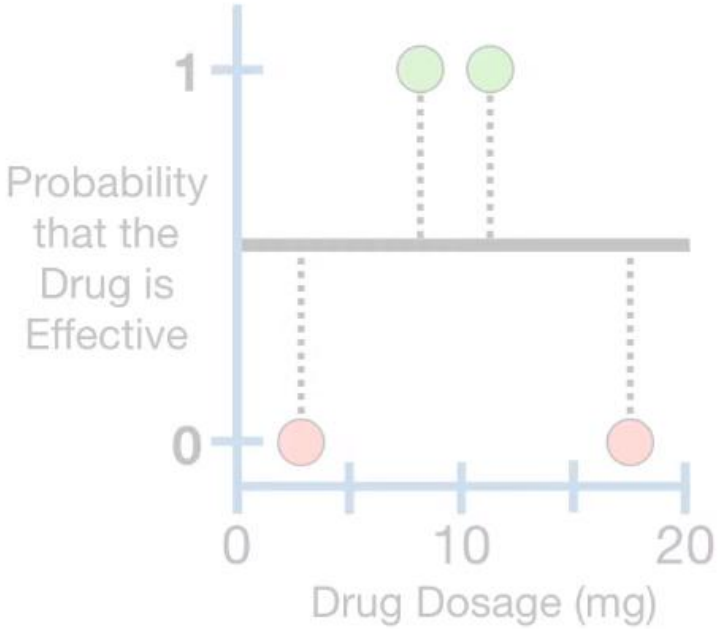
$$(\sum \text{Residual}_i)^2$$

Predicted Drug Effectiveness

0.5



…times **1** minus the previously predicted probability.

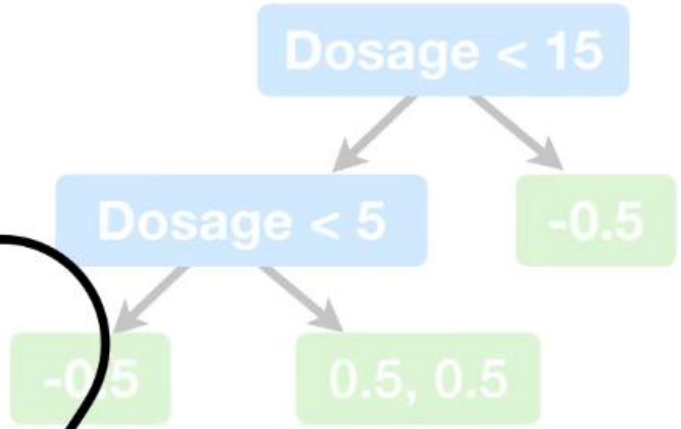


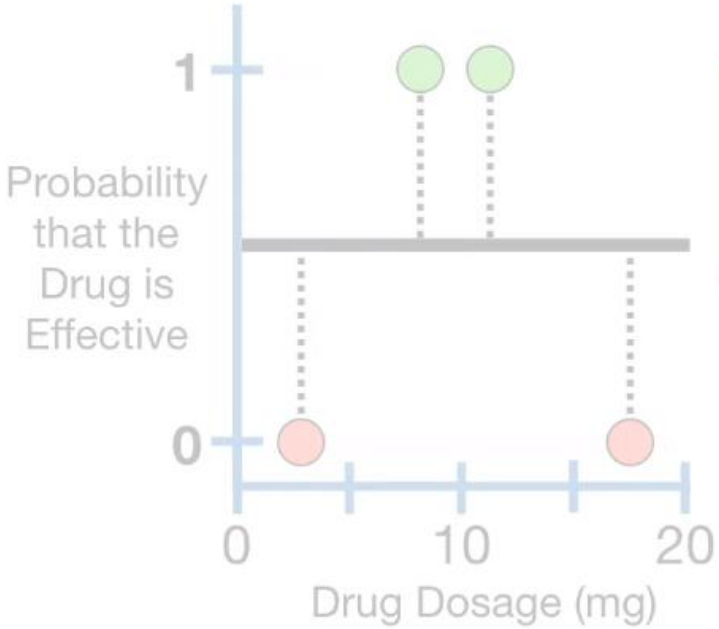
$$\frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

NOTE: Although this formula is different from what **XGBoost** uses for **Regression**, it is very closely related, and we'll show you why in **Part 3** when we get into the nitty gritty details.



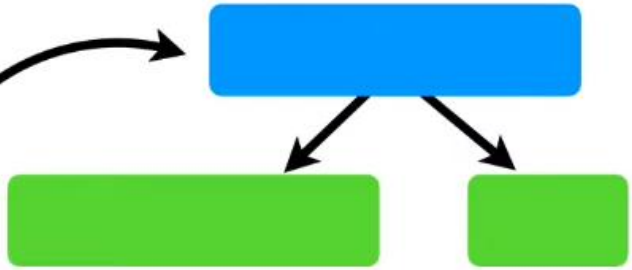


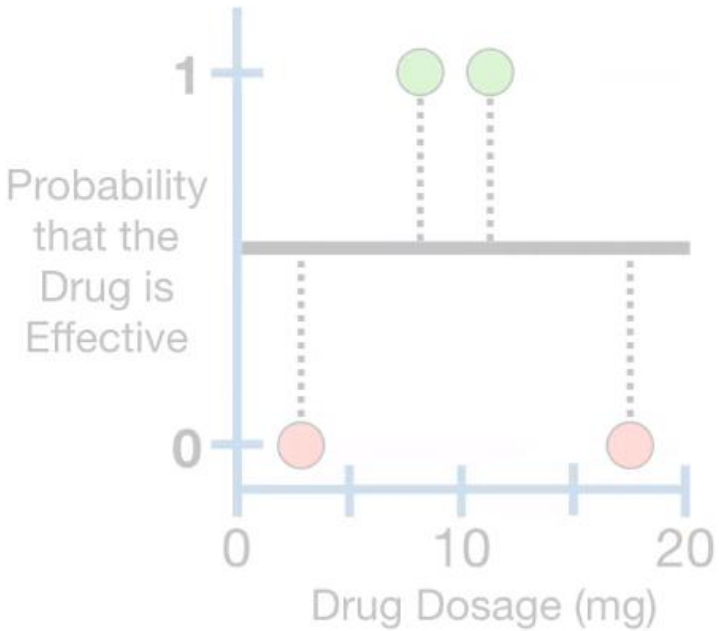
$$\frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

Now let's build a tree!!!



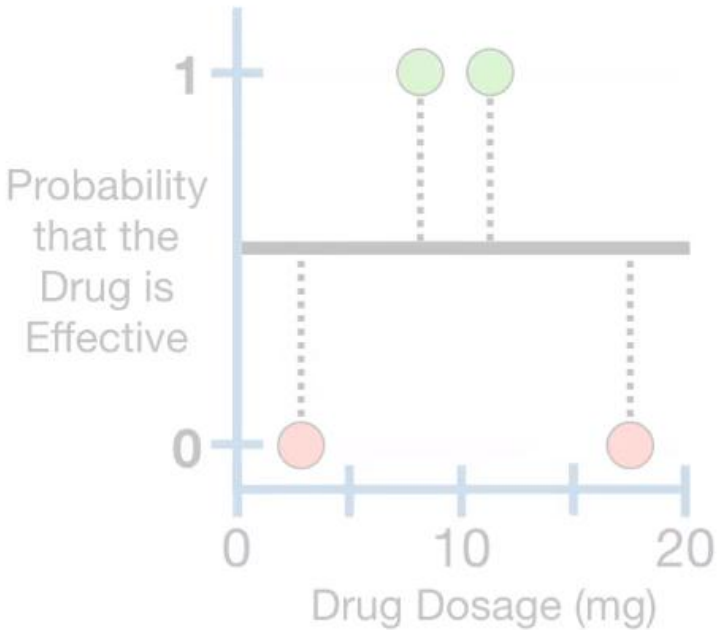


$$\frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

Just like for **Regression**, each tree starts out as a single leaf…



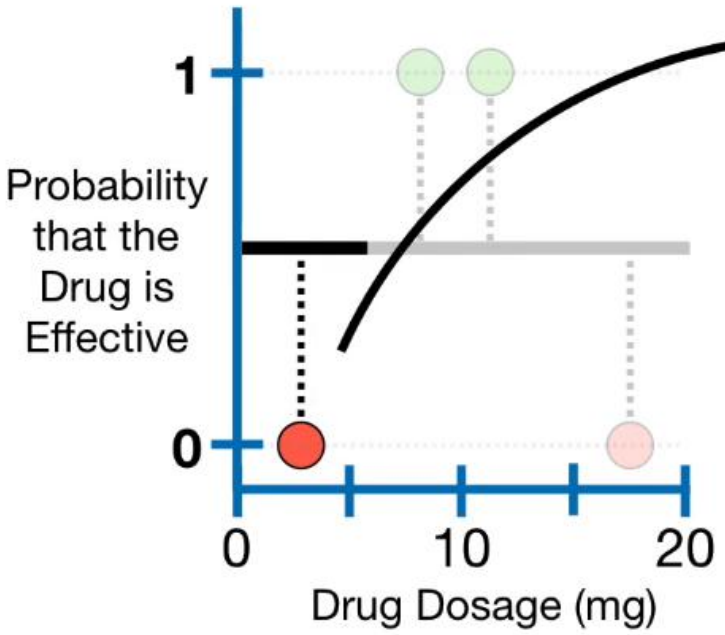
$$\frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

-0.5

…and all of the **Residuals** go to the leaf.



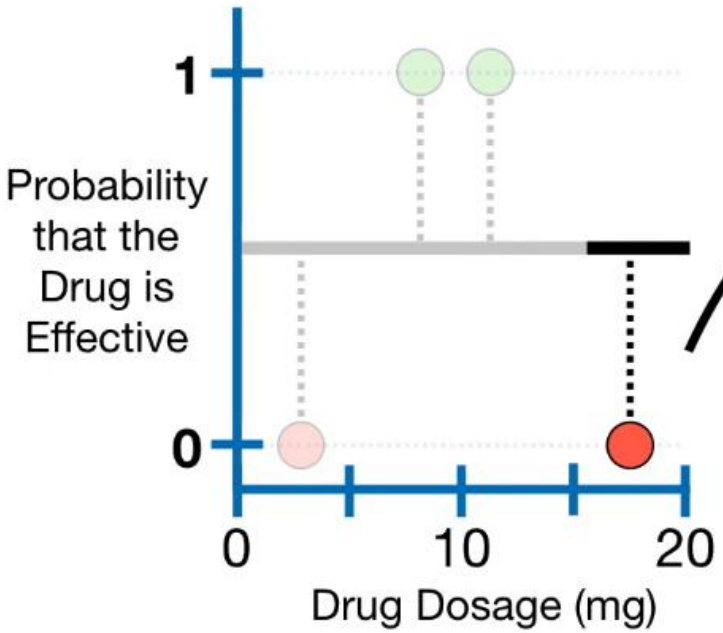
$$\frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

-0.5, 0.5, 0.5, -0.5

…and all of the **Residuals** go to the leaf.



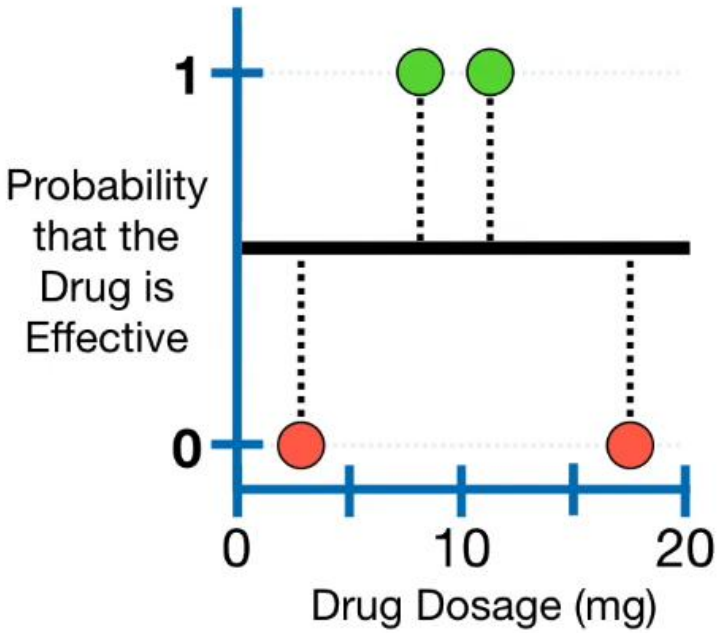
$$\frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

Predicted Drug Effectiveness

0.5

-0.5, 0.5, 0.5, -0.5

Now we need to calculate a **Similarity Score** for the leaf.



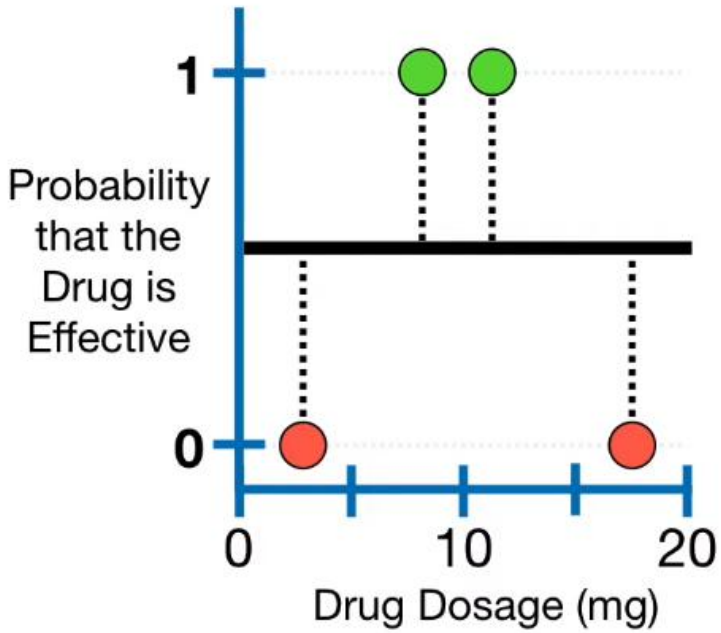
$$\frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

-0.5, 0.5, 0.5, -0.5

And that means we plug all **4 Residuals** into the numerator.



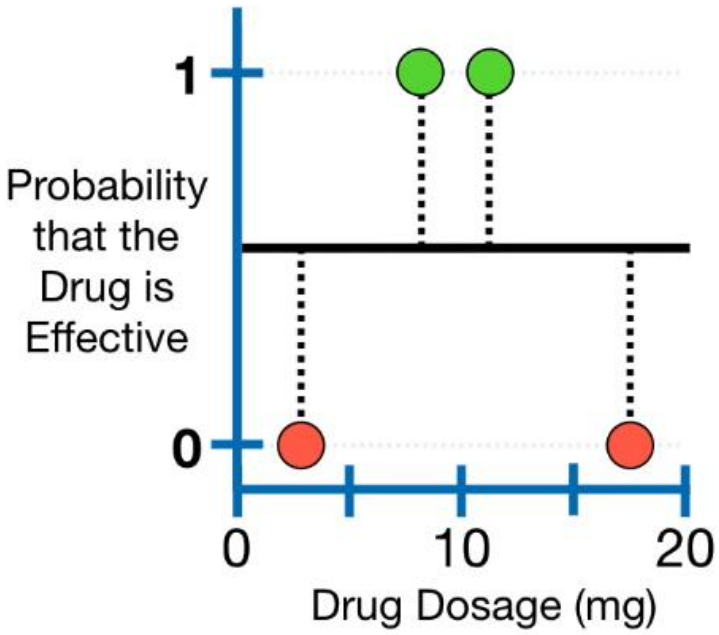
$$\frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

Predicted Drug Effectiveness

0.5

-0.5, 0.5, 0.5, -0.5

NOTE: Because we do not square the **Residuals** before we add them together, they will cancel each other out…



$$\frac{(-0.5 + 0.5 + 0.5 + -0.5)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

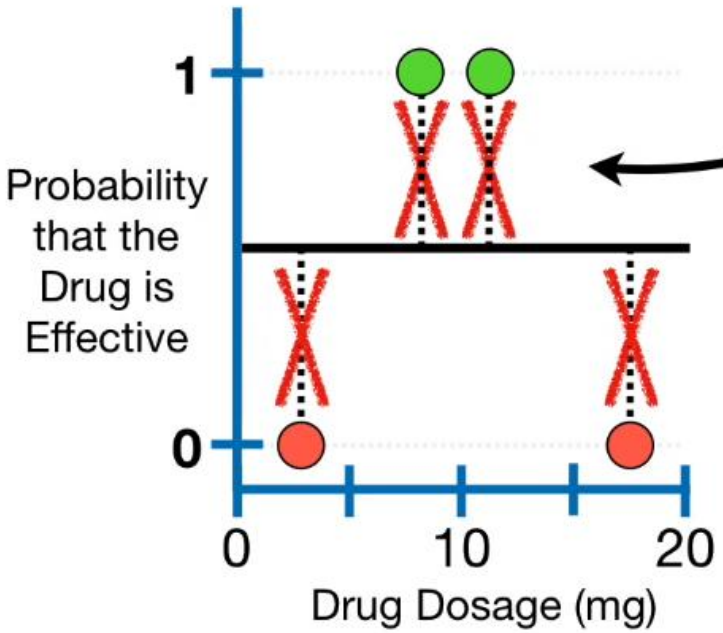
0.5

-0.5, 0.5, 0.5, -0.5

NOTE: Because we do not square the **Residuals** before we add them together, they will cancel each other out…

$$(-0.5 + 0.5 + 0.5 + -0.5)^2$$

$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda$$

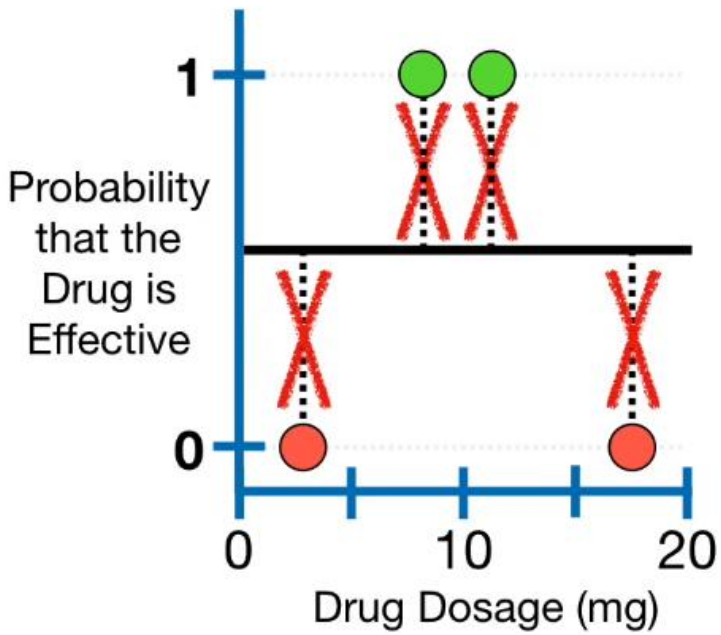


Predicted Drug Effectiveness

0.5

-0.5, 0.5, 0.5, -0.5

…and we will end up with **0** in the numerator and that makes the **Similarity Score = 0**.



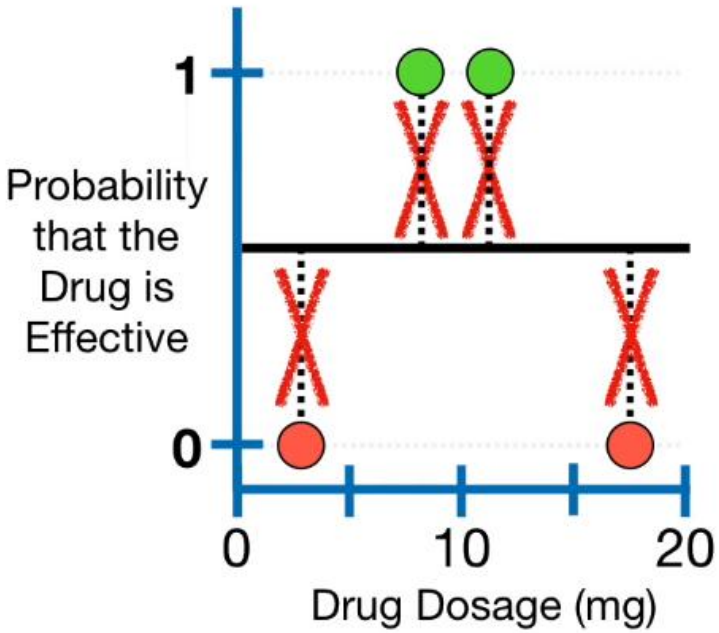
$$\frac{0}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

-0.5, 0.5, 0.5, -0.5

And that's a little bit of a bummer since it doesn't give us a chance to talk about the denominator, which is the interesting part.



0

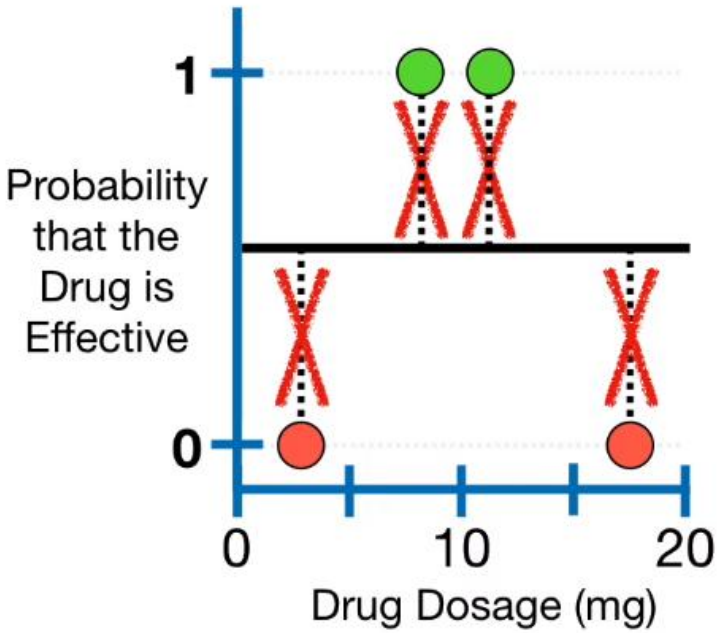
$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda$$

Predicted Drug Effectiveness

0.5

-0.5, 0.5, 0.5, -0.5

However, don't freak out, we'll get to it soon.



$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda$$

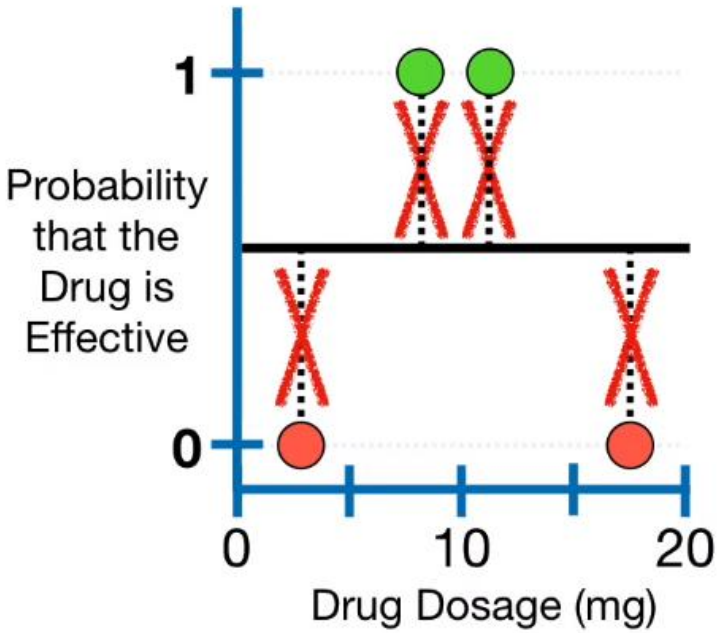
0

Predicted Drug Effectiveness

0.5

Similarity = 0 **-0.5, 0.5, 0.5, -0.5**

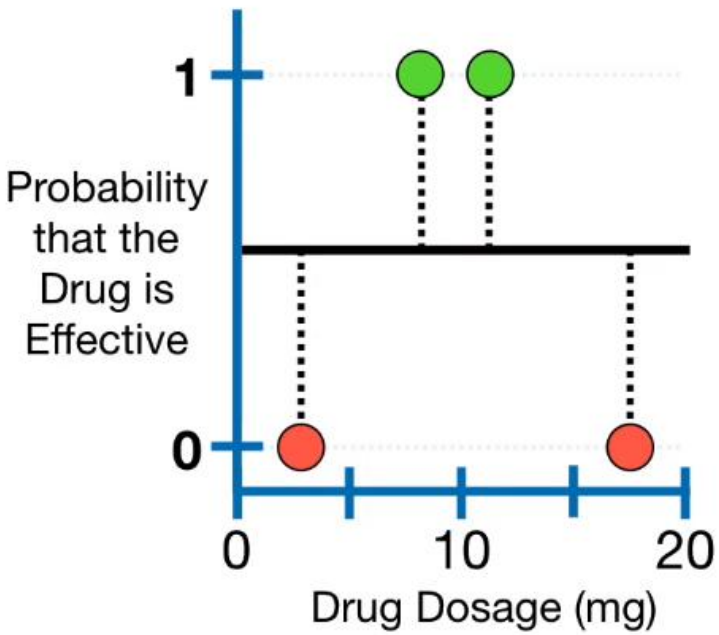
For now, let's just put **Similarity = 0** up here so we can keep track of it.



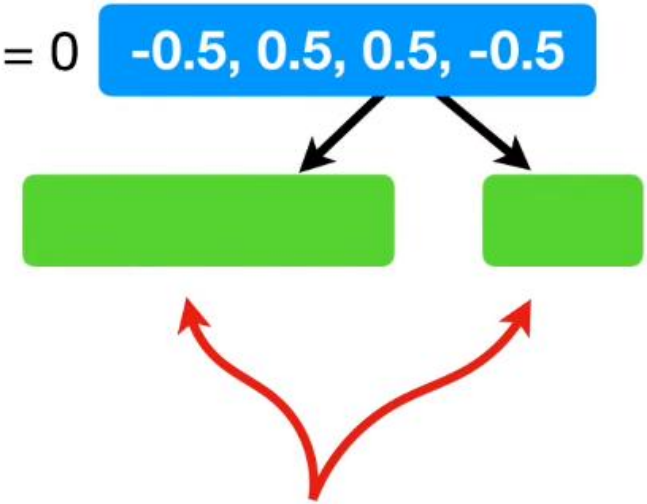
$$\frac{0}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5



Similarity = 0 **-0.5, 0.5, 0.5, -0.5**

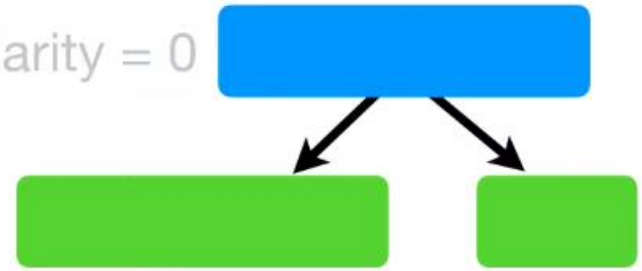


Now we need to decide if we can do a better job clustering similar **Residuals** if we split them into two groups.

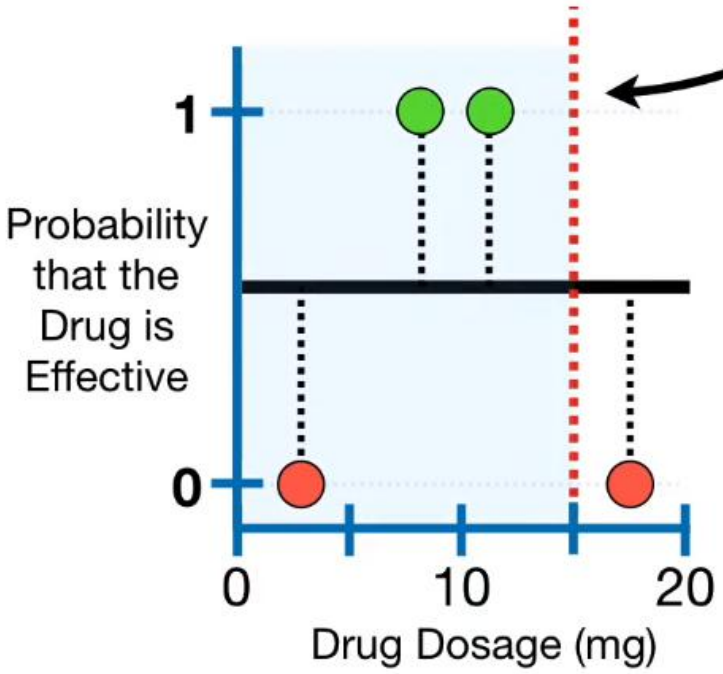
Predicted Drug Effectiveness

0.5

Similarity = 0



We'll start with this threshold, **Dosage < 15**.



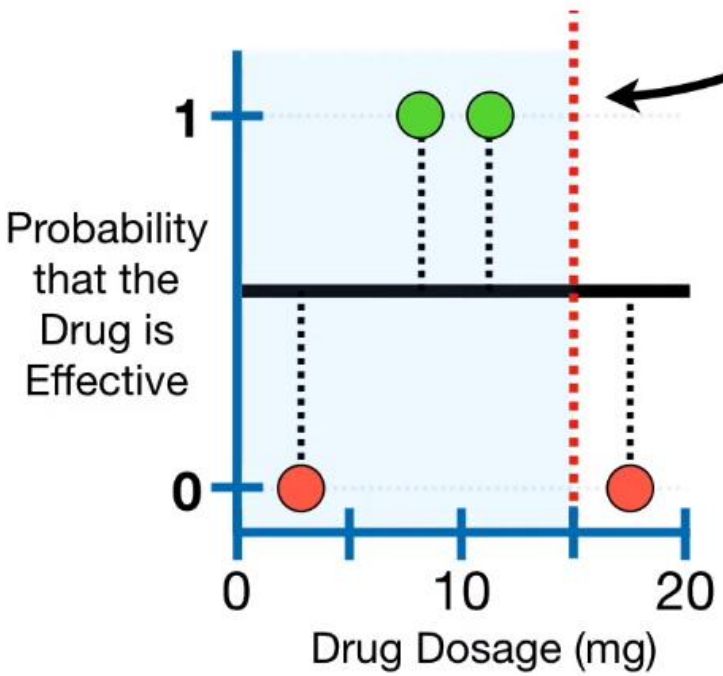
Predicted Drug Effectiveness

0.5

Similarity = 0

Dosage < 15

We'll start with this threshold, **Dosage < 15**.



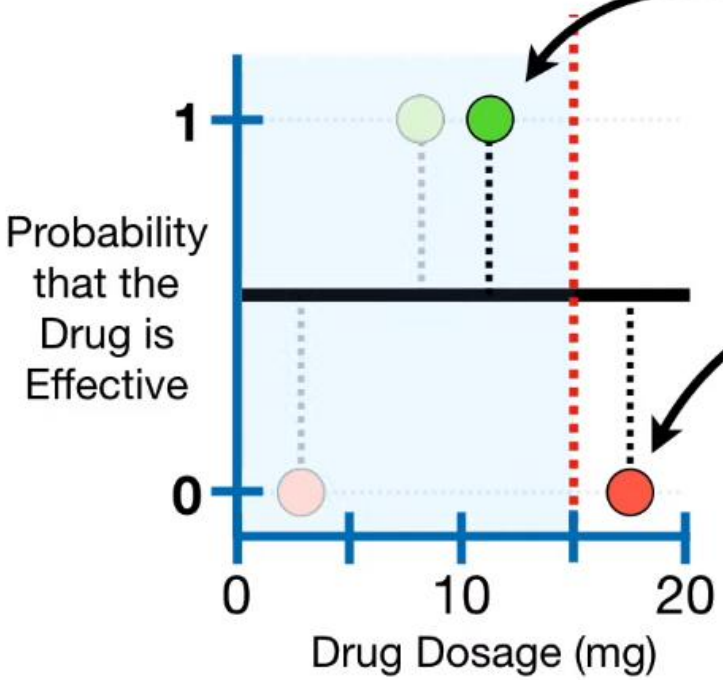
Predicted Drug Effectiveness

0.5

Similarity = 0

Dosage < 15

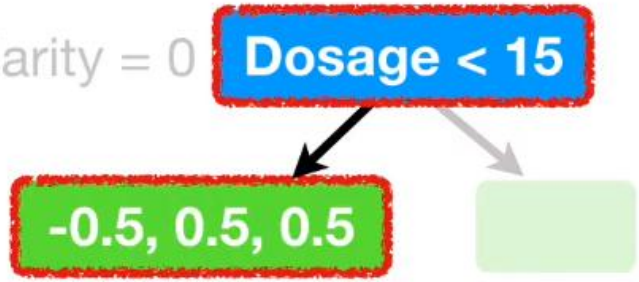
NOTE: We chose the threshold, **Dosage < 15**, because **15** is the average value between last two **2** observations.



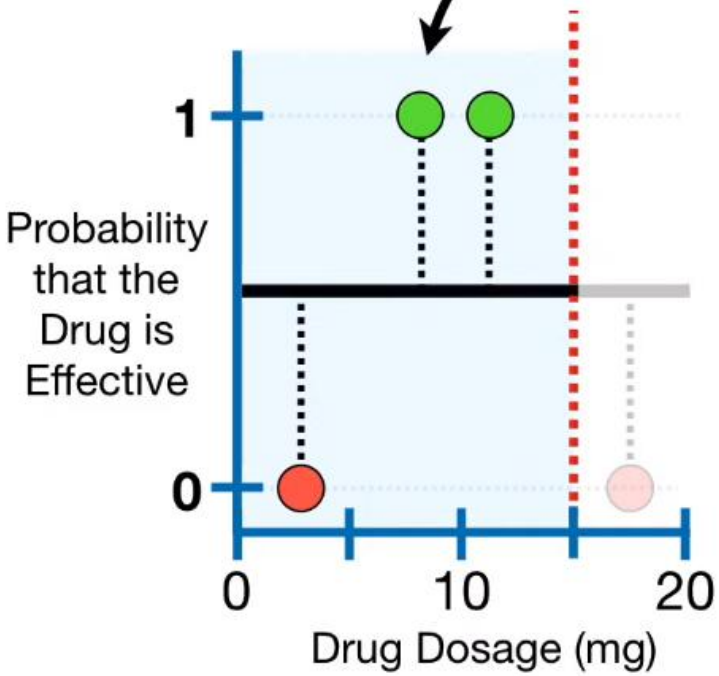
Predicted Drug Effectiveness

0.5

Similarity = 0



Thus, the three **Residuals** with **Dosages < 15** go to the leaf on the left…



Predicted Drug Effectiveness

0.5

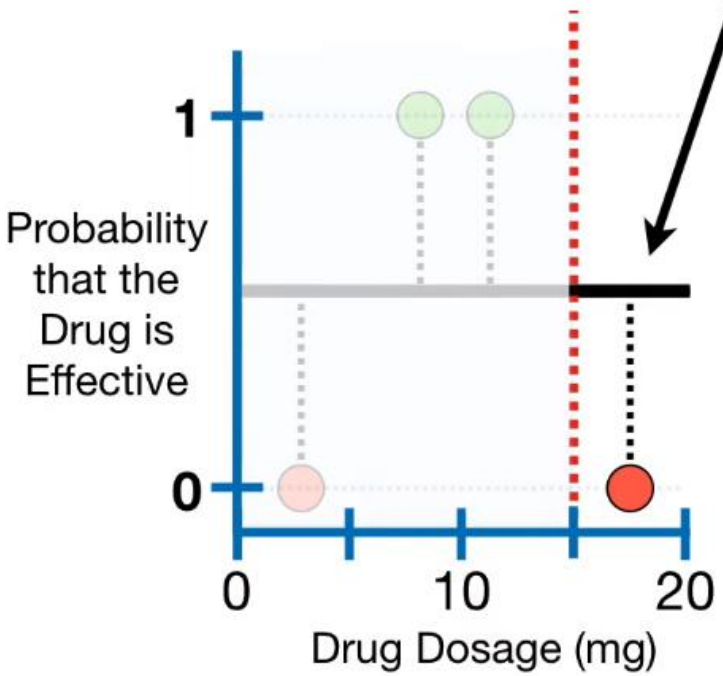
Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

-0.5

…and the one **Residual** with **Dosage > 15** goes to the leaf on the right.



Predicted Drug Effectiveness

0.5

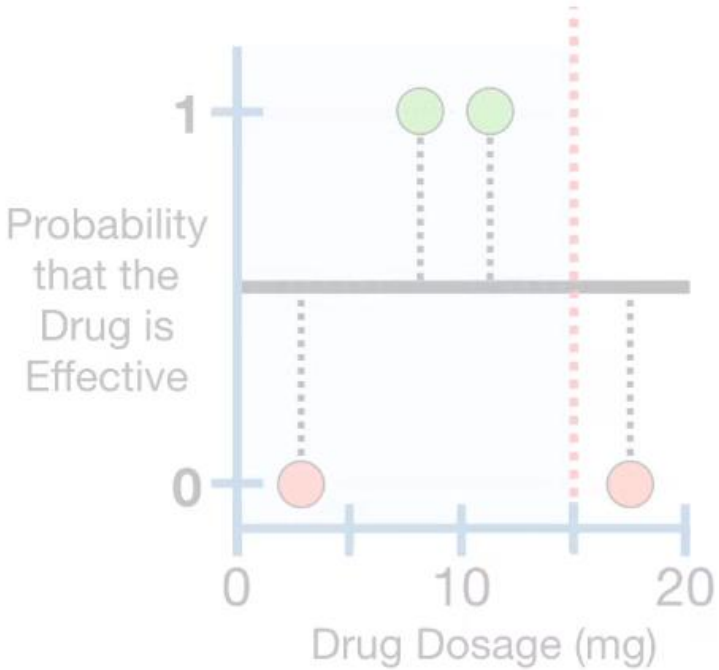
Similarity = 0

Dosage < 15

To calculate the **Similarity Score** for the three **Residuals** that ended up in the leaf on the left…

-0.5, 0.5, 0.5

-0.5



$$\frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

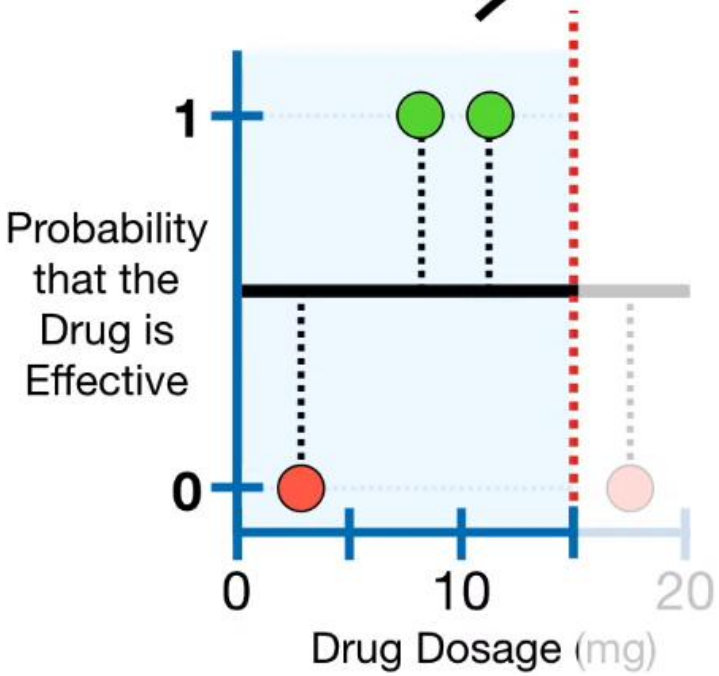
Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

-0.5

…we plug the three **Residuals** into the numerator…



$$\frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

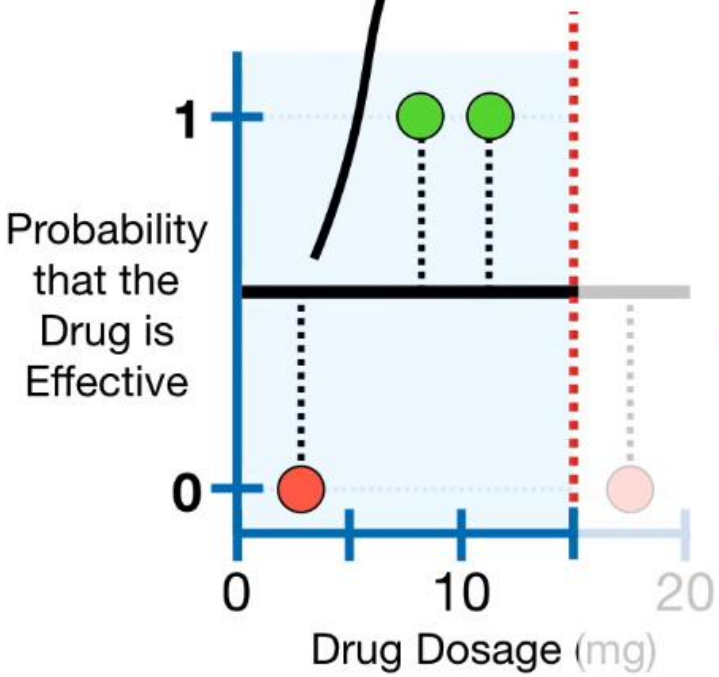
…and, since we are building the first tree, the **Previous Probability** refers to the prediction from the initial leaf…

Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

-0.5



$$\frac{(-0.5 + 0.5 + 0.5)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

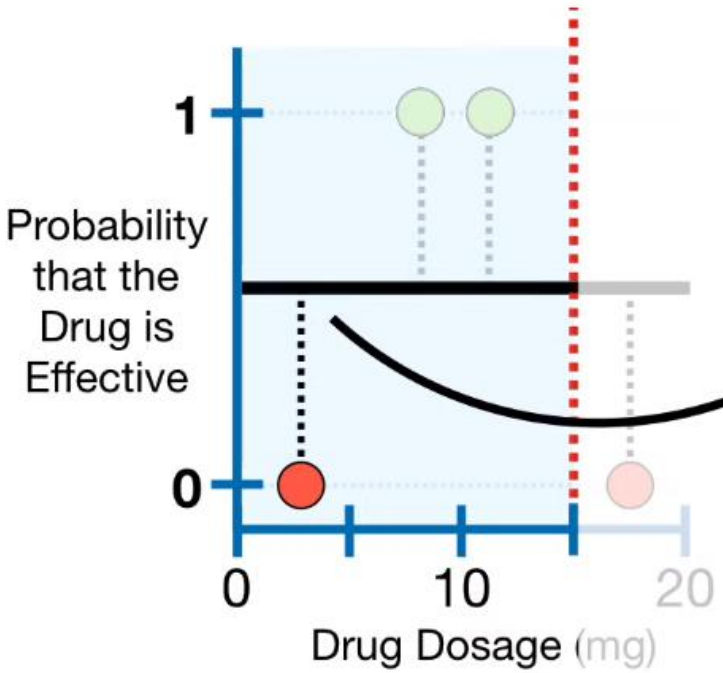
Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

-0.5

…so we plug in **0.5** for each **Residual** that ended up in the left leaf.



$$\frac{(-0.5 + 0.5 + 0.5)^2}{(0.5 \times (1 - \text{Previous Probability}_i) + \ldots + \lambda}$$

Predicted Drug Effectiveness

0.5

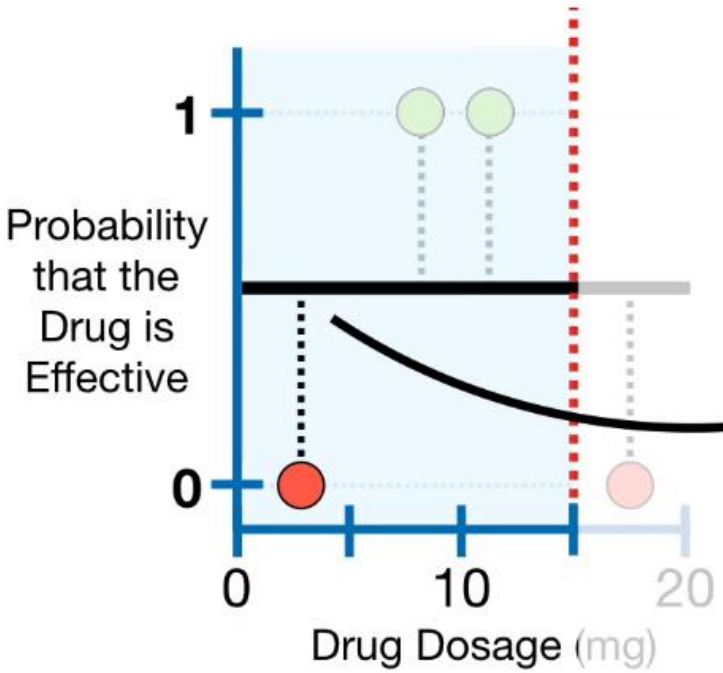
Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

-0.5

…so we plug in **0.5** for each **Residual** that ended up in the left leaf.



$$\frac{(-0.5 + 0.5 + 0.5)^2}{(0.5 \times (1-0.5)) + \ldots + \lambda}$$

Predicted Drug Effectiveness

0.5

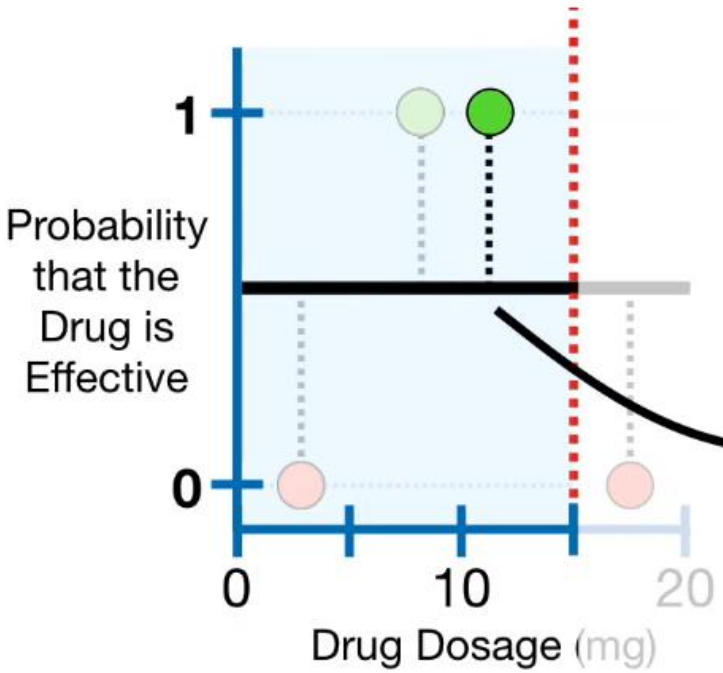
Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

-0.5

…so we plug in **0.5** for each **Residual** that ended up in the left leaf.



$$\frac{(-0.5 + 0.5 + 0.5)^2}{(0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + \lambda}$$

Predicted Drug Effectiveness

0.5

Similarity = 0

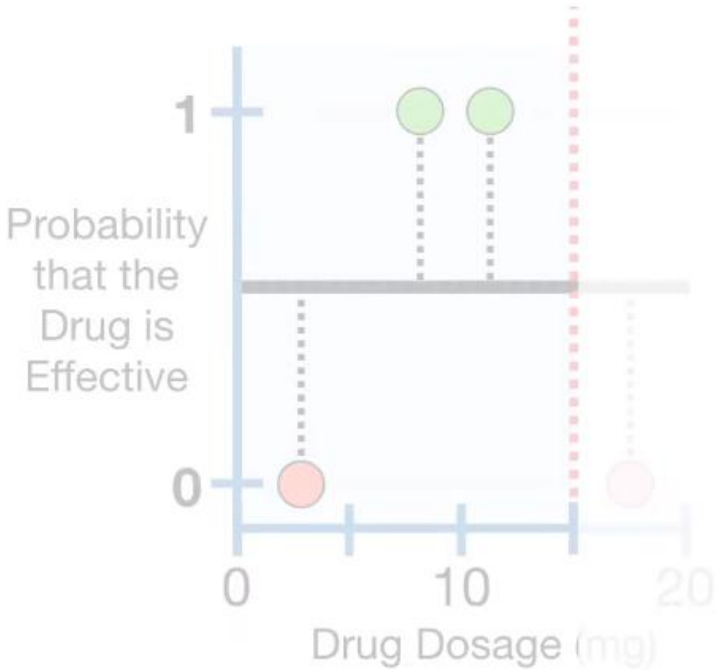
Dosage < 15

-0.5, 0.5

0.5

-0.5

Now, just to keep things simple, we'll let $\lambda = 0$. However, you know from **Part 1** that **λ (lambda)** reduces the **Similarity Score**, which ultimately makes leaves easier to prune.



$$\frac{(-0.5 + 0.5 + 0.5)^2}{(0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + \lambda}$$

Predicted Drug Effectiveness

0.5

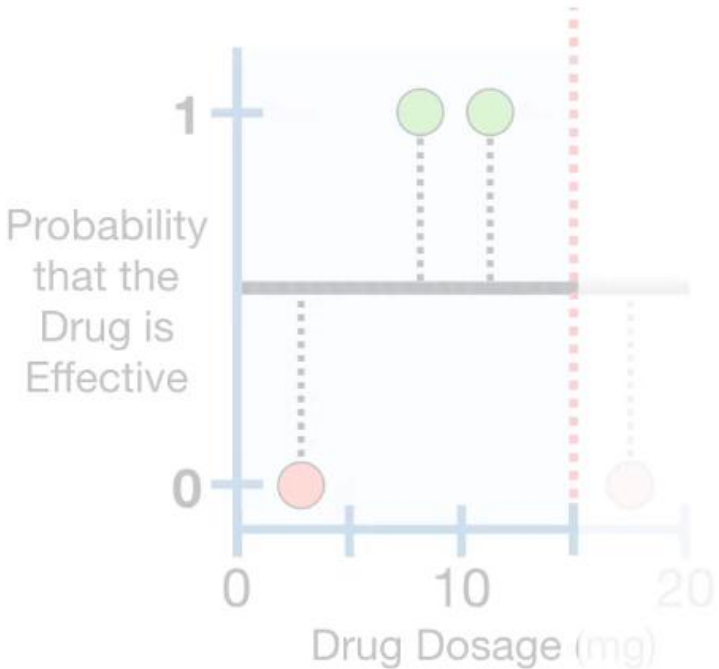
Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

-0.5

Now, just to keep things simple, we'll let $\lambda = 0$. However, you know from **Part 1** that **λ (lambda)** reduces the **Similarity Score**, which ultimately makes leaves easier to prune.



$$\frac{(-0.5 + 0.5 + 0.5)^2}{(0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + 0}$$

Predicted Drug Effectiveness

0.5

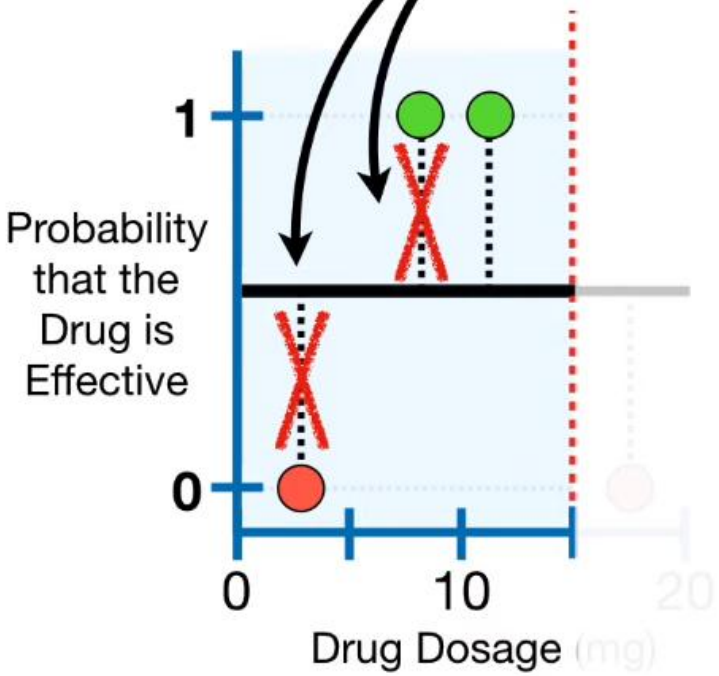
Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

-0.5

Now notice that these two **Residuals** in the numerator cancel each other out…



$$\frac{(-0.5 + 0.5 + 0.5)^2}{(0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + 0}$$

Predicted Drug Effectiveness

0.5

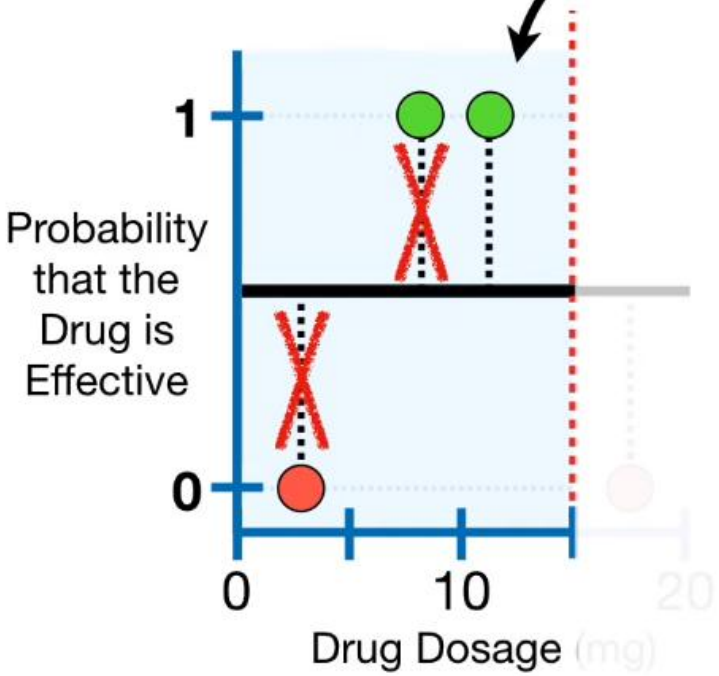
Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

-0.5

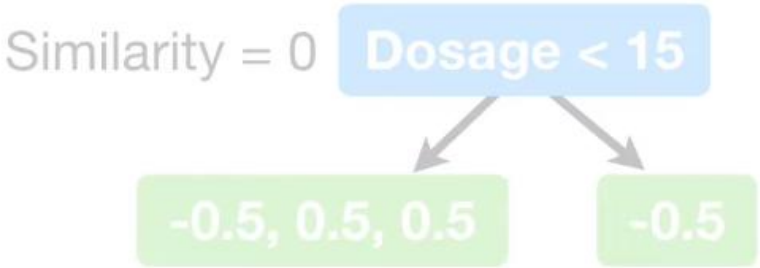
…leaving us with just one **Residual** in the numerator…



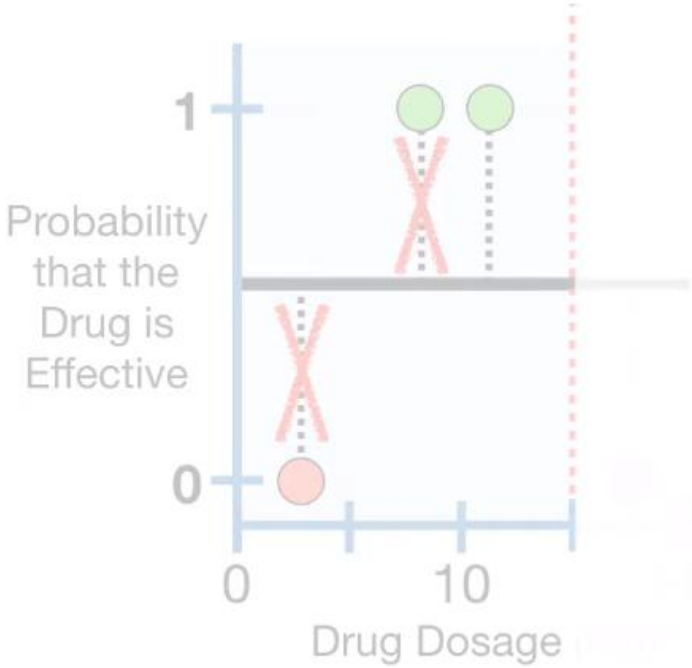
$$\frac{(0.5)^2}{(0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + 0}$$

Predicted Drug Effectiveness

0.5



…and when we do the math…



$$\frac{(0.5)^2}{(0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + \mathbf{0}}$$

Predicted Drug Effectiveness

0.5

Similarity = 0

Dosage < 15

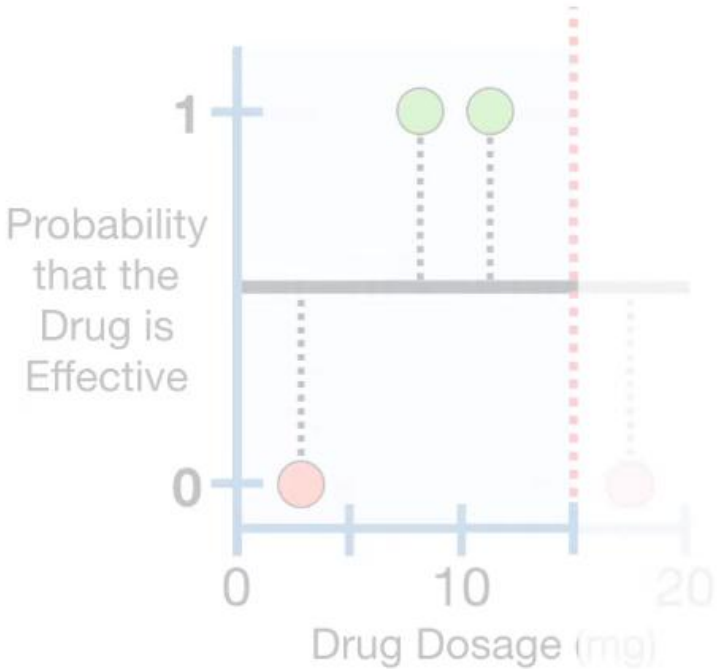
-0.5, 0.5, 0.5

-0.5

…we get **0.33**.

$$\frac{(0.5)^2}{(0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + 0}$$

= 0.33



Predicted Drug Effectiveness

0.5

Similarity = 0

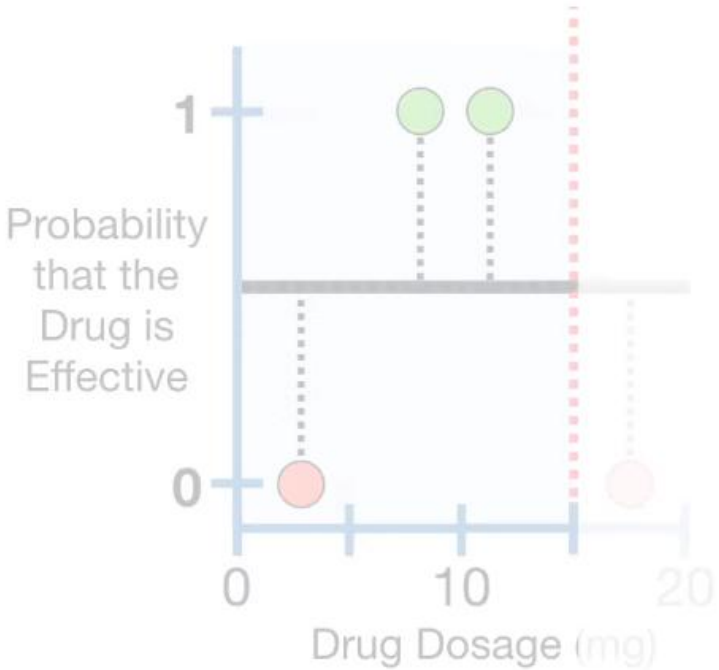
Dosage < 15

-0.5, 0.5, 0.5

-0.5

So let's put **Similarity = 0.33** under this leaf so we can keep track of it.

Similarity = 0.33



$$\frac{(0.5)^2}{(0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + (0.5 \times (1-0.5)) + 0}$$

$$= 0.33$$

Predicted Drug Effectiveness

0.5

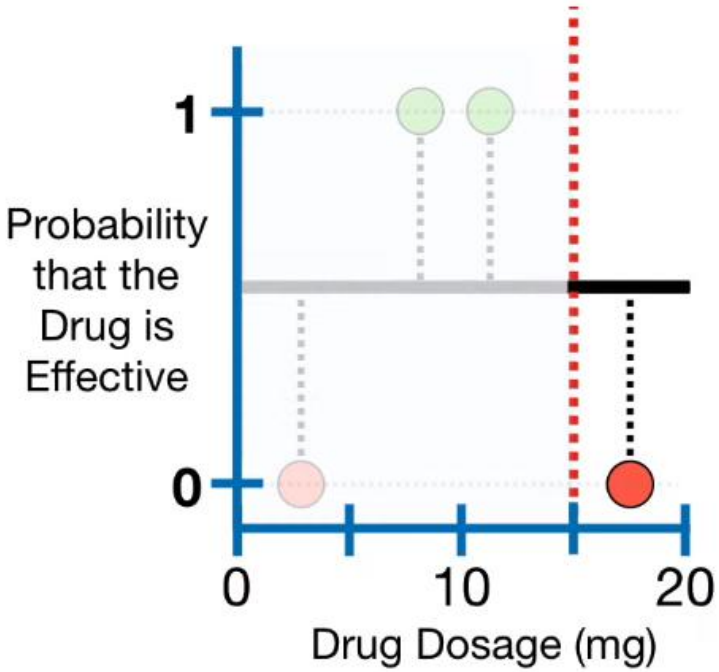
Similarity = 0 **Dosage < 15**

-0.5, 0.5, 0.5

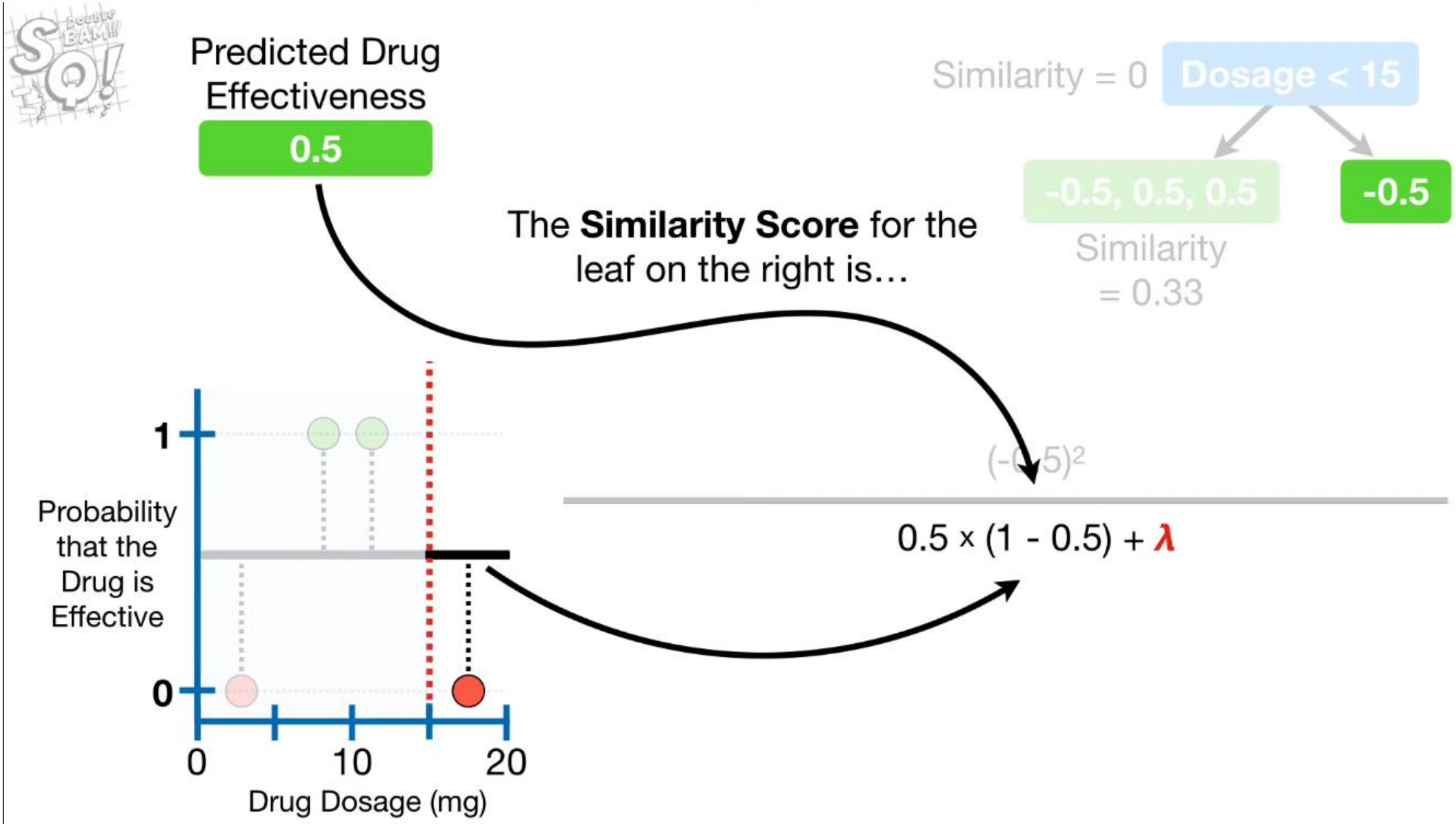
Similarity = 0.33

-0.5

The **Similarity Score** for the leaf on the right is…



$$\frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$


Predicted Drug Effectiveness
0.5
Similarity = 0
Dosage < 15
-0.5, 0.5, 0.5
-0.5
Similarity = 0.33
The Similarity Score for the leaf on the right is…
(-0.5)²
0.5 x (1 - 0.5) + λ
1
0
Probability that the Drug is Effective
0
10
20
Drug Dosage (mg)

Predicted Drug Effectiveness

0.5

Similarity = 0

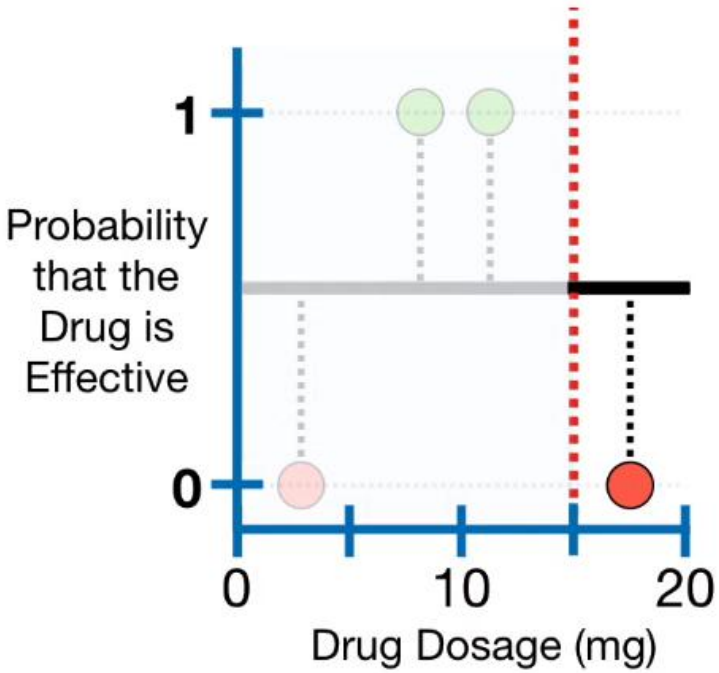
Dosage < 15

-0.5, 0.5, 0.5

-0.5

Similarity = 0.33

The **Similarity Score** for the leaf on the right is…



$$\frac{(-0.5)^2}{0.5 \times (1 - 0.5) + \lambda}$$

Predicted Drug Effectiveness

0.5

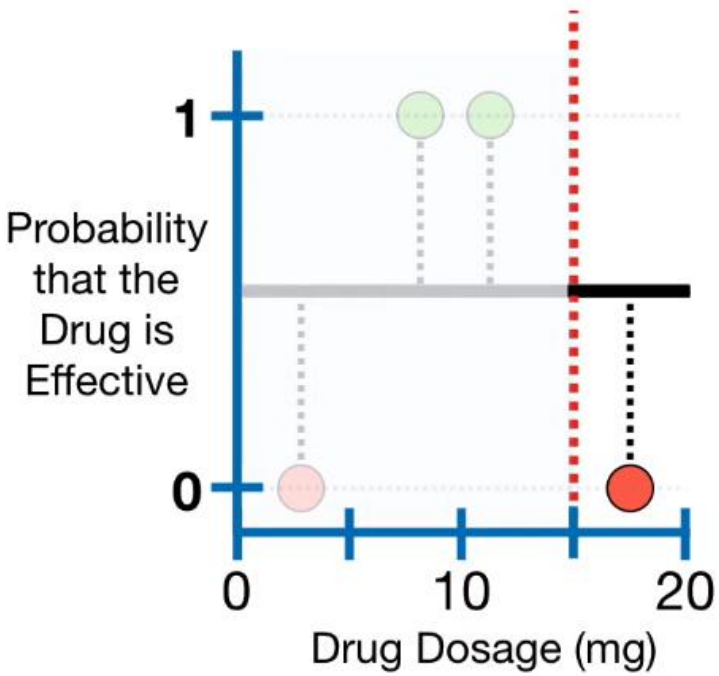
Similarity = 0 | Dosage < 15

-0.5, 0.5, 0.5 | -0.5

Similarity = 0.33

…1, when $\lambda = 0$.

$$\frac{(-0.5)^2}{0.5 \times (1 - 0.5) + \lambda} = 1, \text{ when } \lambda = 0.$$



Predicted Drug Effectiveness

0.5

Similarity = 0 | Dosage < 15

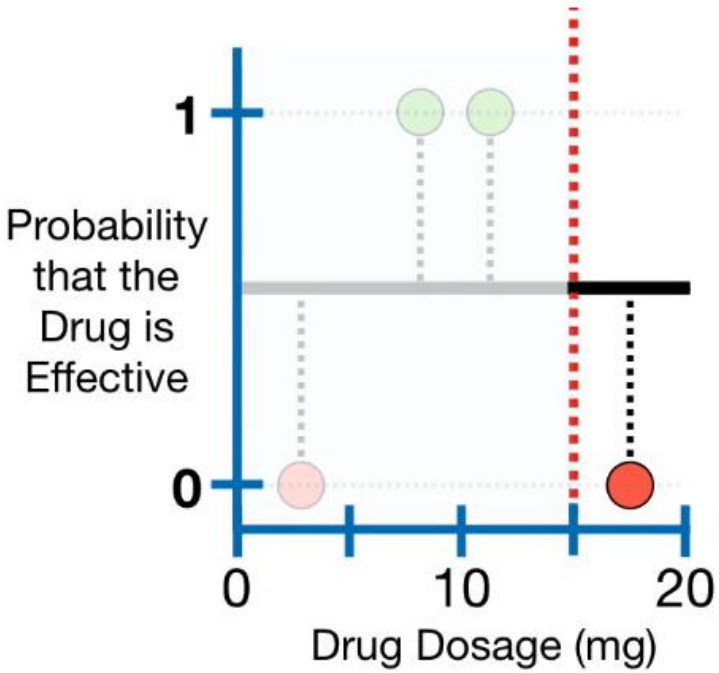
-0.5, 0.5, 0.5

Similarity = 0.33

-0.5

Similarity = 1

So let's put **Similarity = 1** under this leaf to keep track of it.



$$\frac{(-0.5)^2}{0.5 \times (1 - 0.5) + \lambda} = 1, \text{ when } \lambda = 0.$$

Predicted Drug Effectiveness

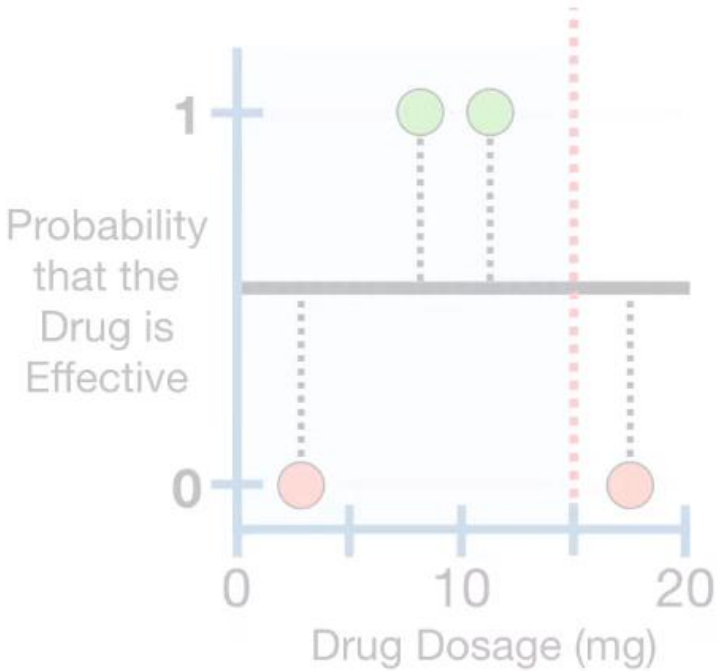
0.5

Similarity = 0 | Dosage < 15

-0.5, 0.5, 0.5 — Similarity = 0.33

-0.5 — Similarity = 1

Now we can calculate the **Gain**, just like we did when we used **XGBoost** for **Regression**.



$$\text{Gain} = \text{Left}_{\text{Similarity}} + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

Predicted Drug Effectiveness

0.5

Similarity = 0

Dosage < 15

-0.5, 0.5, 0.5

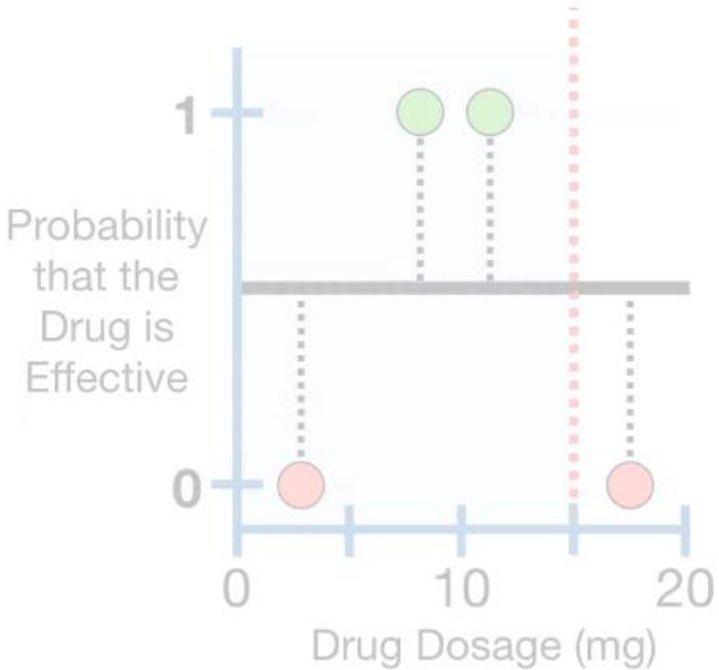
Similarity = 0.33

-0.5

Similarity = 1

…and get **1.33**.

Gain = 0.33 + 1 - 0 = 1.33



Predicted Drug Effectiveness

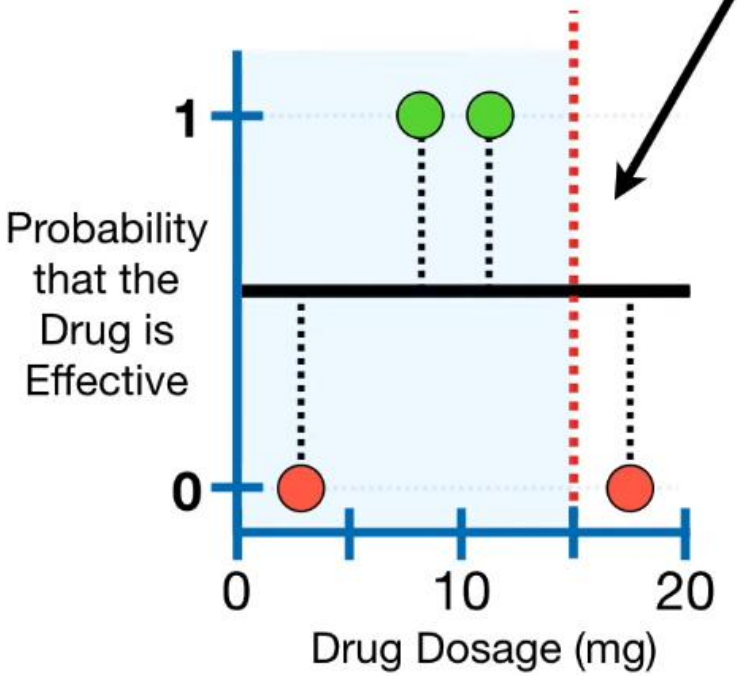
0.5

Similarity = 0 **Dosage < 15**

-0.5, 0.5, 0.5 Similarity = 0.33

-0.5 Similarity = 1

So when we split the **Observations** based on the threshold **Dosage < 15**, **Gain = 1.33**.



Gain = 0.33 + 1 - 0 = 1.33

Predicted Drug Effectiveness

0.5

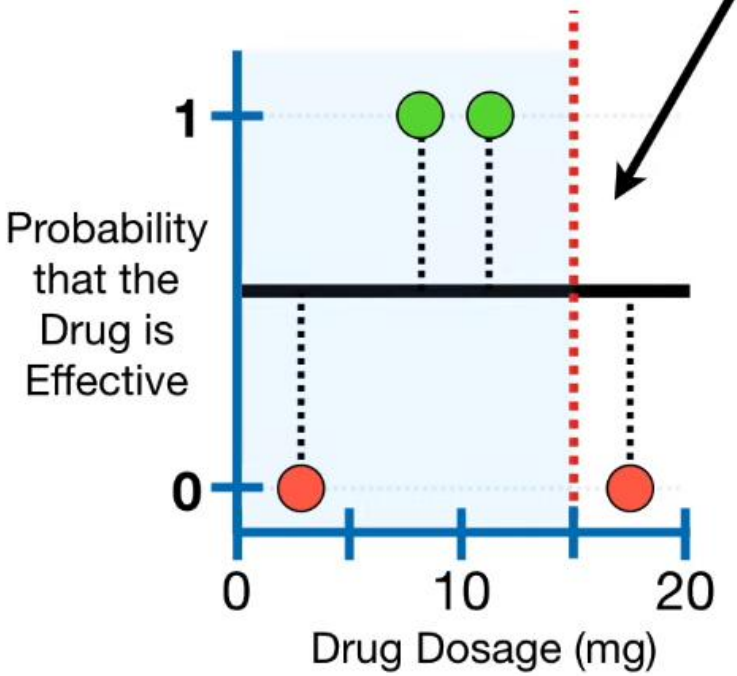
Similarity = 0 **Dosage < 15**

-0.5, 0.5, 0.5 Similarity = 0.33

-0.5 Similarity = 1

So when we split the **Observations** based on the threshold **Dosage < 15**, **Gain = 1.33**.

Gain = 0.33 + 1 - 0 = 1.33



Predicted Drug Effectiveness

0.5

Similarity = 0

Dosage < 15

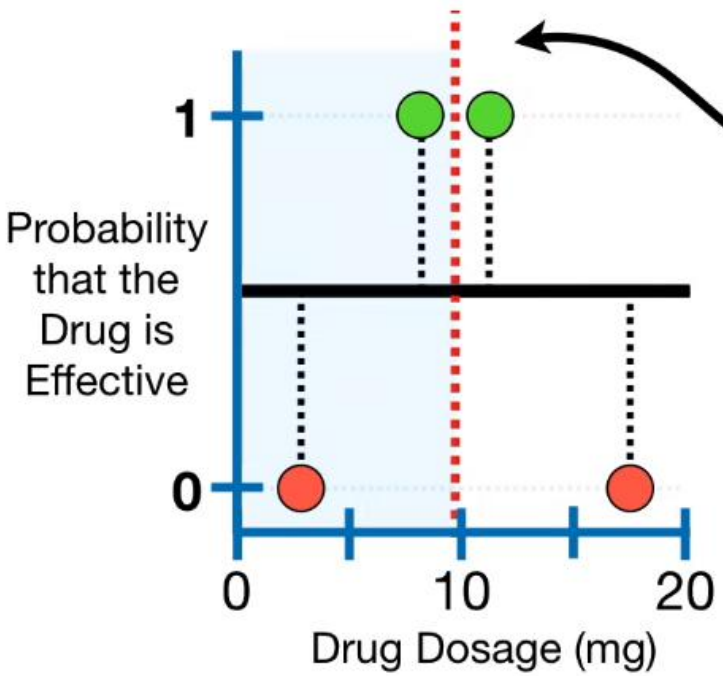
-0.5, 0.5, 0.5

Similarity = 0.33

-0.5

Similarity = 1

Since I'm such a nice guy, I'm going to tell you that no other threshold gives us a larger **Gain** value…



Predicted Drug Effectiveness

0.5

Similarity = 0

Dosage < 15

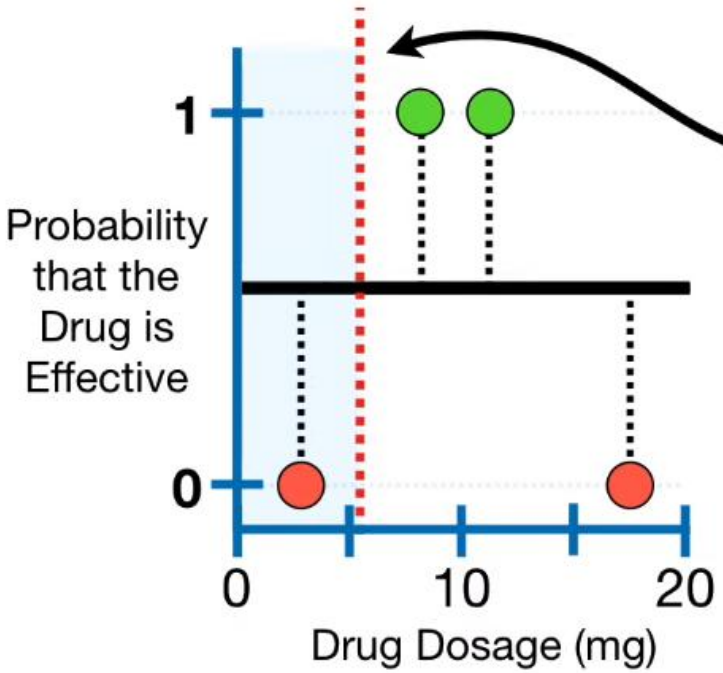
-0.5, 0.5, 0.5

Similarity = 0.33

-0.5

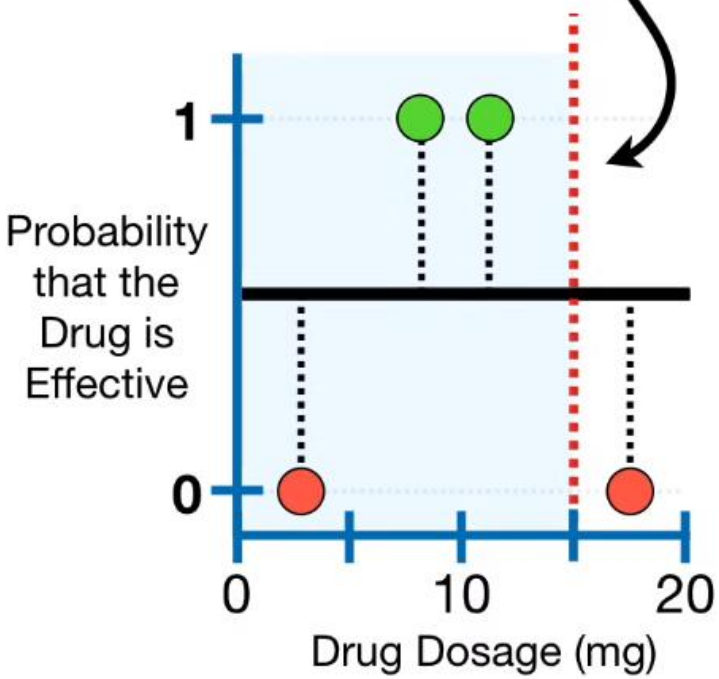
Similarity = 1

Since I'm such a nice guy, I'm going to tell you that no other threshold gives us a larger **Gain** value…

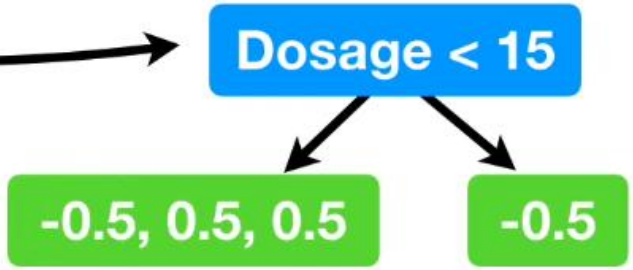


Predicted Drug Effectiveness

0.5



…and that means **Dosage < 15** will be the first branch in our tree.



Predicted Drug Effectiveness

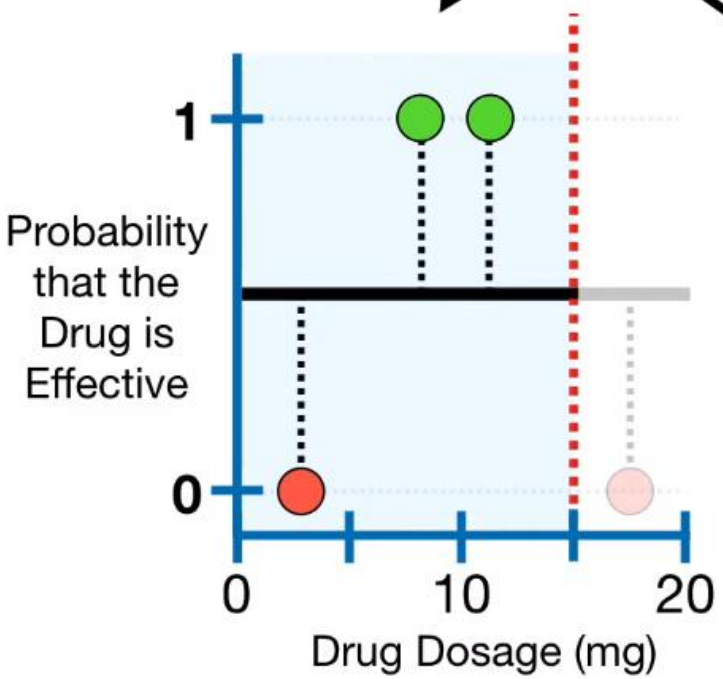
0.5

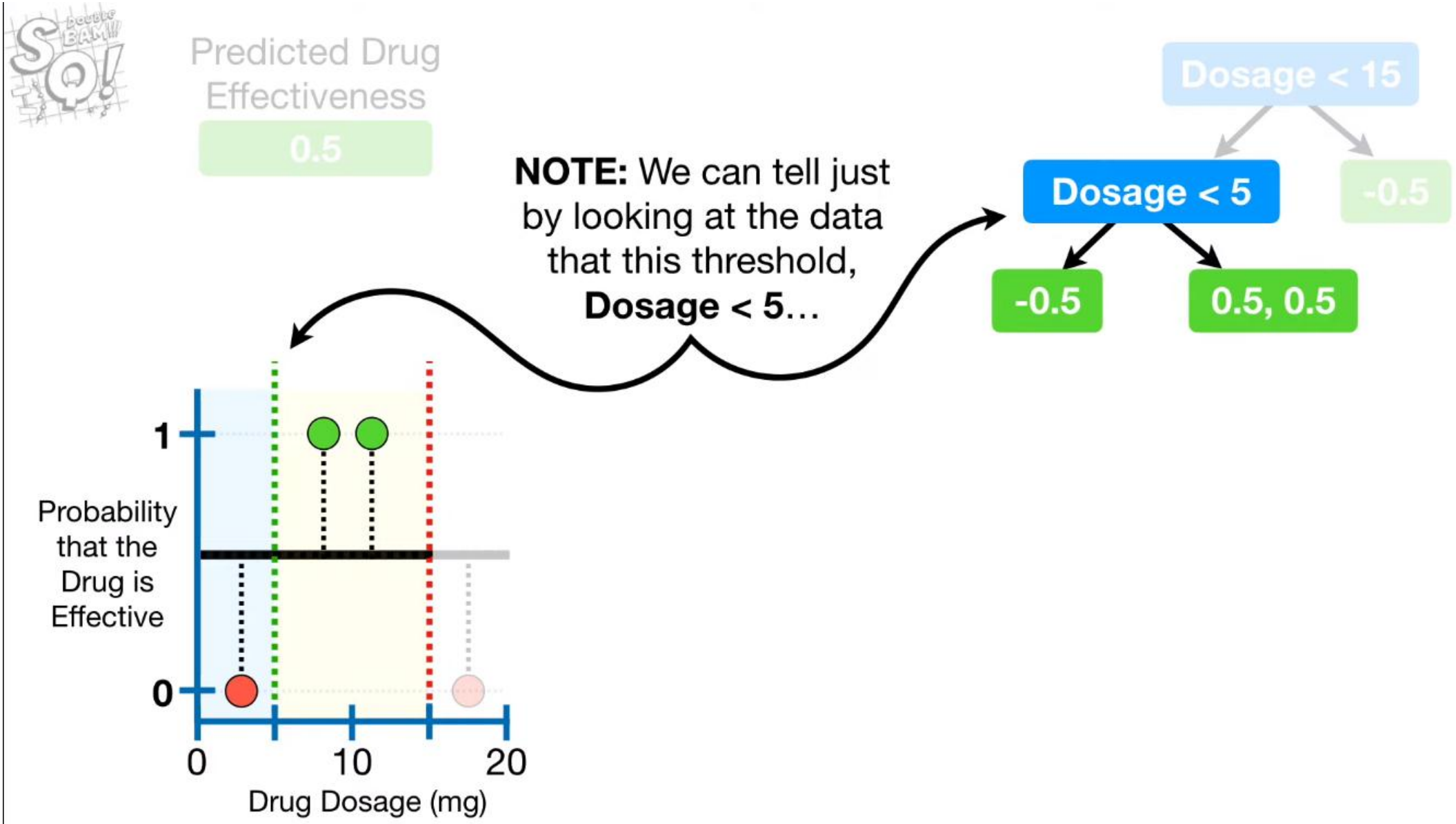
Dosage < 15

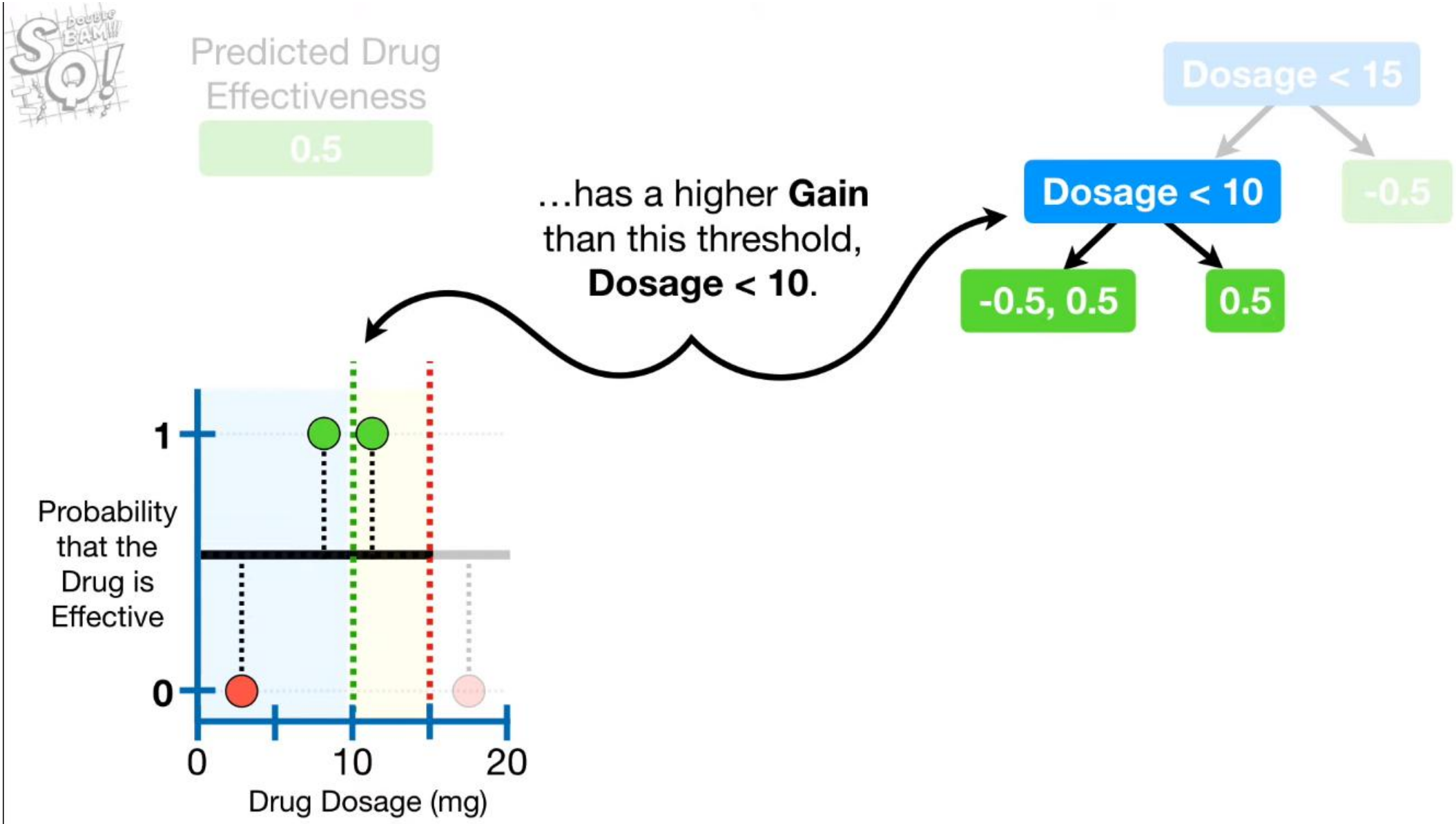
-0.5, 0.5, 0.5

-0.5

Now we will focus on splitting these **Residuals** into two leaves.



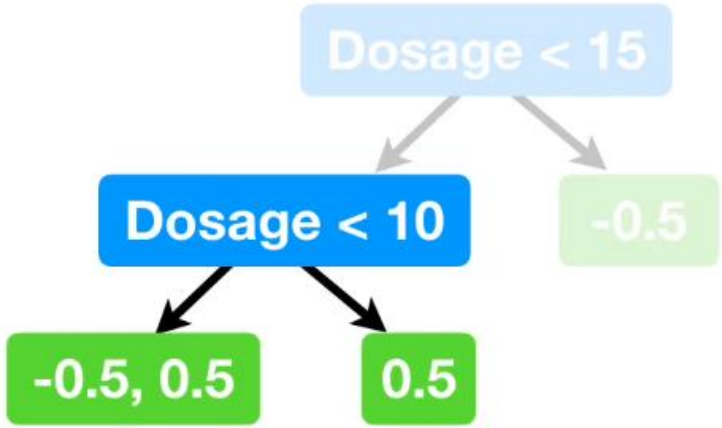

Predicted Drug Effectiveness
0.5
NOTE: We can tell just by looking at the data that this threshold, Dosage < 5…
Dosage < 15
Dosage < 5
-0.5
-0.5
0.5, 0.5
1
0
Probability that the Drug is Effective
0
10
20
Drug Dosage (mg)

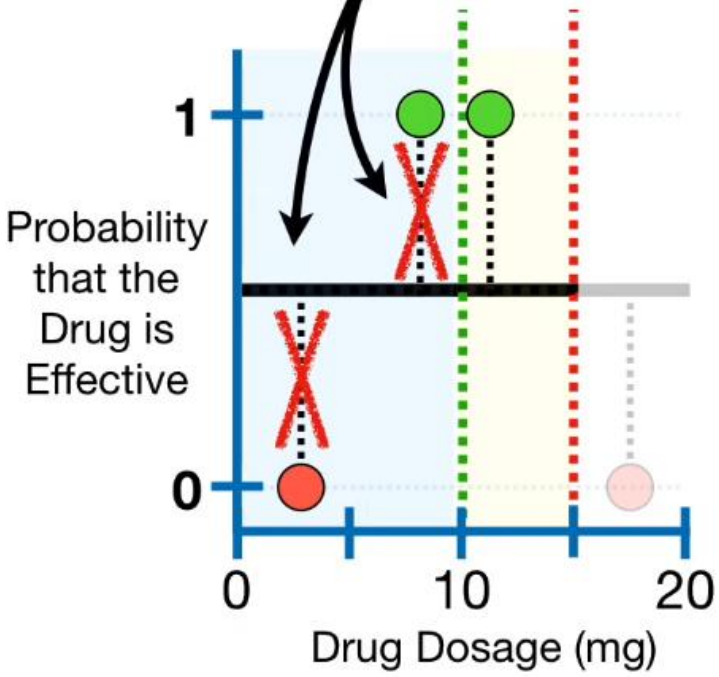

Predicted Drug Effectiveness
0.5
Dosage < 15
Dosage < 10
-0.5
-0.5, 0.5
0.5
…has a higher Gain than this threshold, Dosage < 10.
1
0
Probability that the Drug is Effective
0
10
20
Drug Dosage (mg)

Predicted Drug Effectiveness

0.5

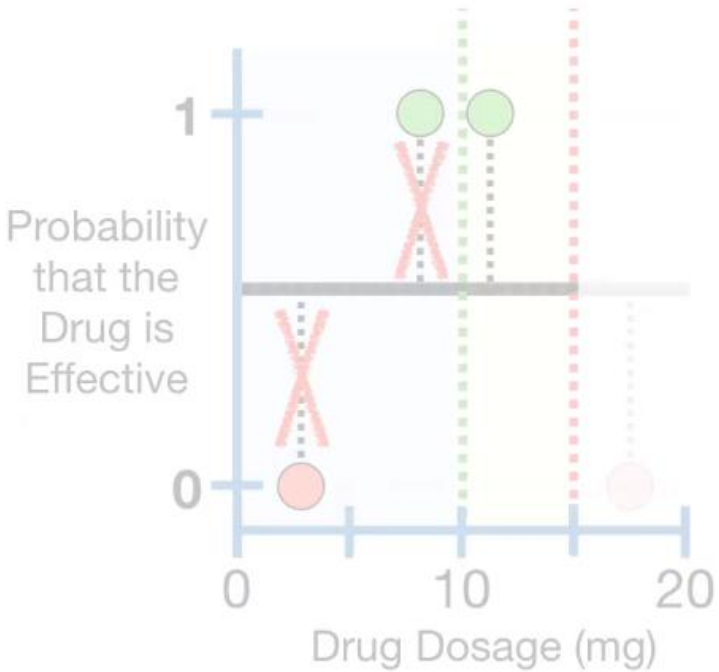
This is because when the threshold is **Dosage < 10**, these two **Residuals** will cancel each other out.



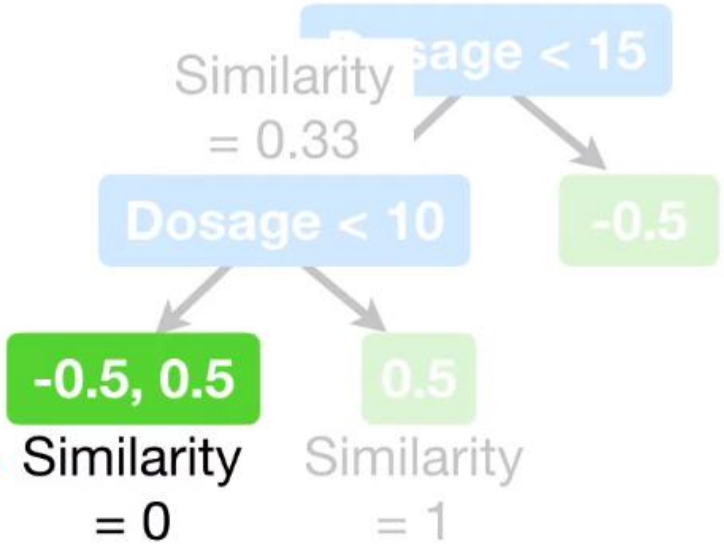


Predicted Drug Effectiveness

0.5



So when we calculate the **Gain**, we get…

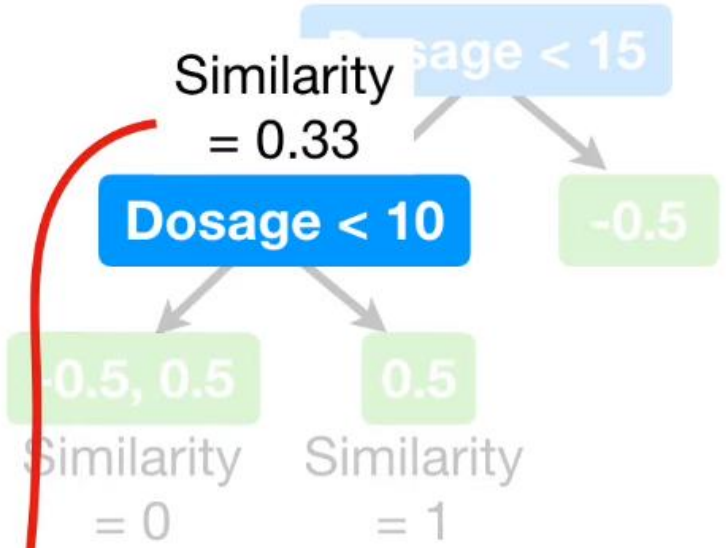


$$\text{Gain} = \text{Left}_{\text{Similarity}} + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

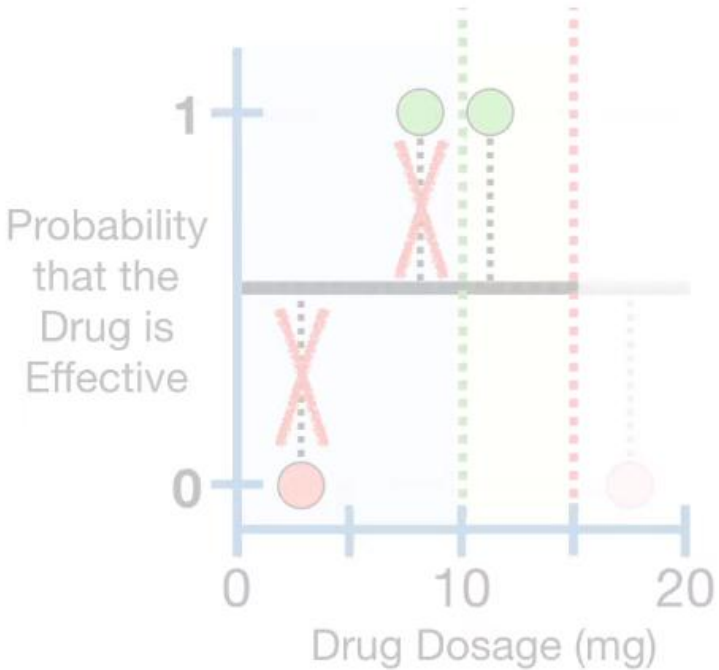
Predicted Drug Effectiveness

0.5

So when we calculate the **Gain**, we get…

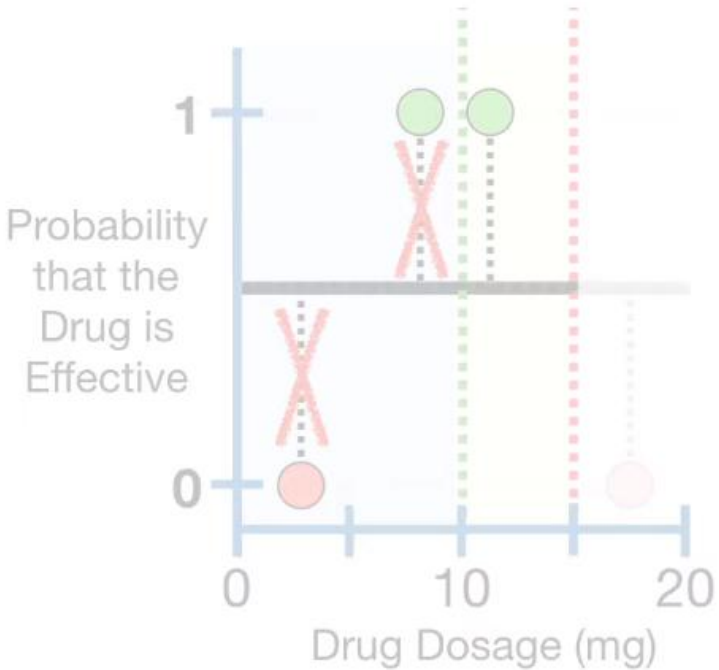


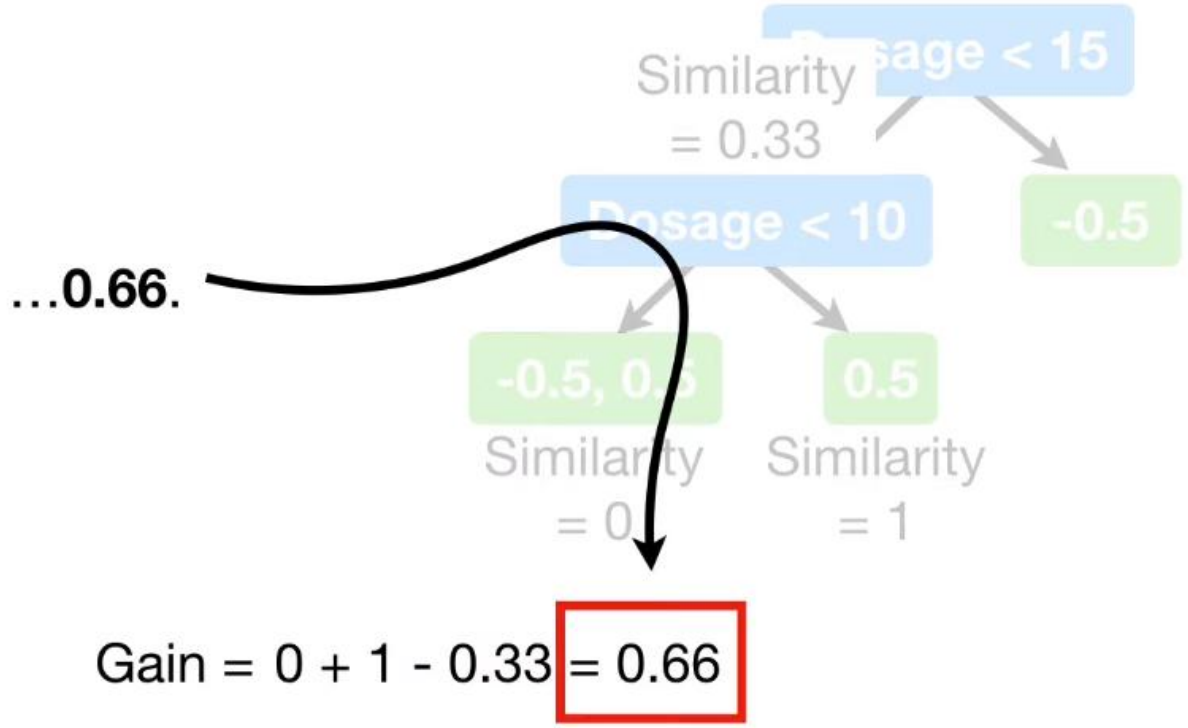
Gain = 0 + 1 - 0.33



Predicted Drug Effectiveness

0.5



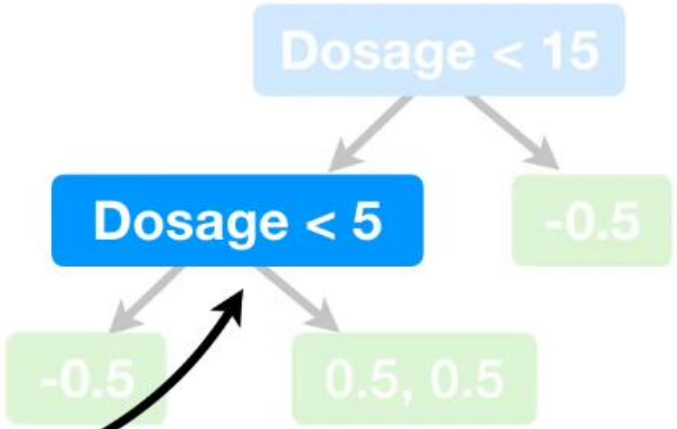


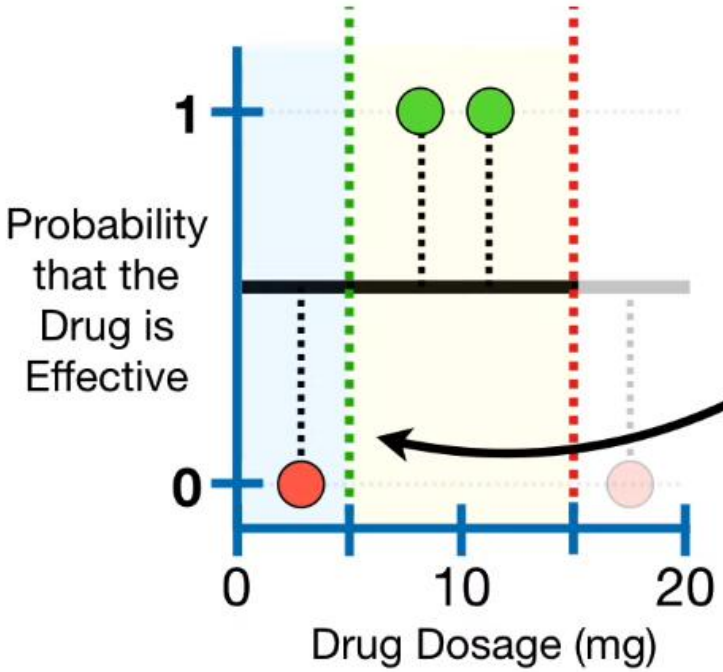
Gain = 0 + 1 - 0.33 = 0.66

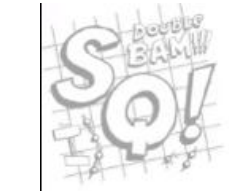
Predicted Drug Effectiveness

0.5

Now let's compare that to the **Gain** we get when the threshold is **Dosage < 5**.

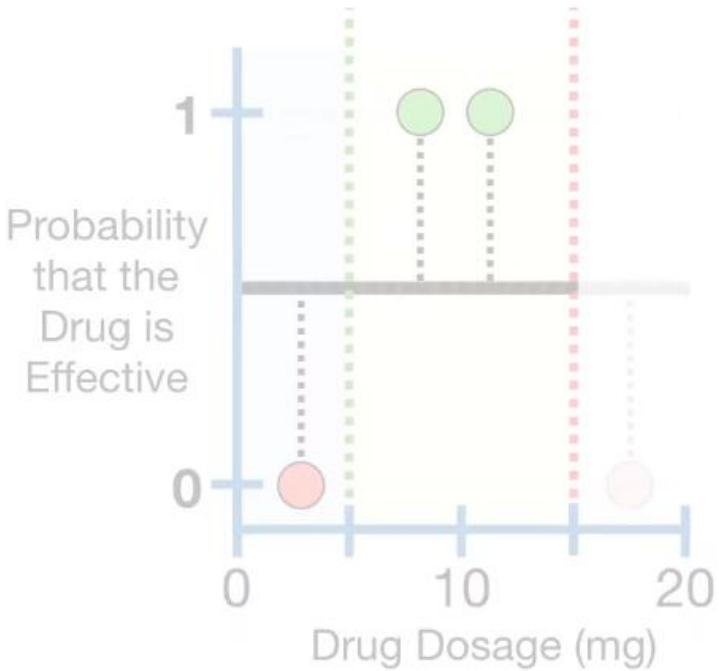


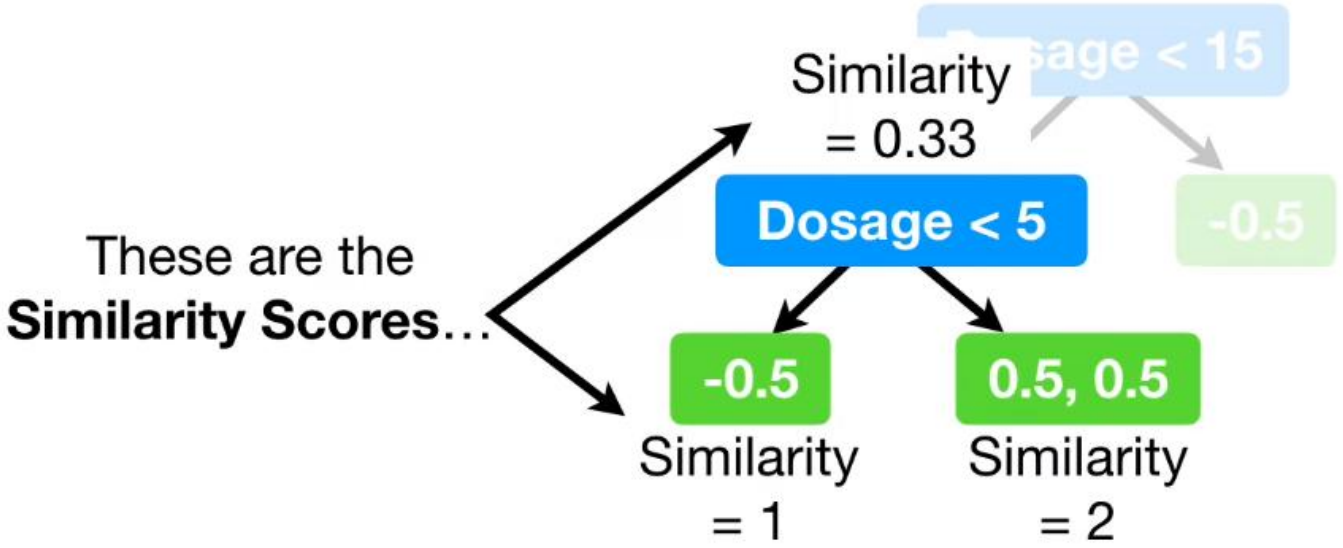




Predicted Drug Effectiveness

0.5



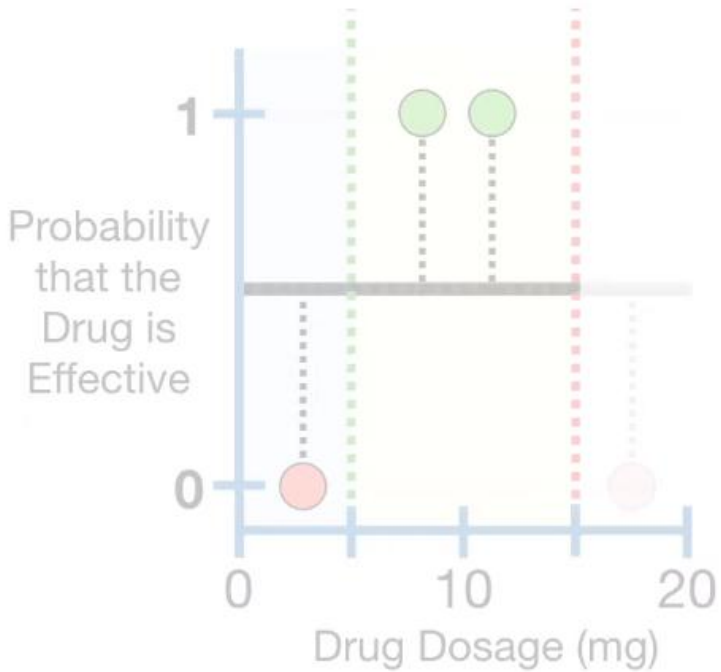


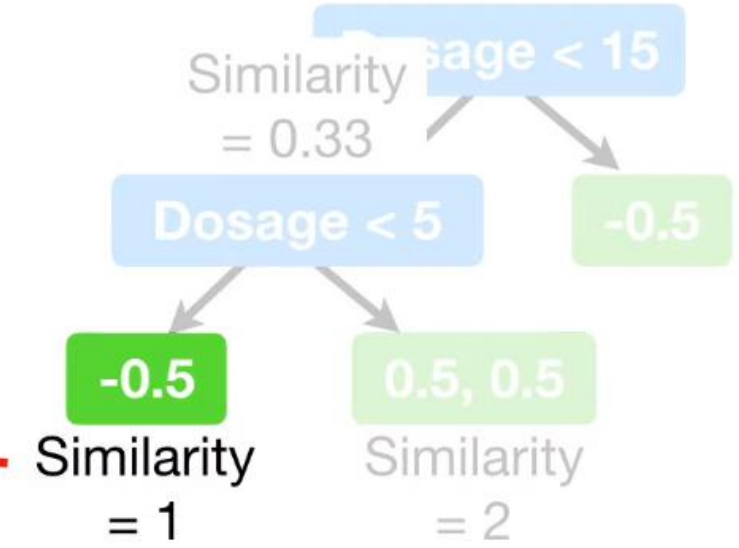
$$\text{Similarity} = \frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

…and when we plug them into the equation for the **Gain**…



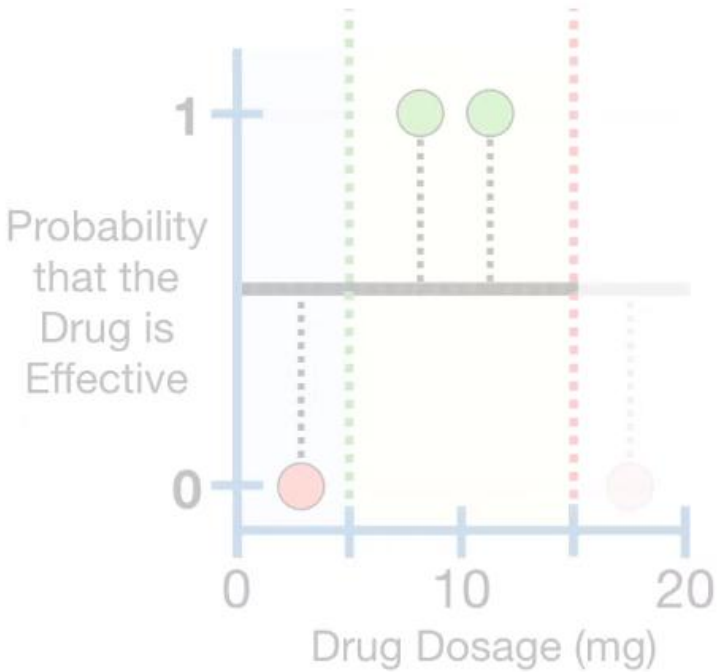


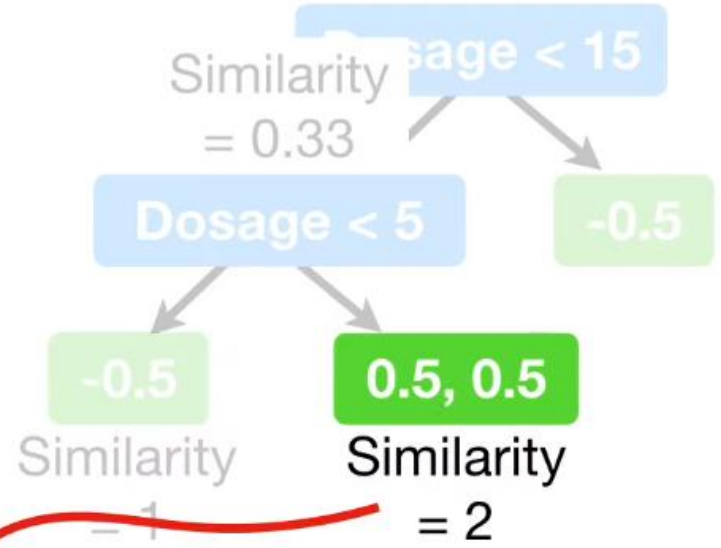
$$\text{Gain} = \text{Left}_{\text{Similarity}} + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

Predicted Drug Effectiveness

0.5

…and when we plug them into the equation for the **Gain**…



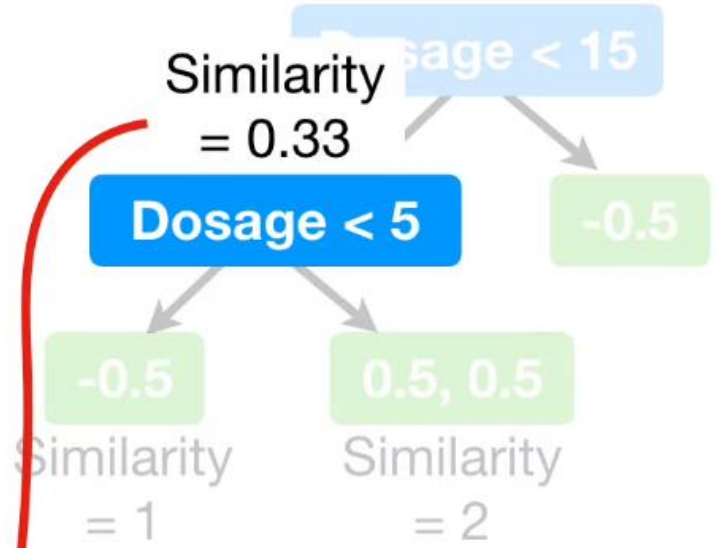


$$\text{Gain} = 1 + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

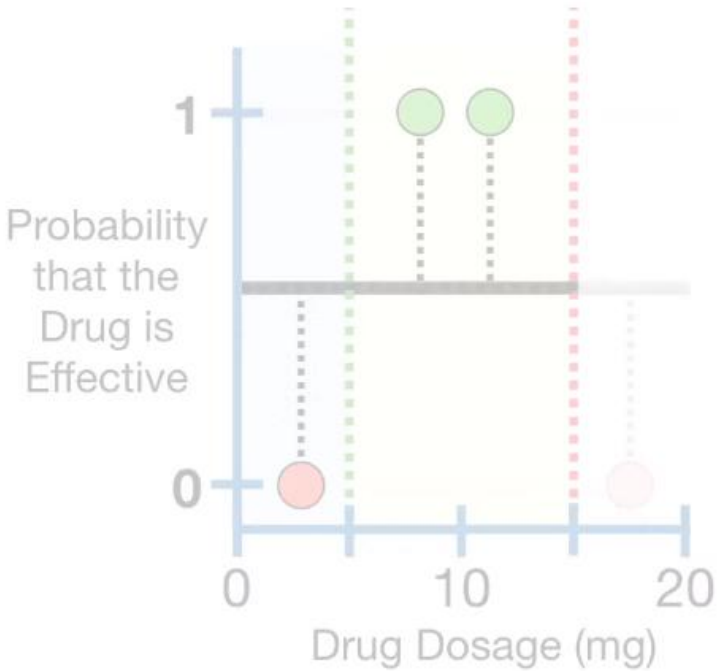
Predicted Drug Effectiveness

0.5

…and when we plug them into the equation for the **Gain**…

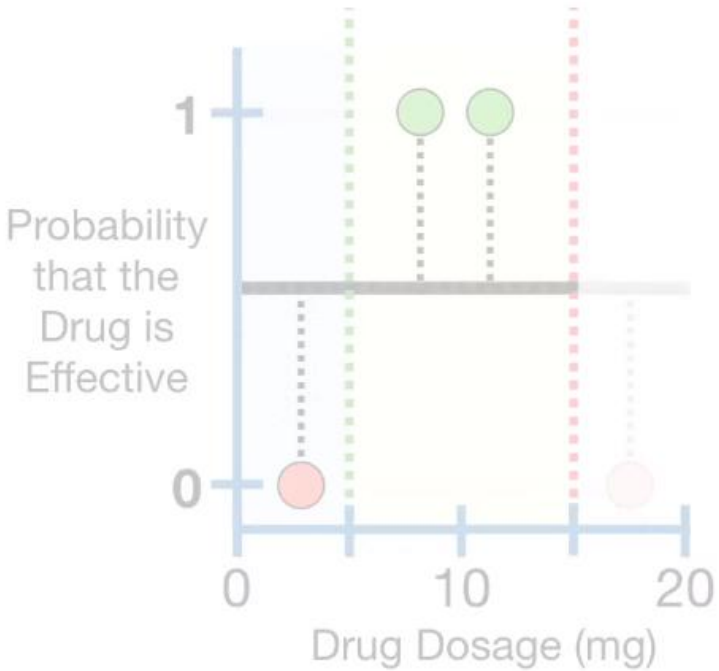


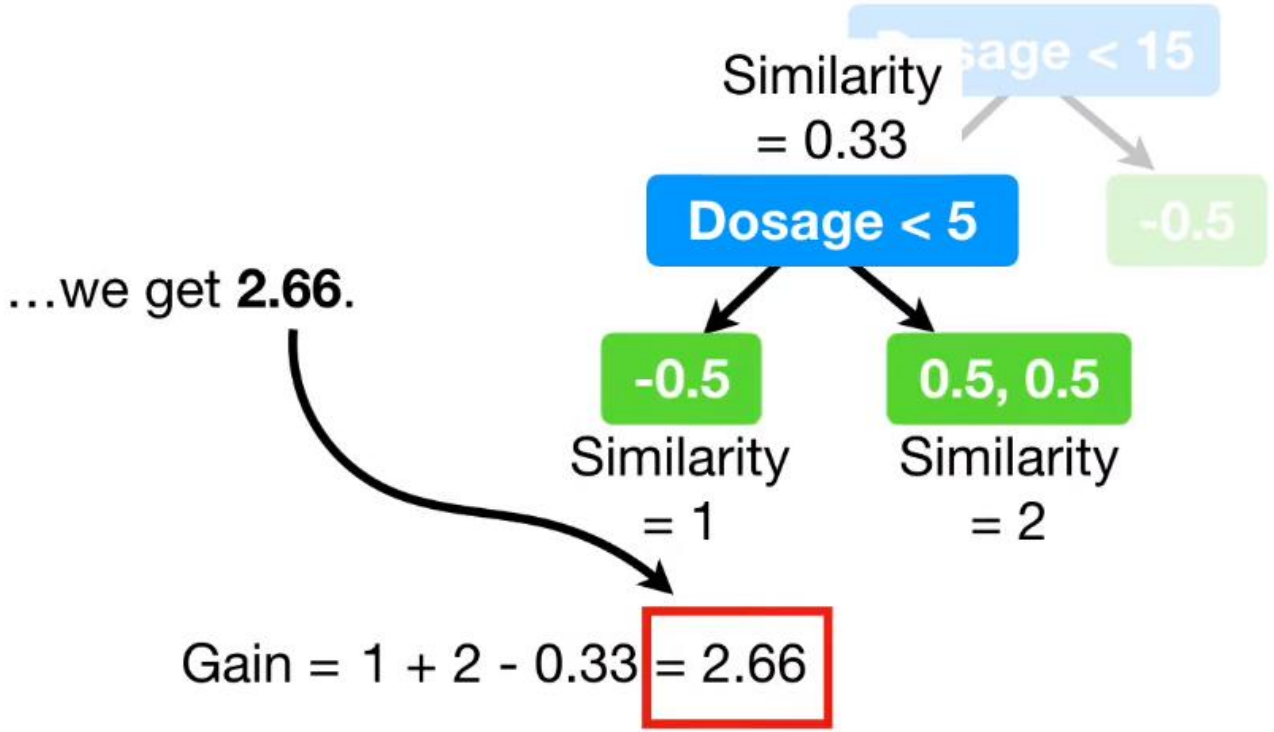
Gain = 1 + 2 - 0.33



Predicted Drug Effectiveness

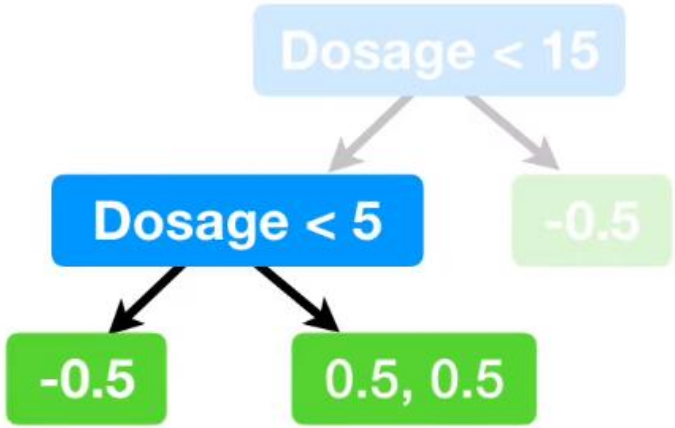
0.5



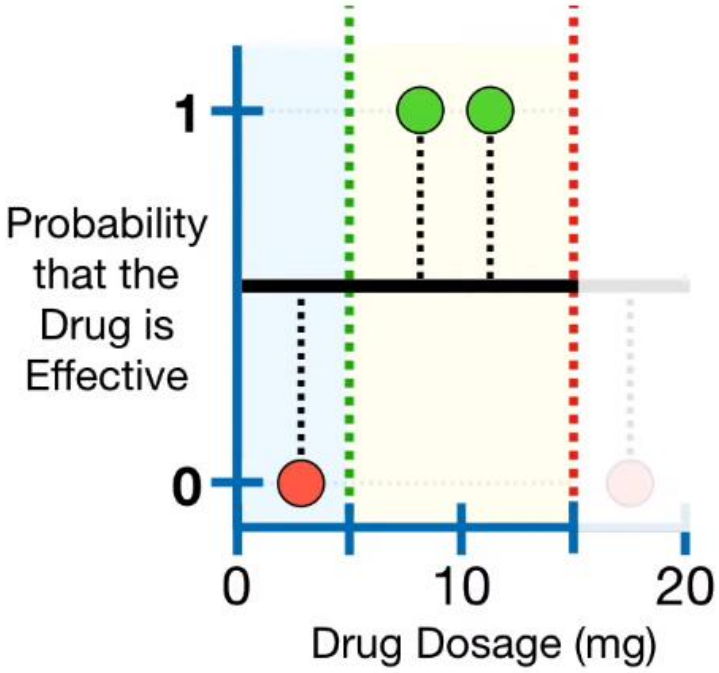


Predicted Drug Effectiveness

0.5

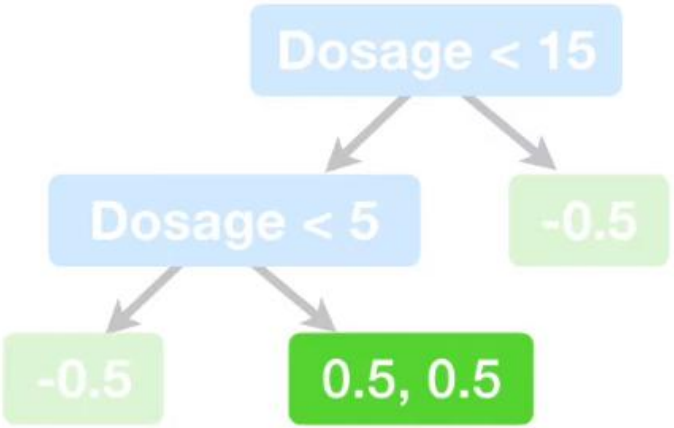


And since **2.66 > 0.66**, we will use **Dosage < 5** as the threshold for this branch.

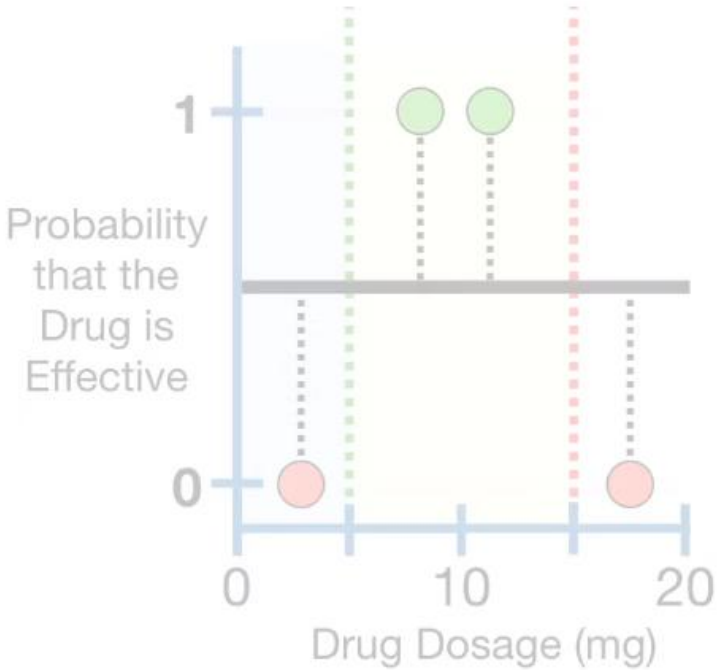


Predicted Drug Effectiveness

0.5



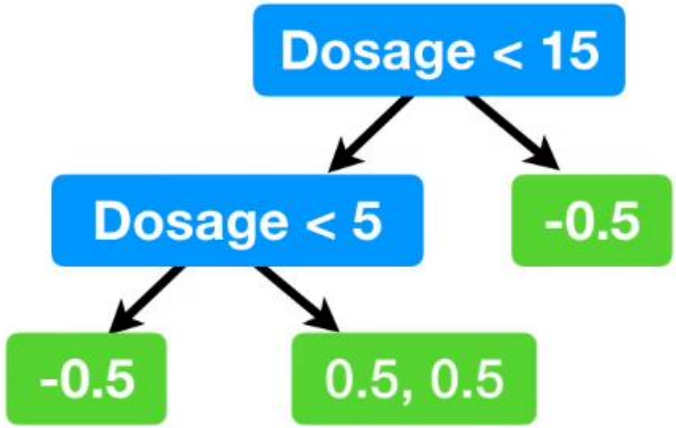
Now, since I'm limiting trees to **2** levels, we will not split this leaf any further, and we are done building this tree.

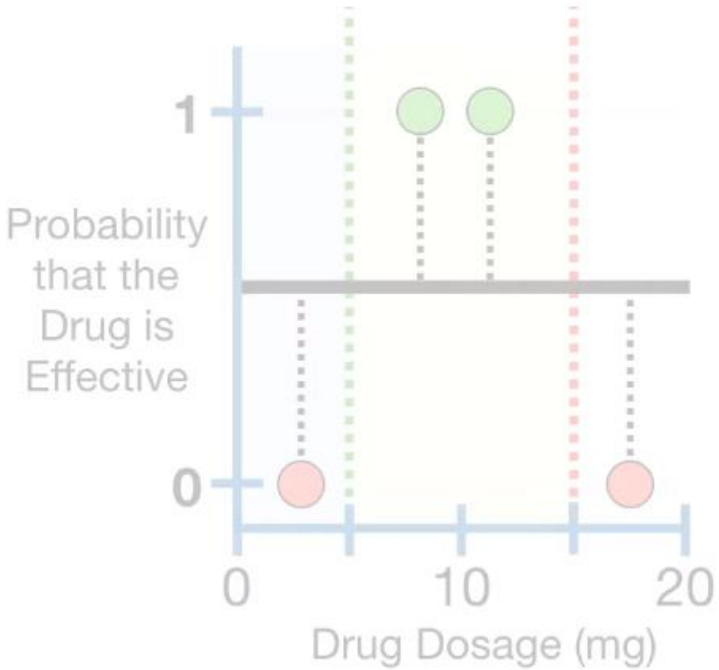


Predicted Drug Effectiveness

0.5

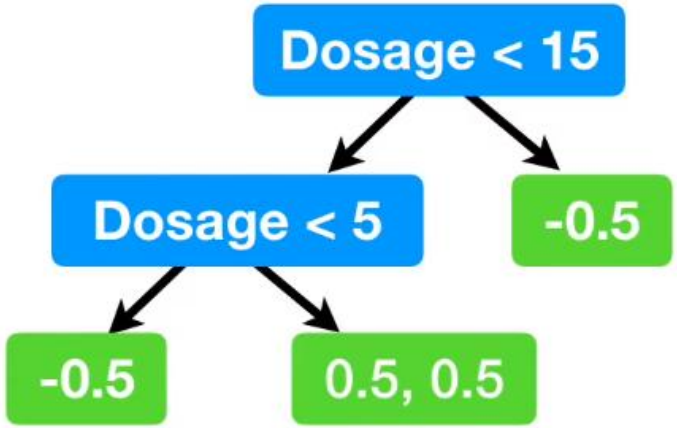
BAM!!!



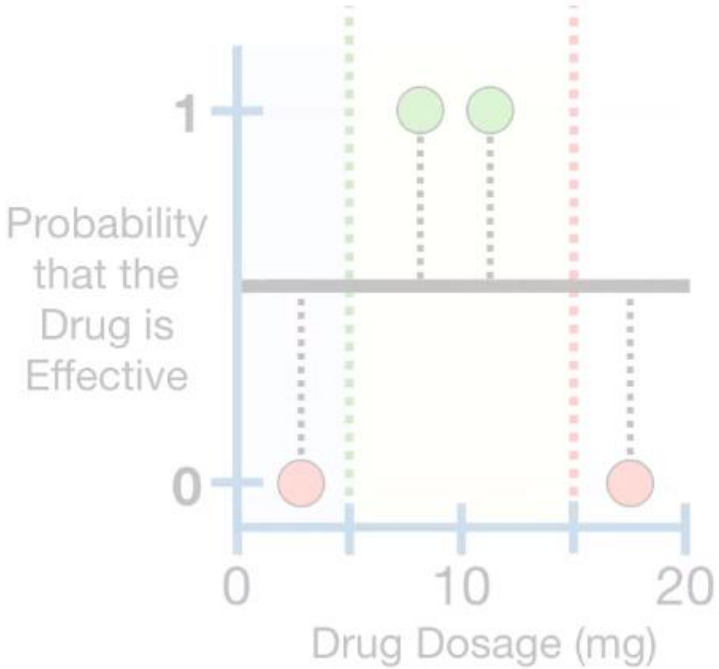


Predicted Drug Effectiveness

0.5

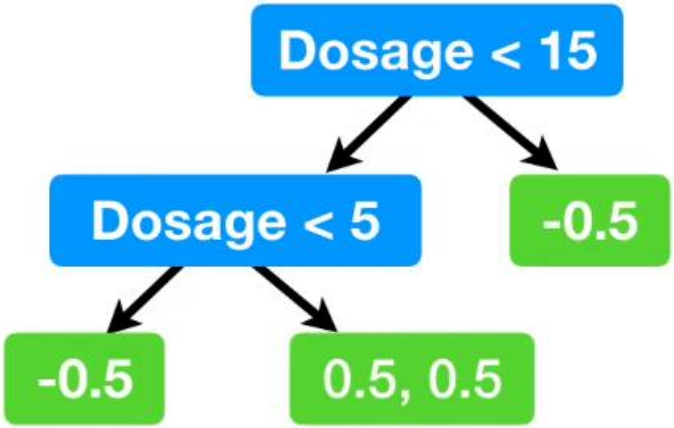


NOTE: We stopped growing this tree because we limited the number of levels to **2**...

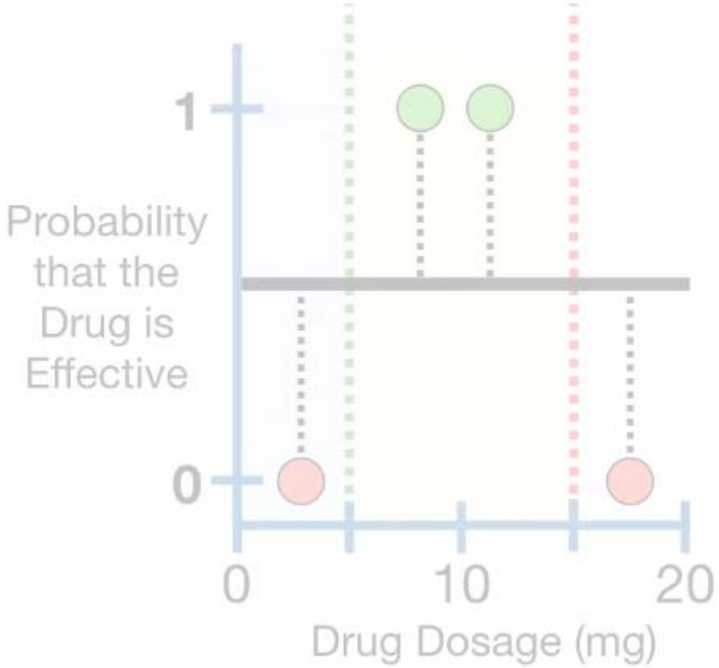


Predicted Drug Effectiveness

0.5

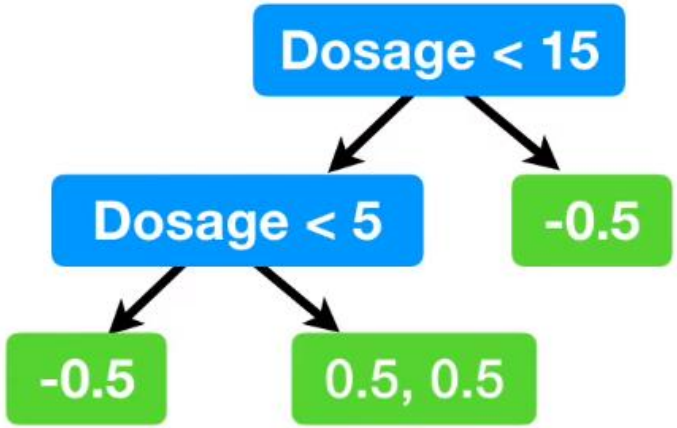


…however, **XGBoost** also has a threshold for the minimum number of **Residuals** in each leaf.

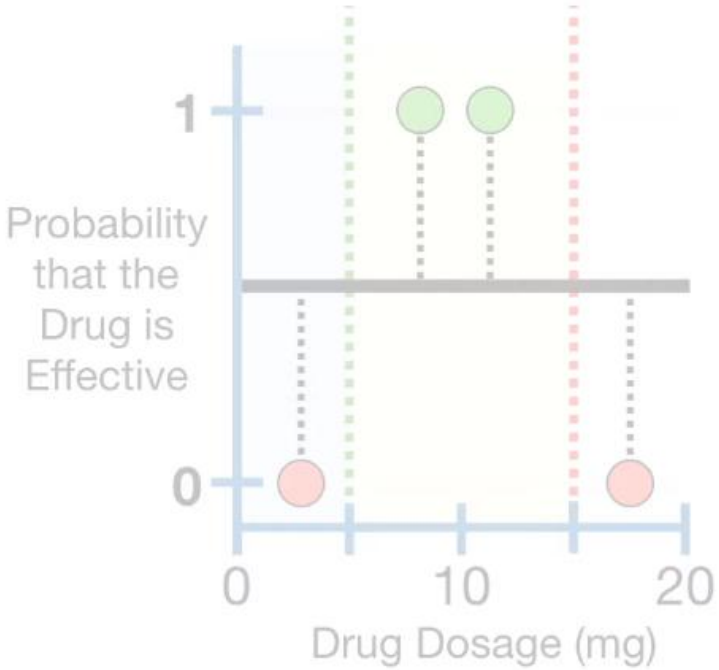


Predicted Drug Effectiveness

0.5

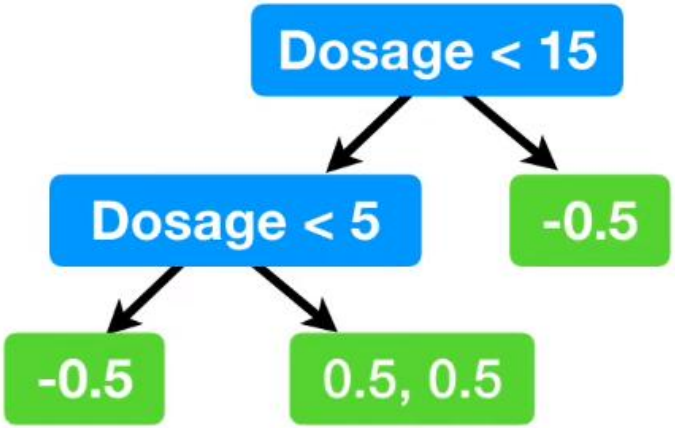


WARNING!!! It's time for some tedious detail and terminology!!!

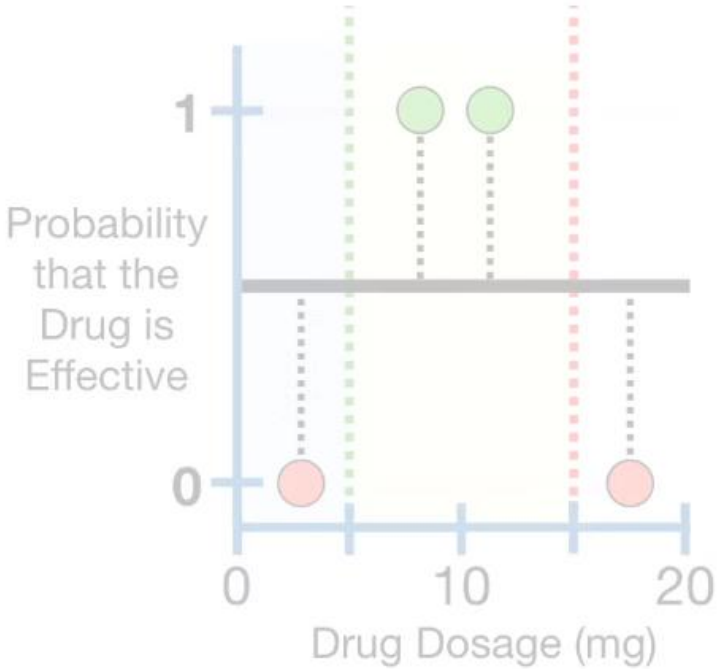


Predicted Drug Effectiveness

0.5

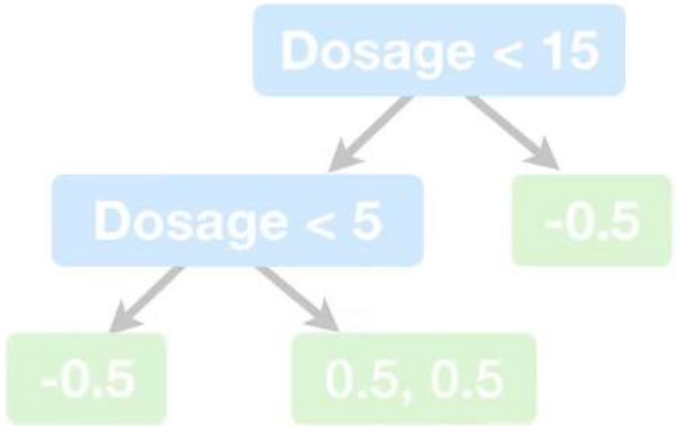


The minimum number of **Residuals** in each leaf is determined by calculating something called **Cover**.

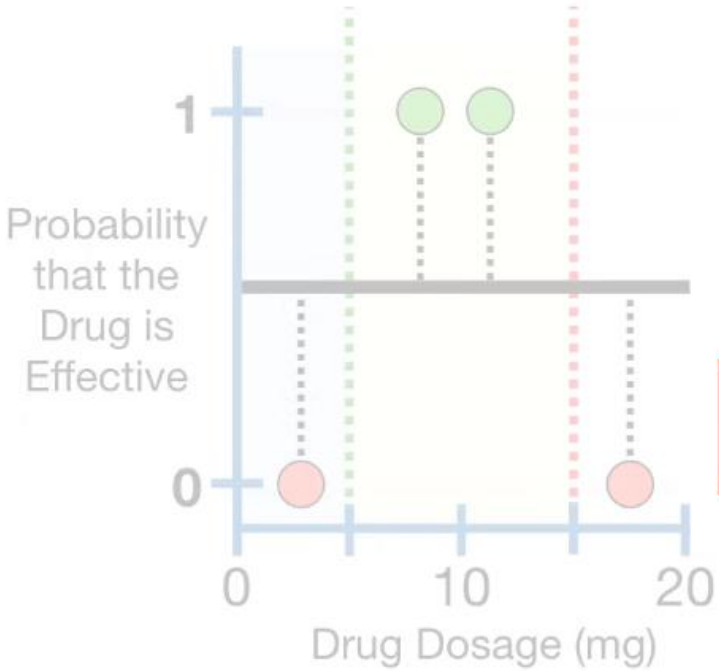


Predicted Drug Effectiveness

0.5



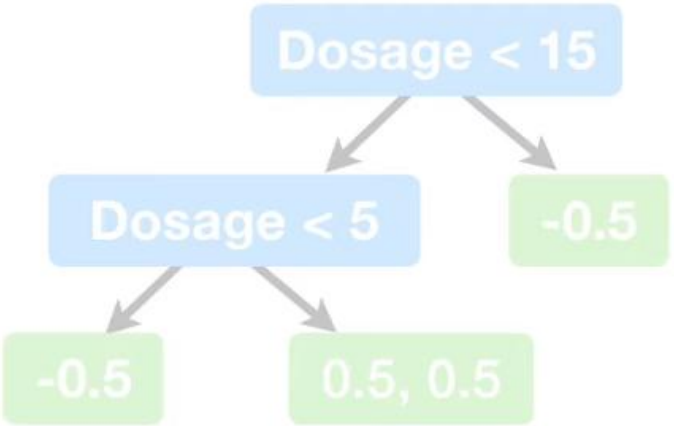
Cover is defined as the denominator of the **Similarity Score** minus **λ (lambda)**.



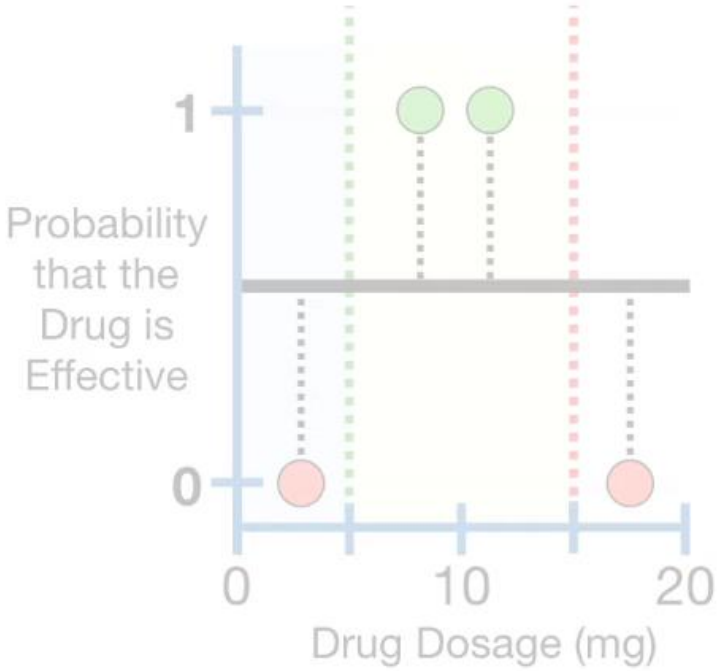
$$\text{Similarity} = \frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

Predicted Drug Effectiveness

0.5



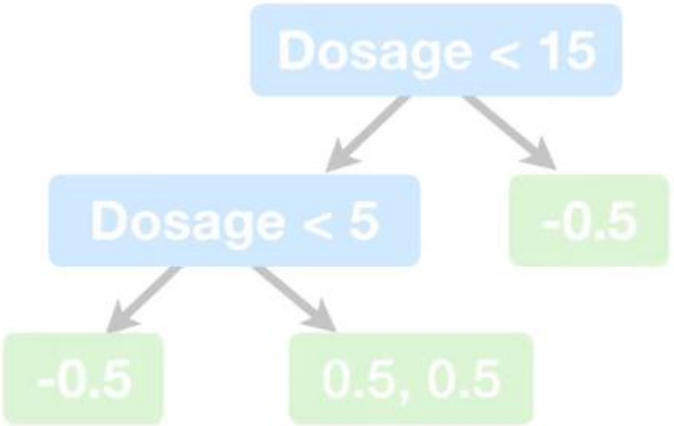
In other words, when we are using **XGBoost** for **Classification**, **Cover** is equal to…



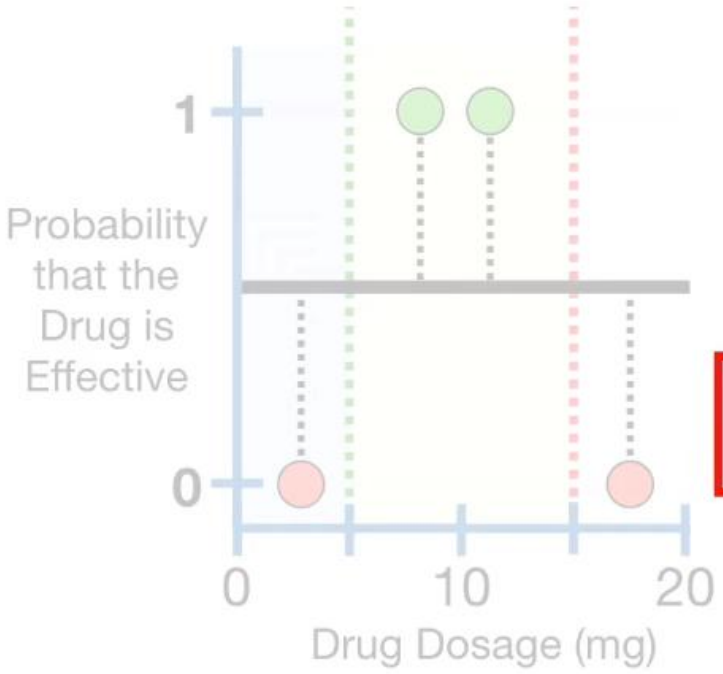
$$\text{Similarity} = \frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5



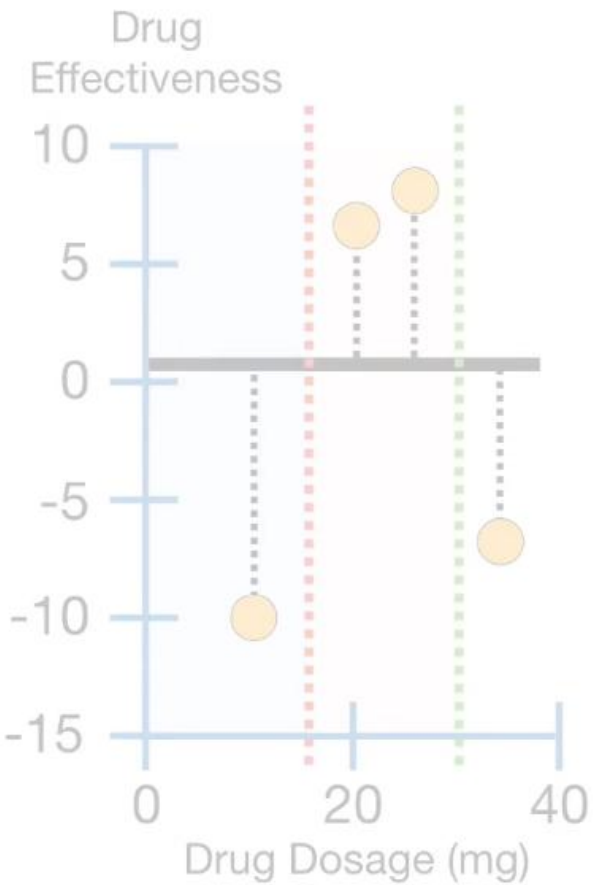
In other words, when we are using **XGBoost** for **Classification**, **Cover** is equal to…



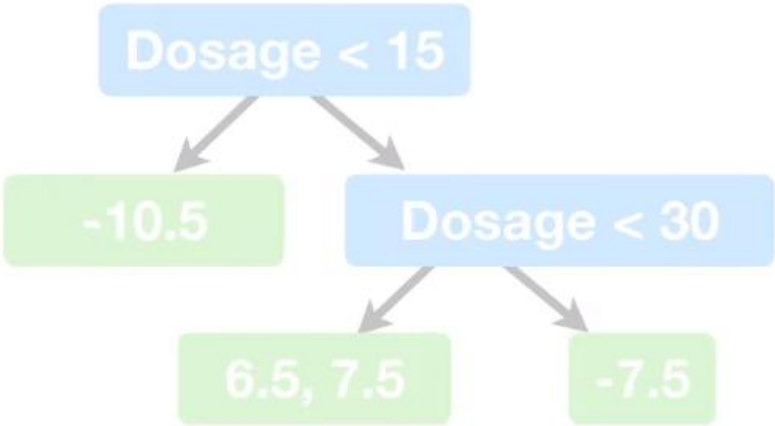
$$\text{Similarity} = \frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5



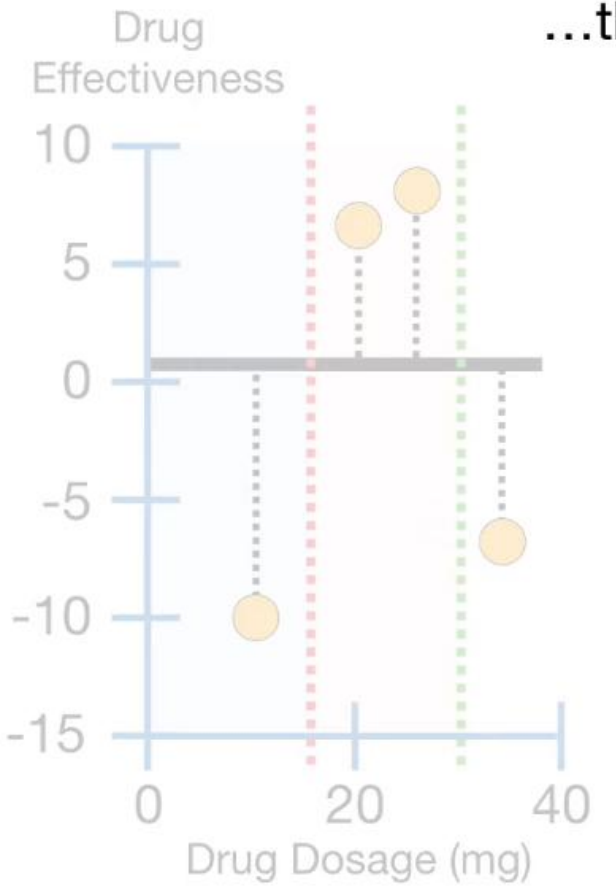
In contrast, when **XGBoost** is used for **Regression** and we are using this formula for the **Similarity Score**…



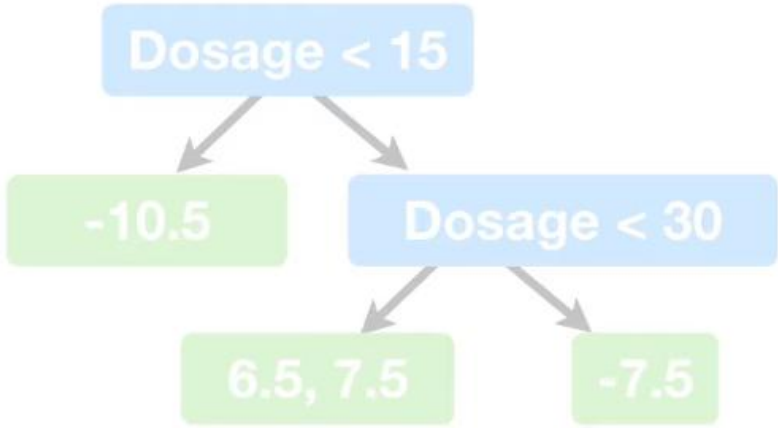
$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

Predicted Drug Effectiveness

0.5



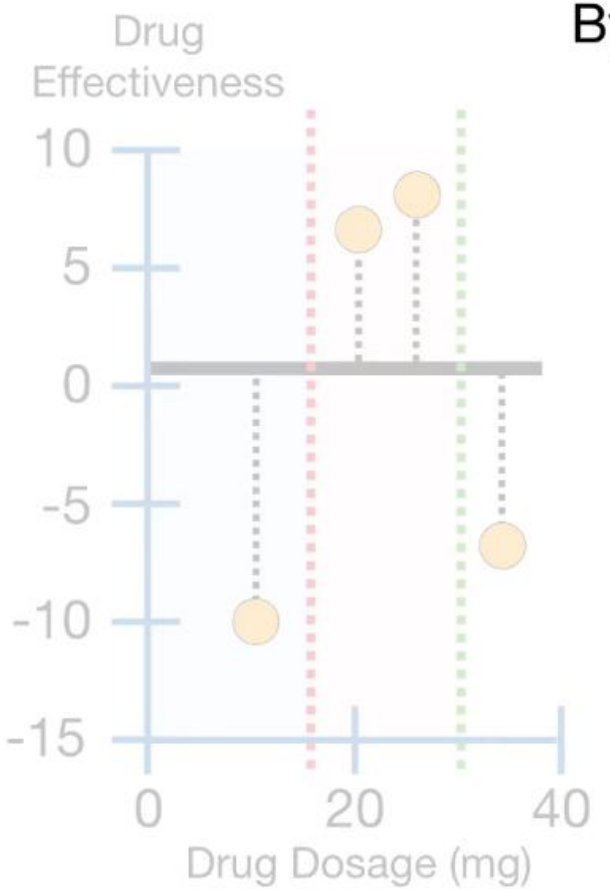
…then **Cover** is equal to…



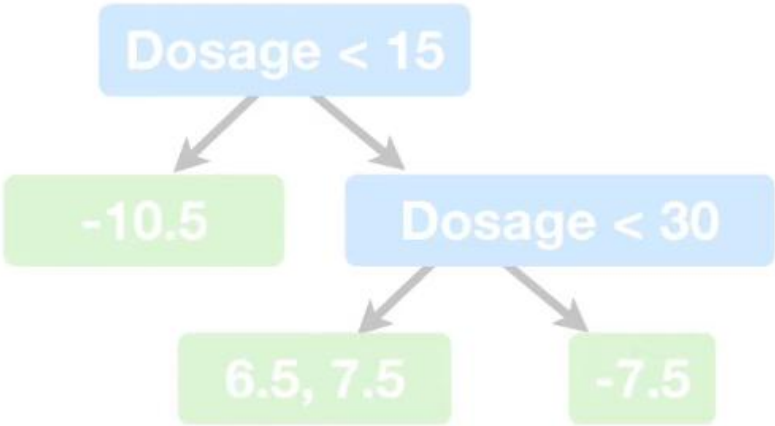
$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

Predicted Drug Effectiveness

0.5



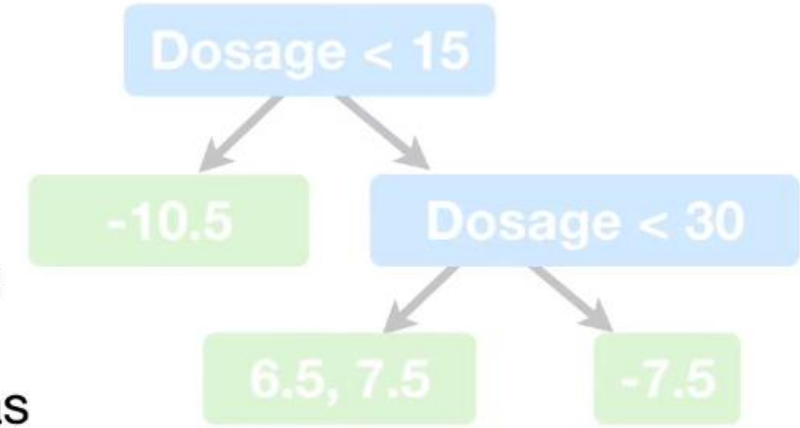
By default, the minimum value for **Cover** is **1**.

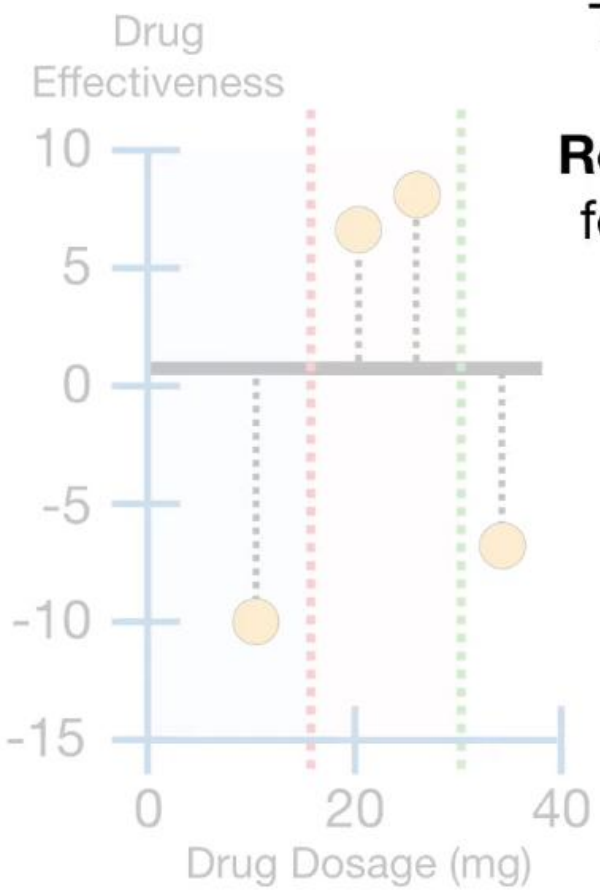


$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

Predicted Drug Effectiveness

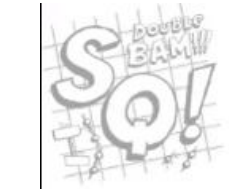
0.5





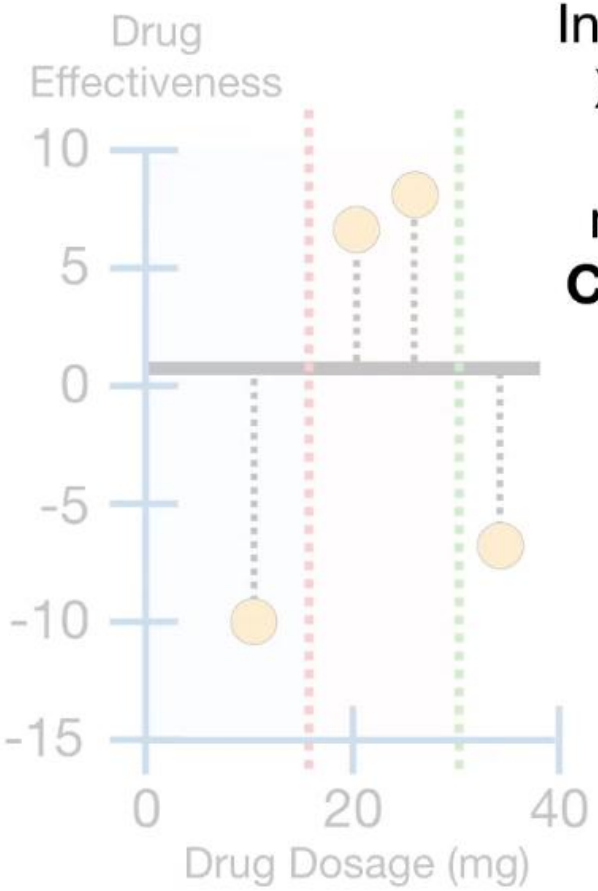
Thus, by default, when we use **XGBoost** for **Regression**, we can have as few as **1 Residual** per leaf.

$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

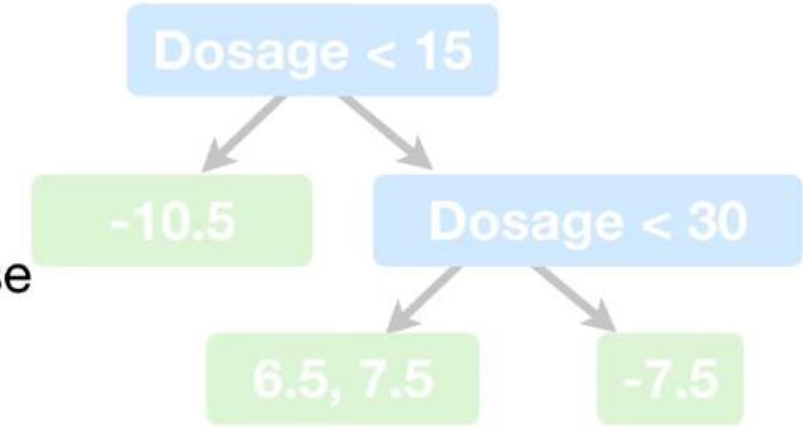


Predicted Drug Effectiveness

0.5



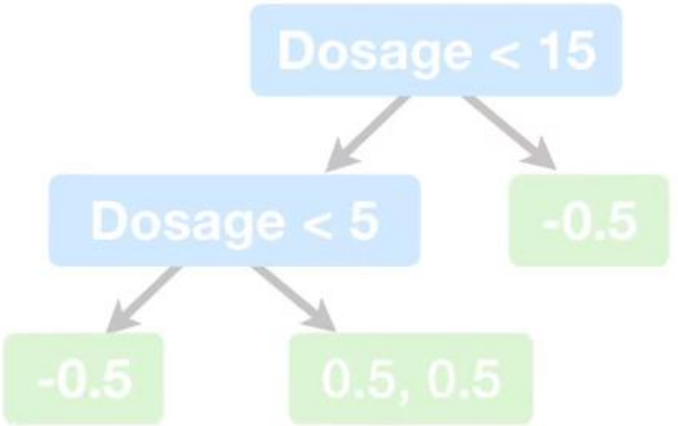
In other words, when we use **XGBoost** for **Regression** and use the default minimum value for **Cover**, **Cover** has no effect on how we grow the tree.



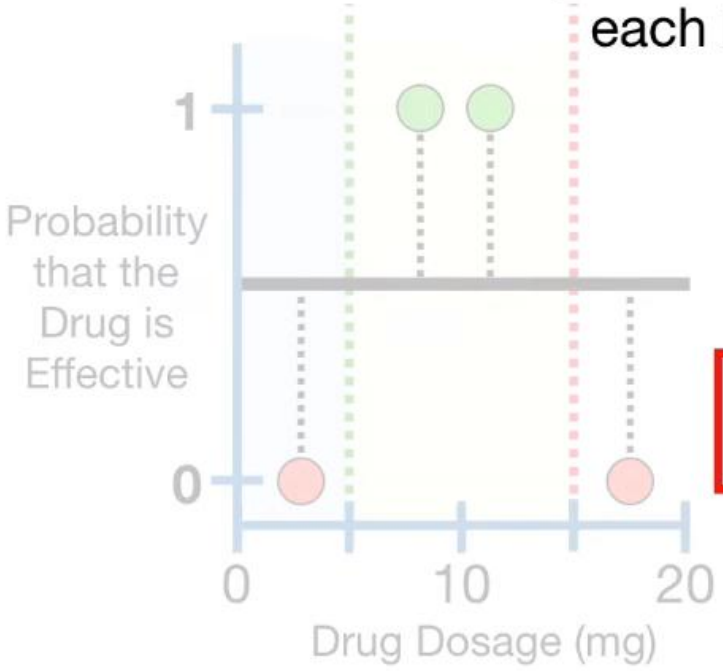
$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

Predicted Drug Effectiveness

0.5



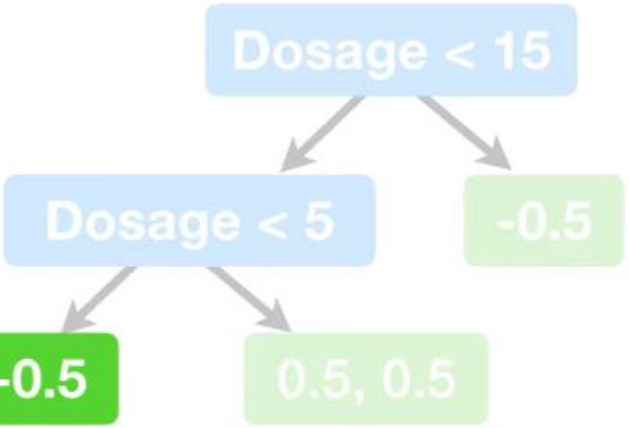
In contrast, things are way more complicated when we use **XGBoost** for **Classification** because **Cover** depends on the previously predicted probability of each **Residual** in a leaf.



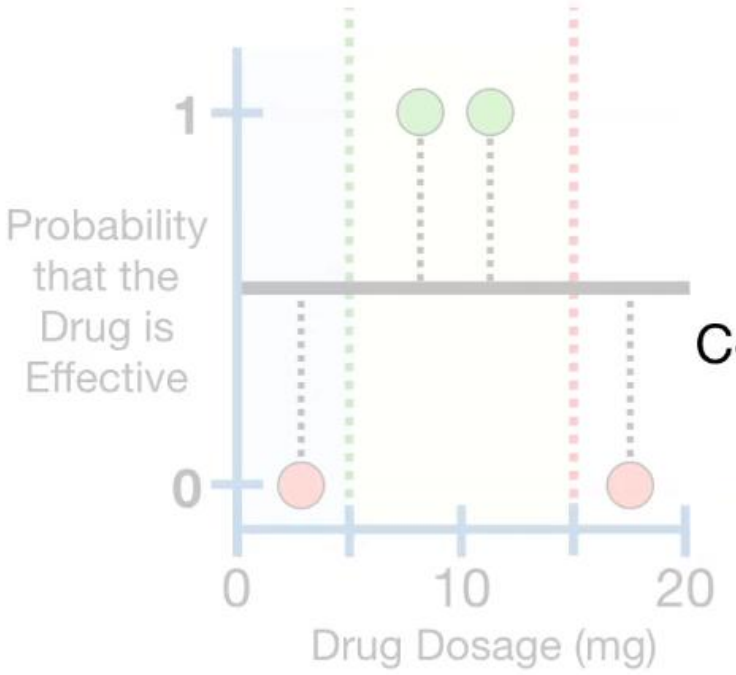
$$\text{Similarity} = \frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

Predicted Drug Effectiveness

0.5



For example, the **Cover** for this leaf is…



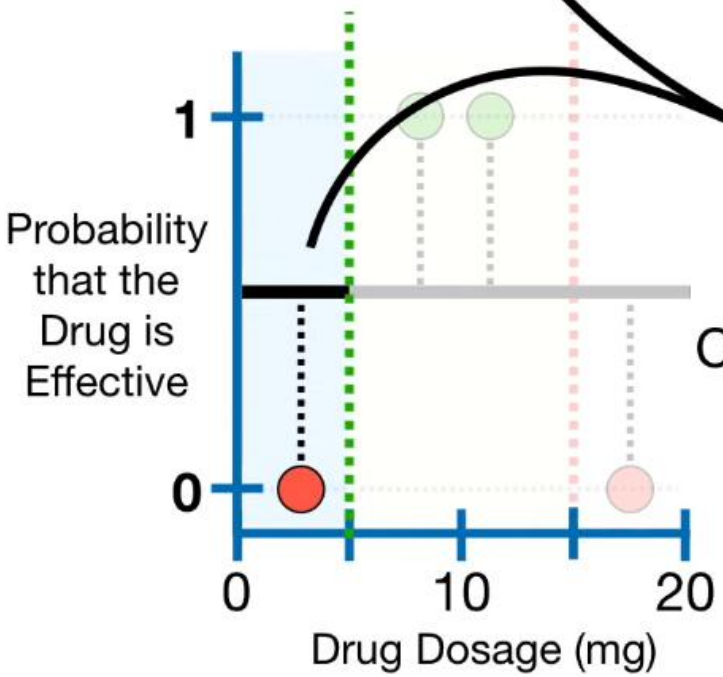
Cover =

$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5

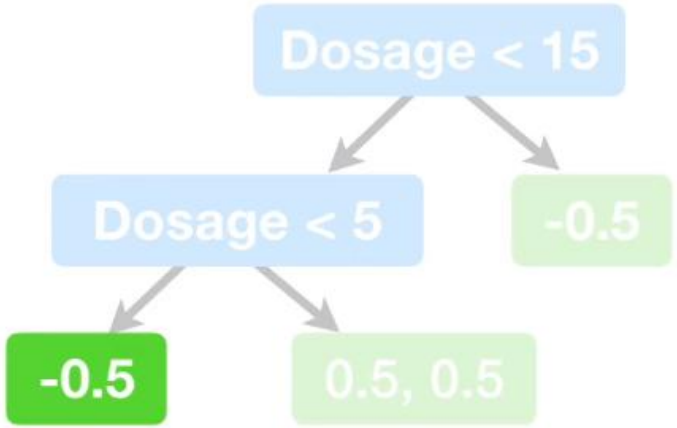
…the previously predicted probability for this observation, which was **0.5**…



Cover = 0.5

Cover =

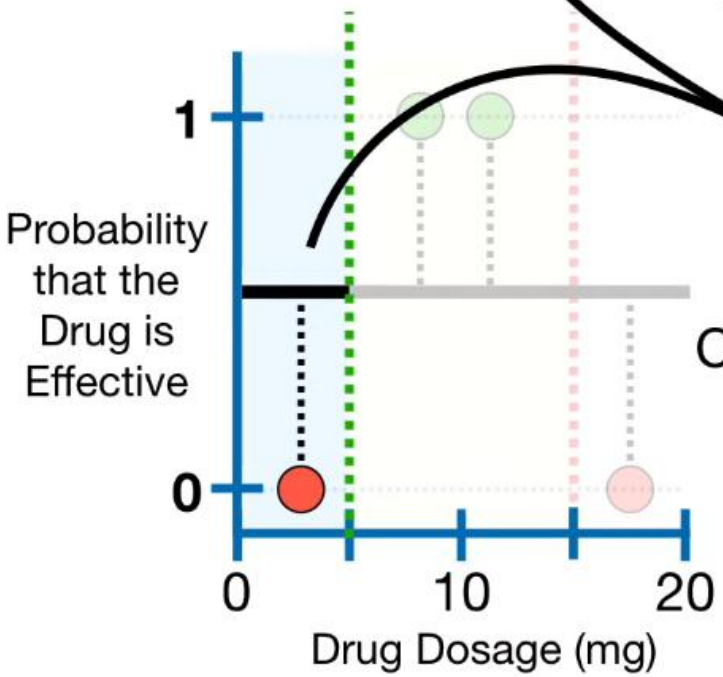
$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

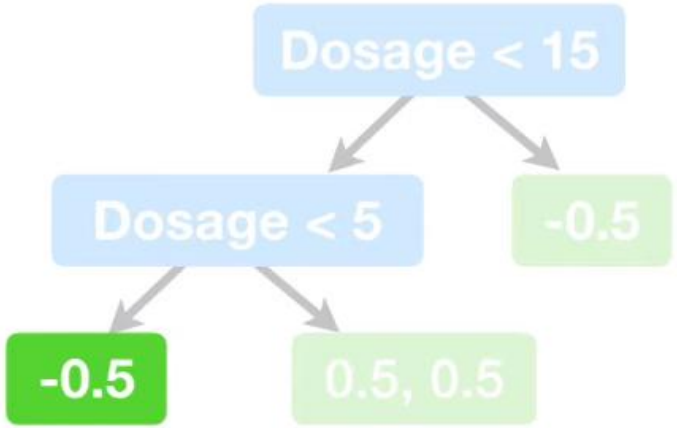


Predicted Drug Effectiveness

0.5

…times **1** minus the previously predicted probability…





Cover = 0.5 x (1 - 0.5)

Cover =

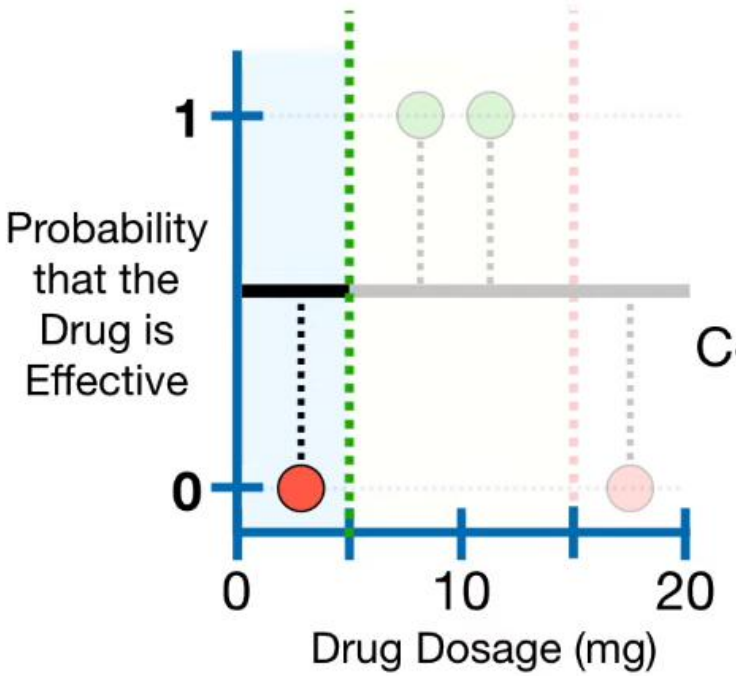
$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i) \right]$$

Predicted Drug Effectiveness

0.5

…which is **0.25**.

Cover = 0.5 x (1 - 0.5) = 0.25

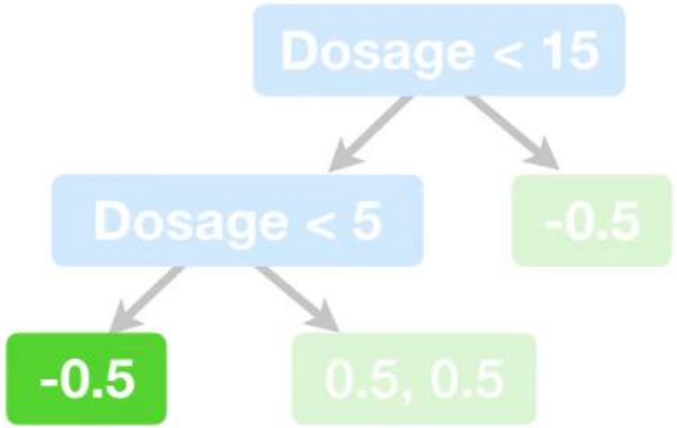


Cover =

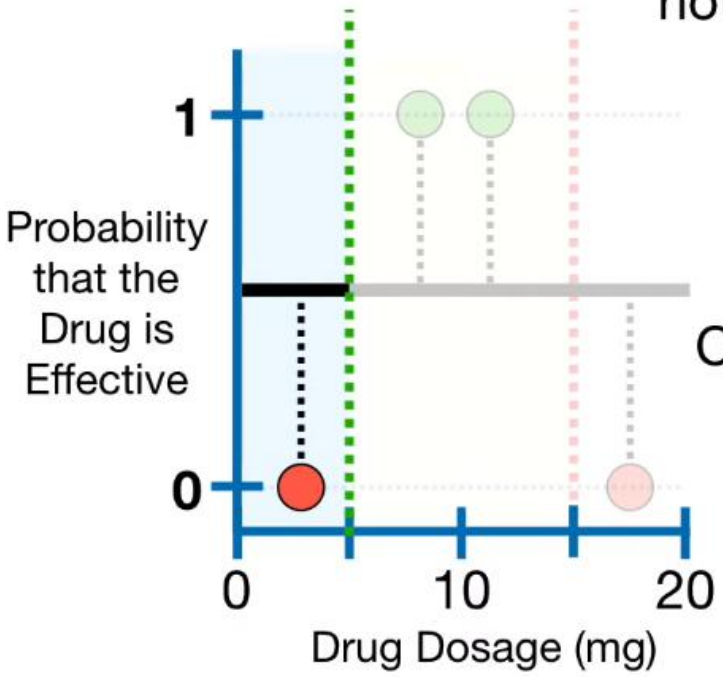
$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5



And and since the default value for the minimum **Cover** is **1**, **XGBoost** would not allow this leaf.

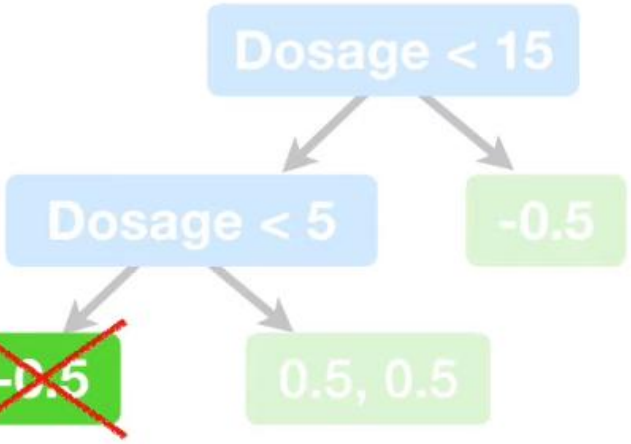


$$\text{Cover} = 0.5 \times (1 - 0.5) = 0.25$$

$$\text{Cover} = \sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

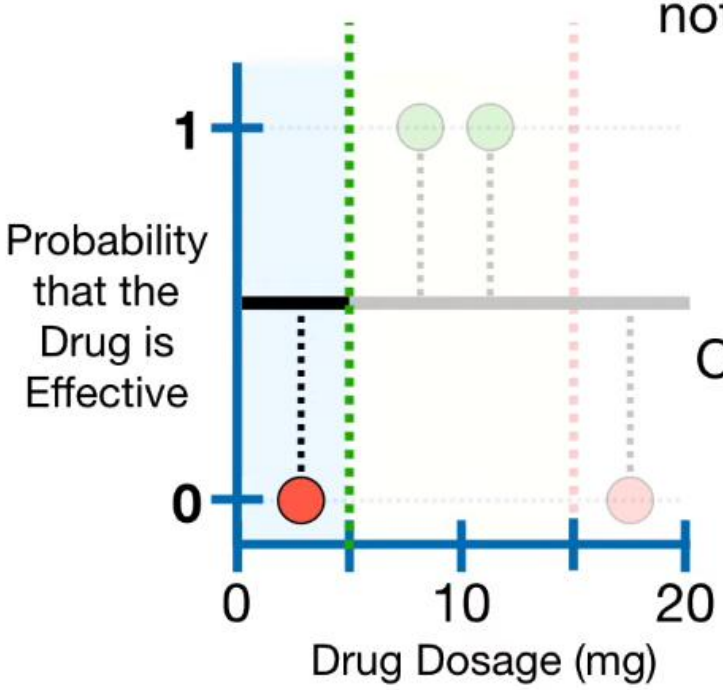
Predicted Drug Effectiveness

0.5



And and since the default value for the minimum **Cover** is **1**, **XGBoost** would not allow this leaf.

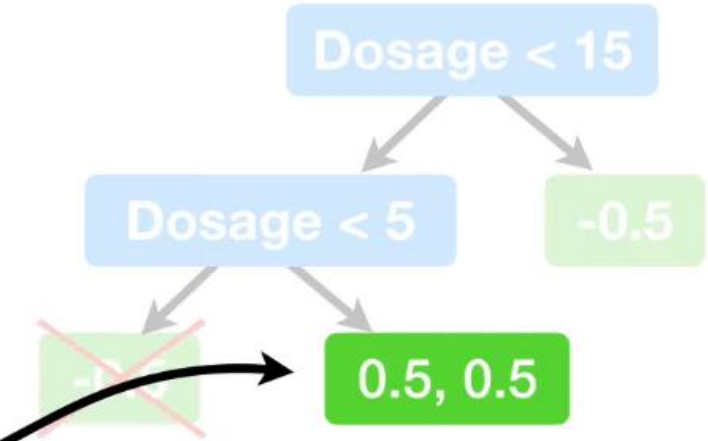
Cover = 0.5 x (1 - 0.5) = 0.25



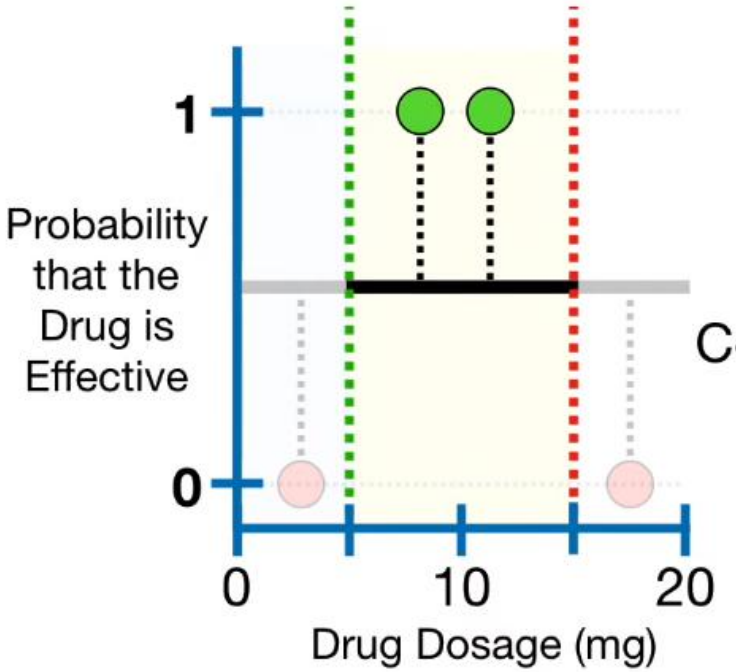
$$\text{Cover} = \sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5



Likewise, the **Cover** for this leaf…

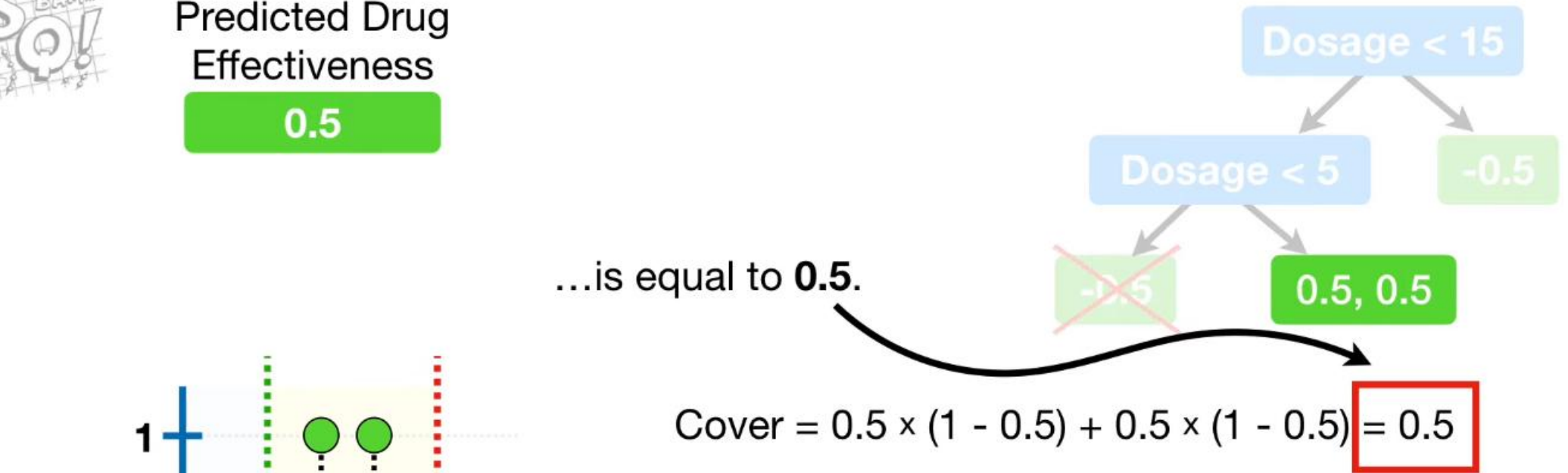


Cover =

$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

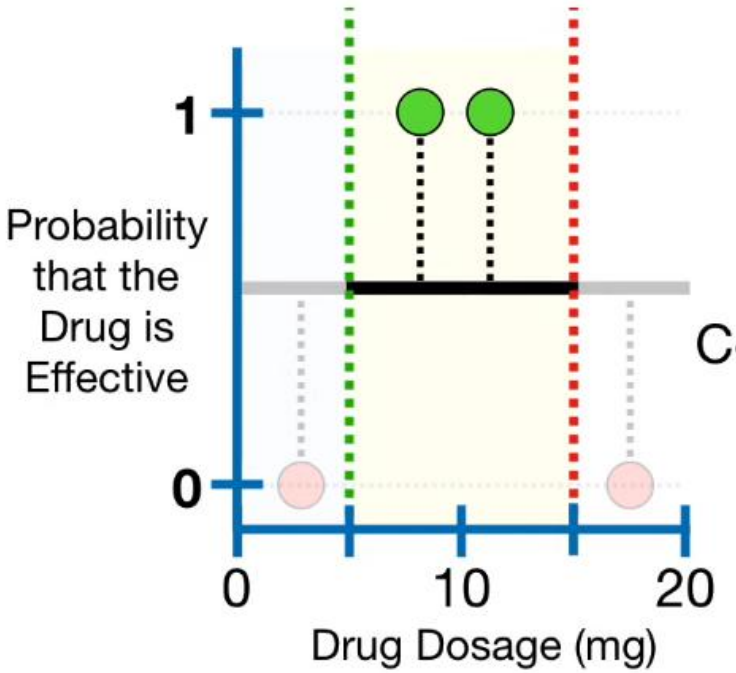
Predicted Drug Effectiveness

0.5



…is equal to **0.5**.

Cover = 0.5 x (1 - 0.5) + 0.5 x (1 - 0.5) = 0.5

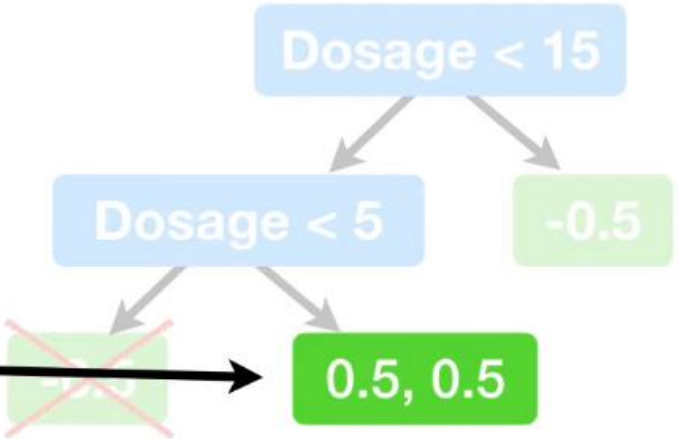


Cover =

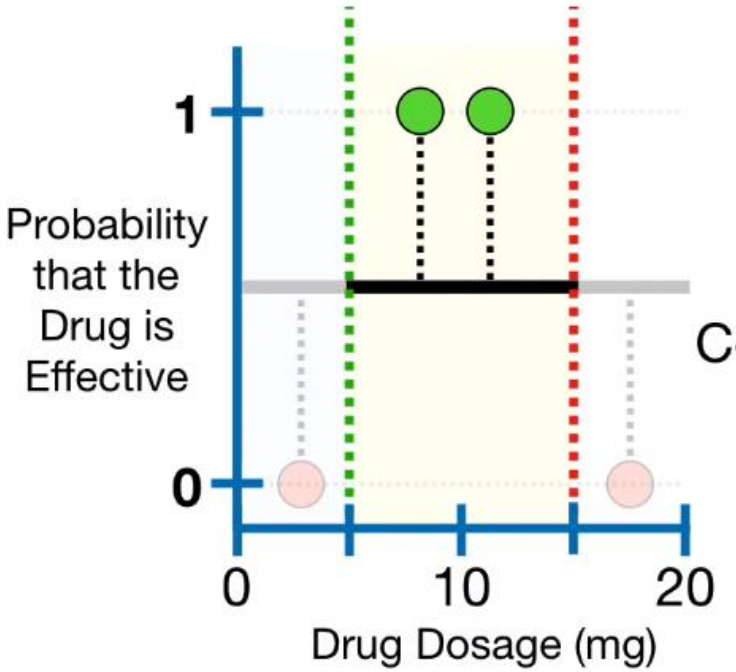
$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5



So, by default, **XGBoost** would not allow this leaf either.



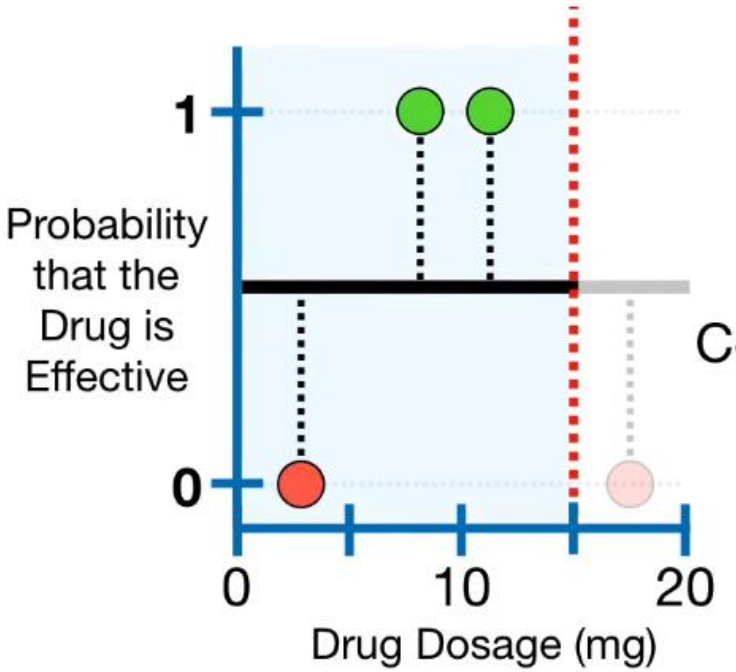
$$\text{Cover} = 0.5 \times (1 - 0.5) + 0.5 \times (1 - 0.5) = 0.5$$

$$\text{Cover} = \sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5

Since these leaves are not allowed, let's removed them…



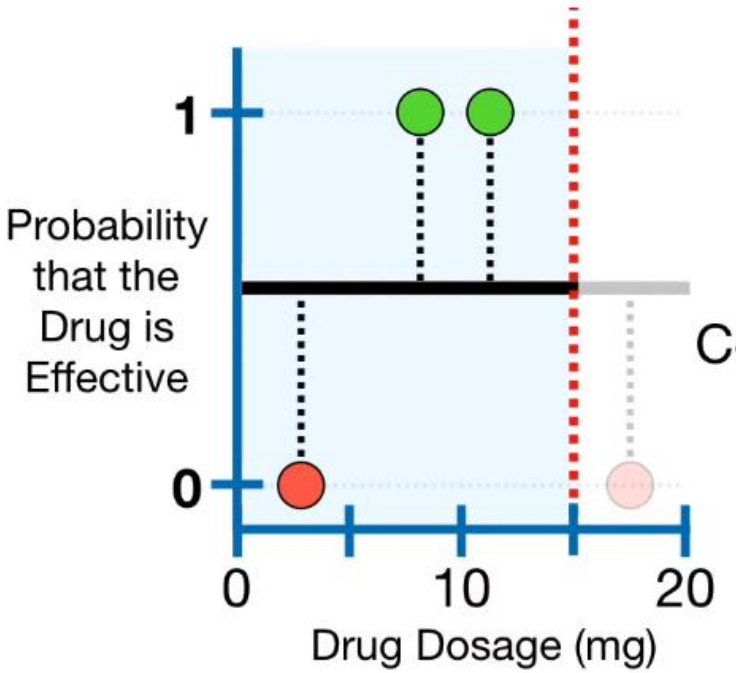
Cover =

$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5

…and go back to this leaf.



Cover =

$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

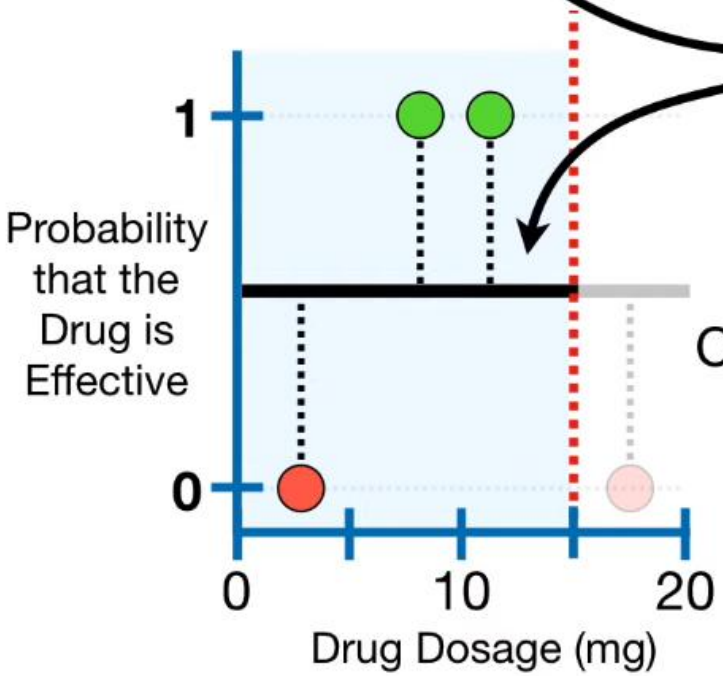
0.5

Dosage < 15

-0.5, 0.5, 0.5

-0.5

Because the previously predicted probability is the same for all three of these **Residuals**…



Cover =

$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5

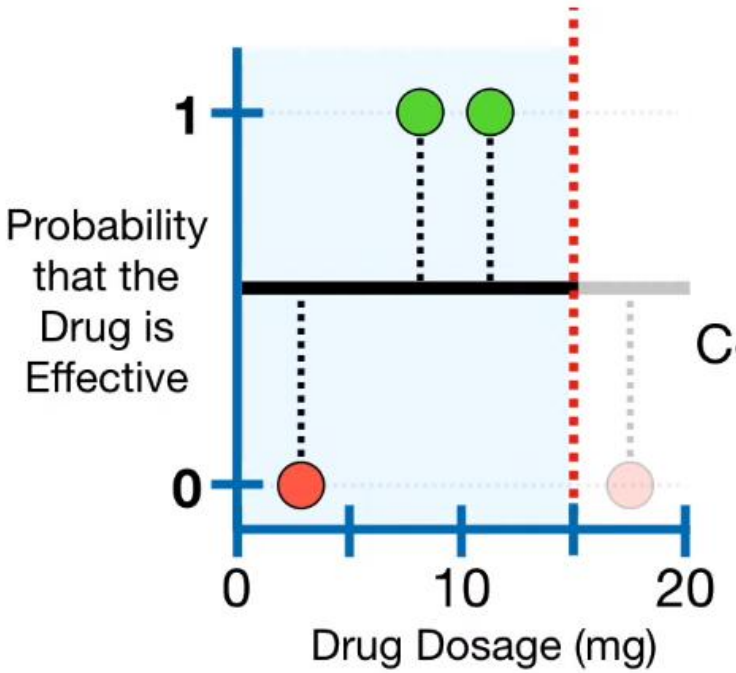
Dosage < 15

-0.5, 0.5, 0.5

-0.5

…**Cover** is just **3** times the **Cover** for one of the **Residuals**…

Cover = 3 × [0.5 × (1 - 0.5)]



Cover =

$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

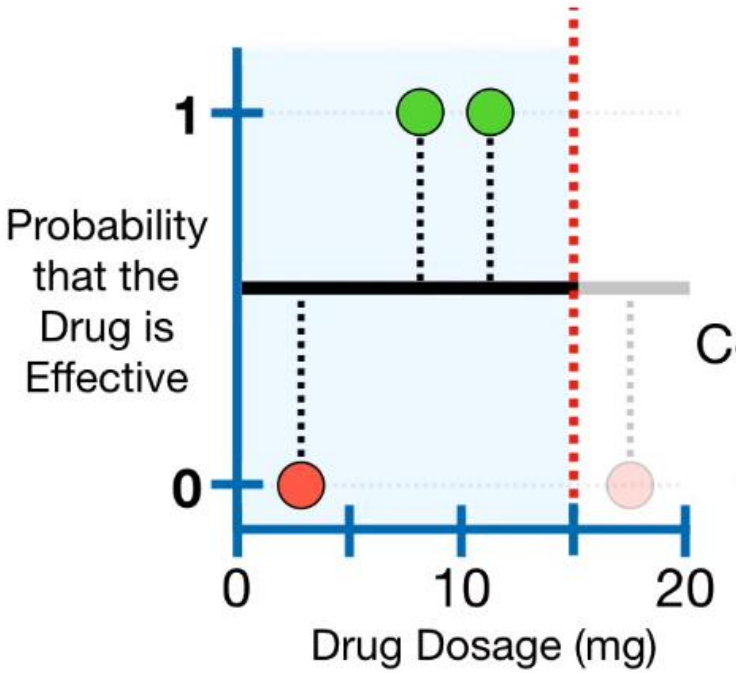
0.5

Dosage < 15

-0.5, 0.5, 0.5

-0.5

…and that means **Cover = 0.75**…



$$\text{Cover} = 3 \times [0.5 \times (1 - 0.5)] = 0.75$$

$$\text{Cover} = \sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5

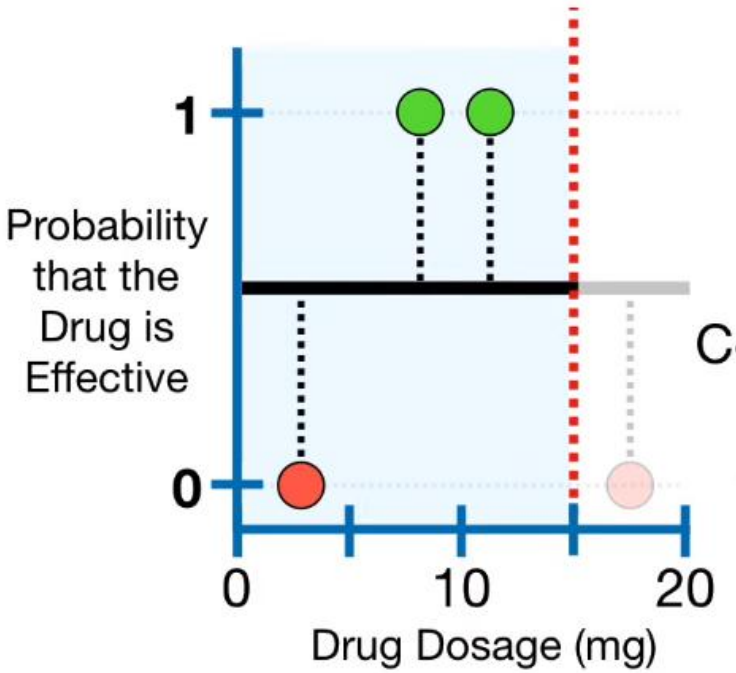
Dosage < 15

-0.5, 0.5, 0.5

-0.5

…so **XGBoost** would not allow this leaf either.

Cover = 3 × [0.5 × (1 - 0.5)] = 0.75



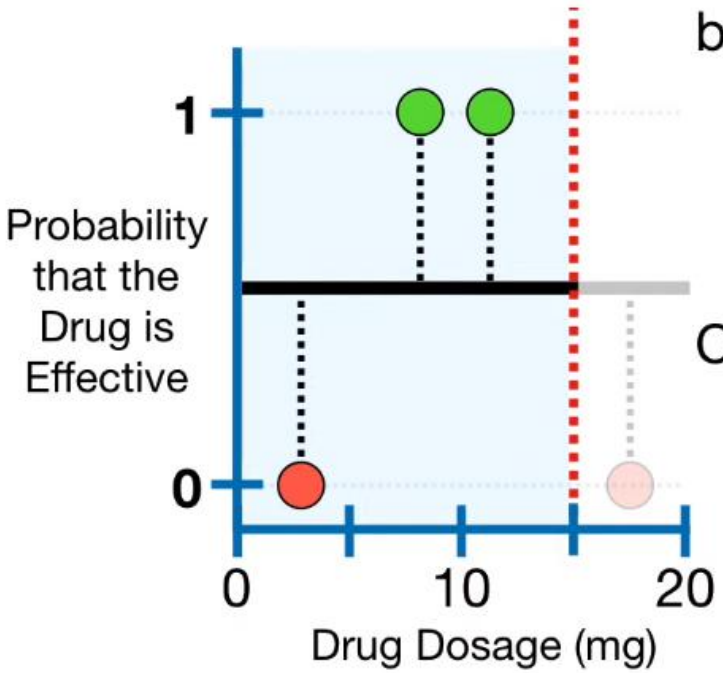
Cover =

$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5

Ultimately, if we used the default minimum value for **Cover**, **1**, then we would be left with the **Root**, and **XGBoost** requires trees to be larger than just the **Root**.



$$\text{Cover} = \sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

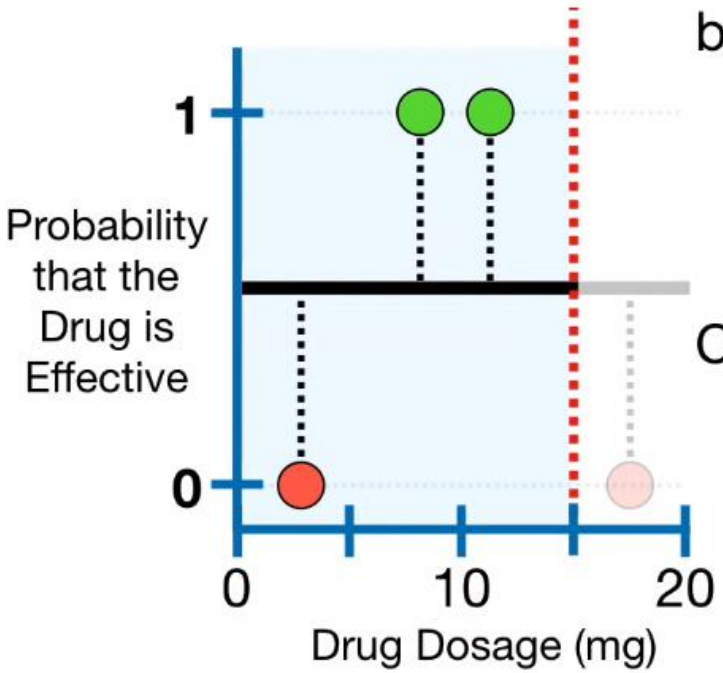
Predicted Drug Effectiveness

0.5

Dosage < 15

-0.5

Ultimately, if we used the default minimum value for **Cover**, **1**, then we would be left with the **Root**, and **XGBoost** requires trees to be larger than just the **Root**.



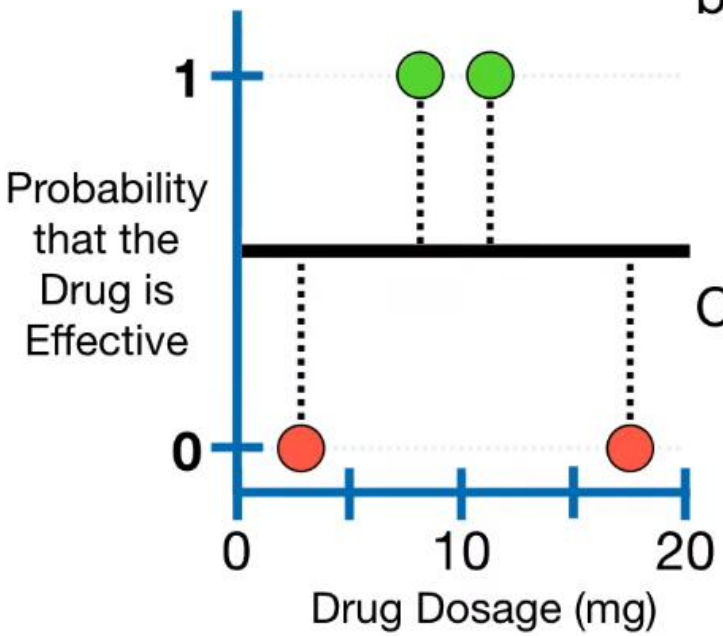
$$\text{Cover} = \sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5

-0.5, 0.5, 0.5, -0.5

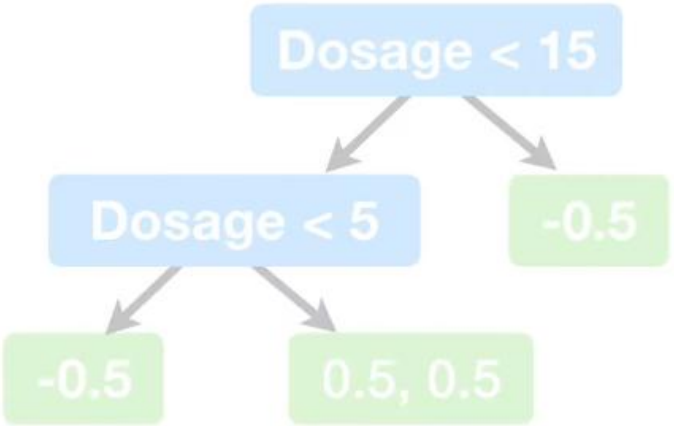
Ultimately, if we used the default minimum value for **Cover**, **1**, then we would be left with the **Root**, and **XGBoost** requires trees to be larger than just the **Root**.



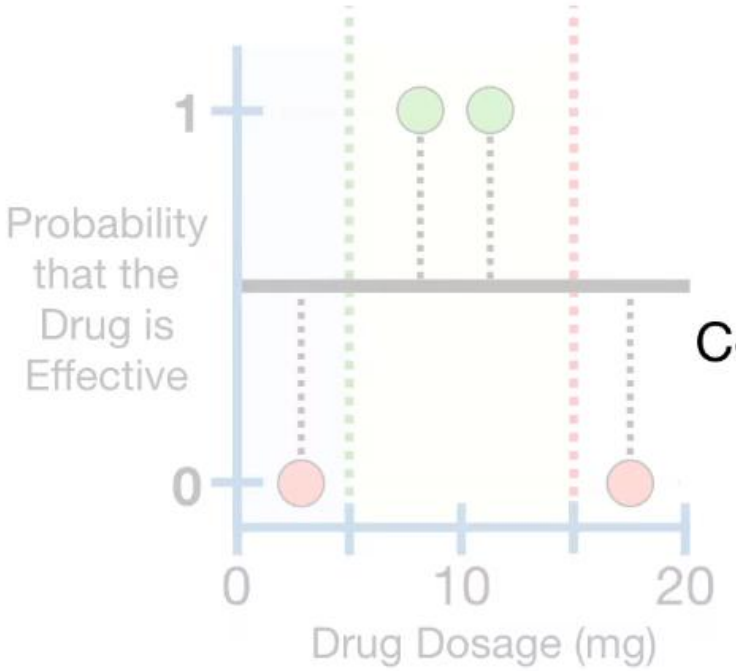
$$\text{Cover} = \sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5



So, in order to prevent this from being the worst example ever, let's set the minimum value for **Cover = 0**.

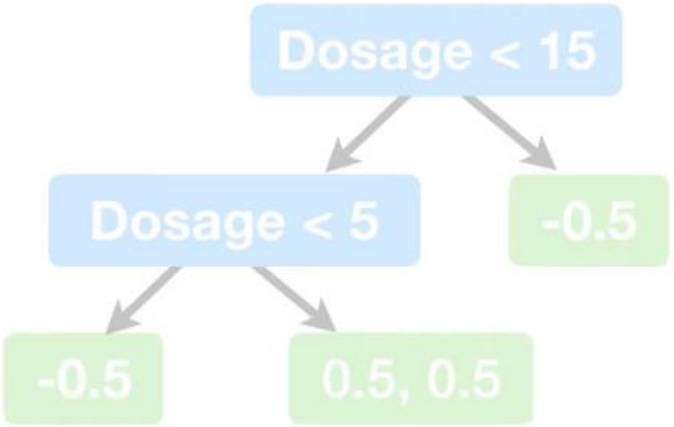


Cover =

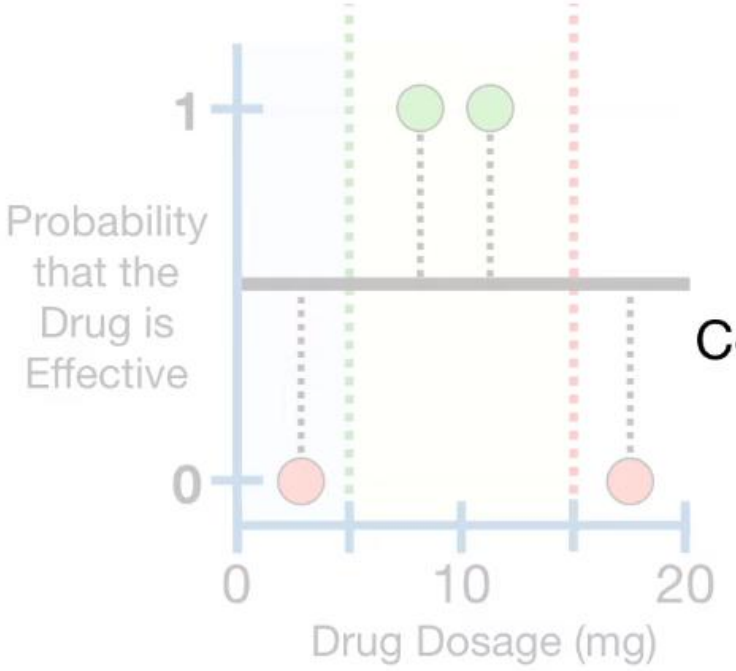
$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5



That means setting the **min_child_weight** parameter equal to **0**.

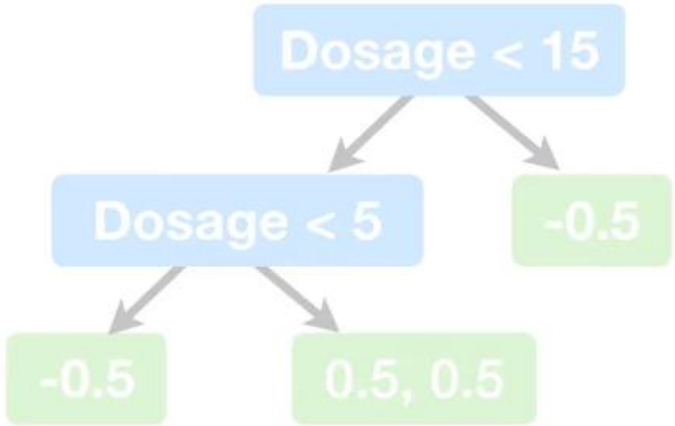


Cover =

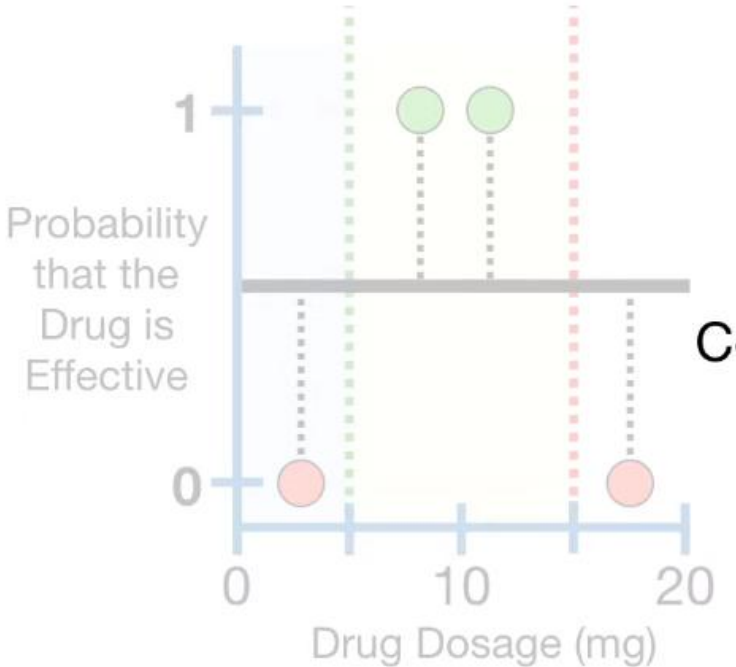
$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5



Small Bam.



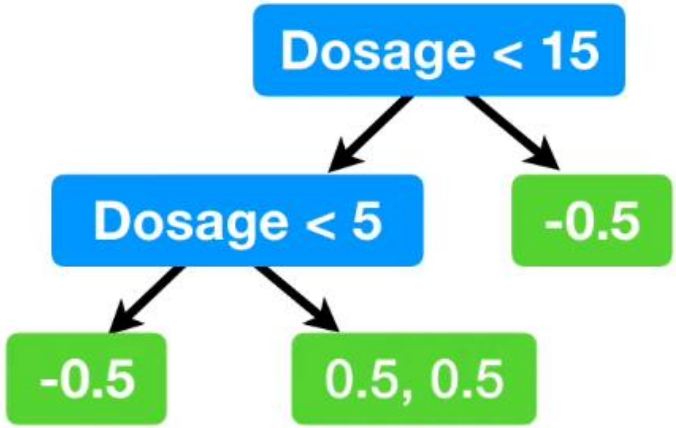
Cover =

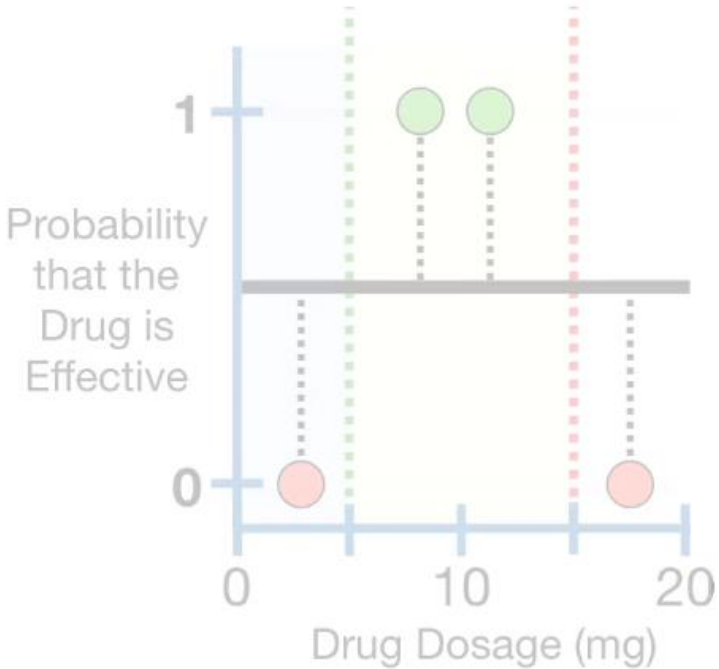
$$\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right]$$

Predicted Drug Effectiveness

0.5

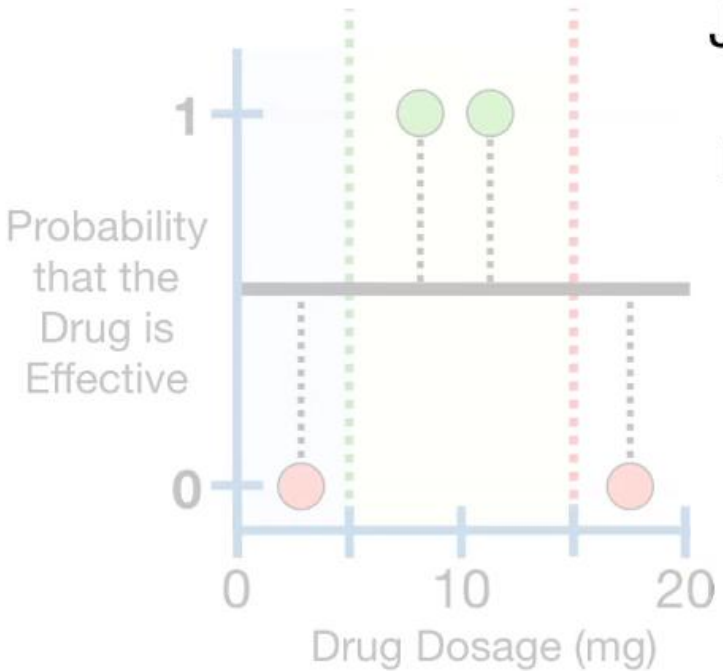
Now we can talk about how to **Prune** the tree.

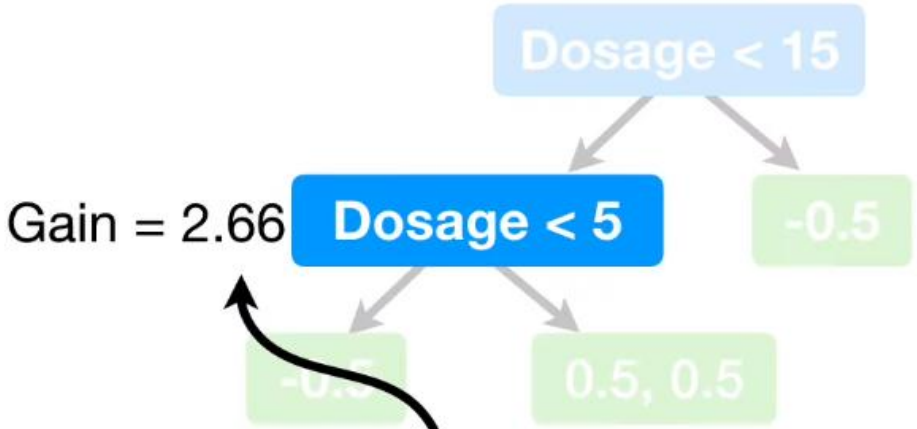




Predicted Drug Effectiveness

0.5



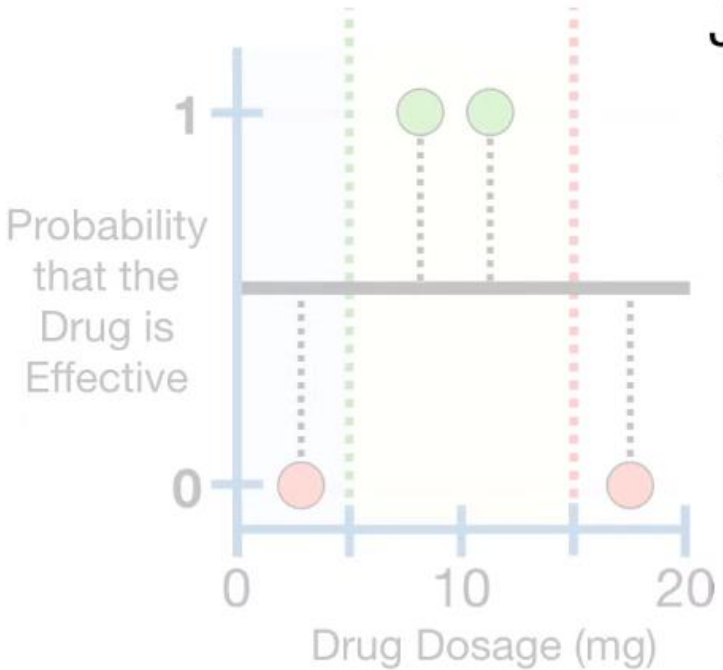


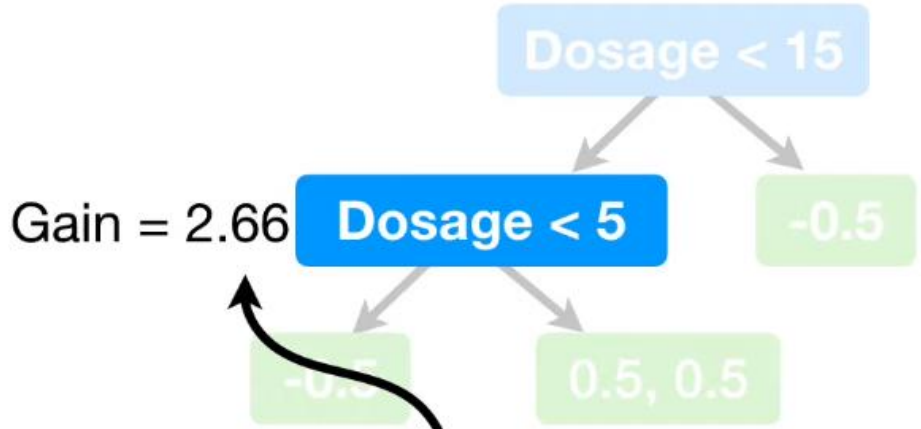
Just like we did in **Part 1**, we prune by calculating the difference between the **Gain** associated with the lowest branch and a number we pick for ***γ* (gamma)**.

Gain - ***γ*** =

Predicted Drug Effectiveness

0.5



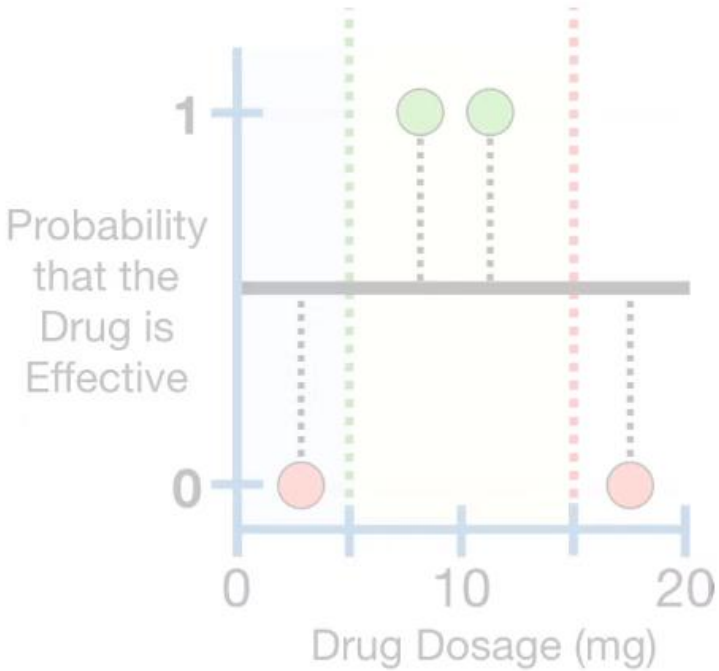


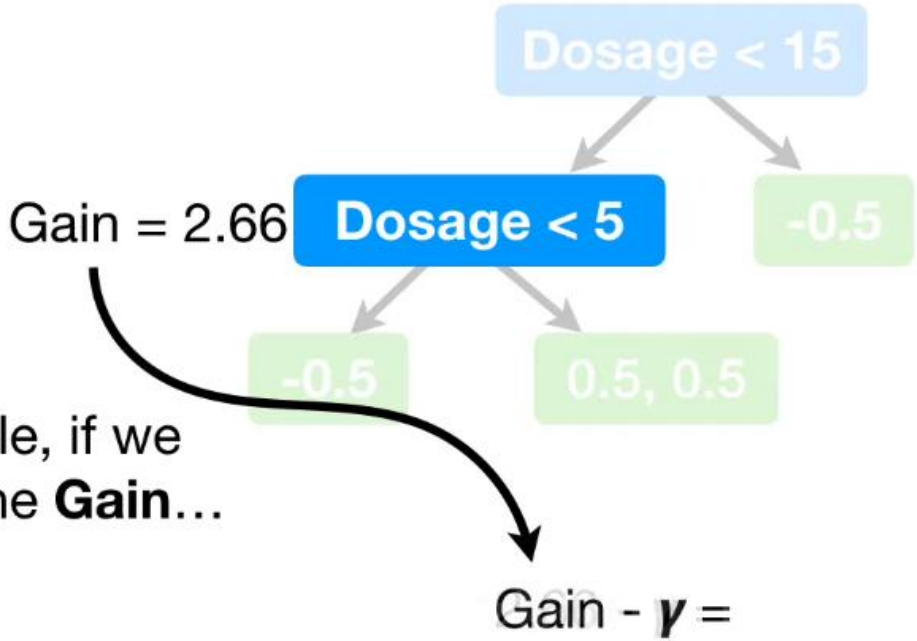
Just like we did in **Part 1**, we prune by calculating the difference between the **Gain** associated with the lowest branch and a number we pick for ***γ* (gamma)**.

Gain - ***γ*** =

Predicted Drug Effectiveness

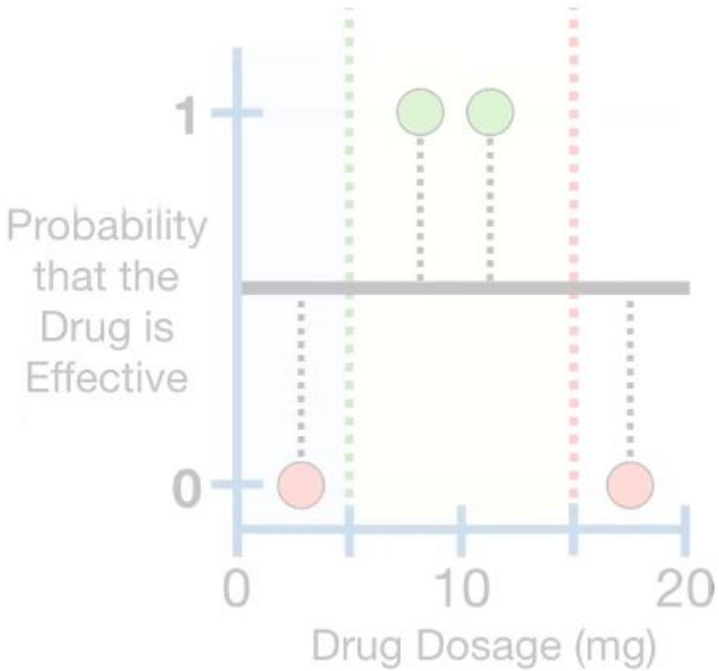
0.5

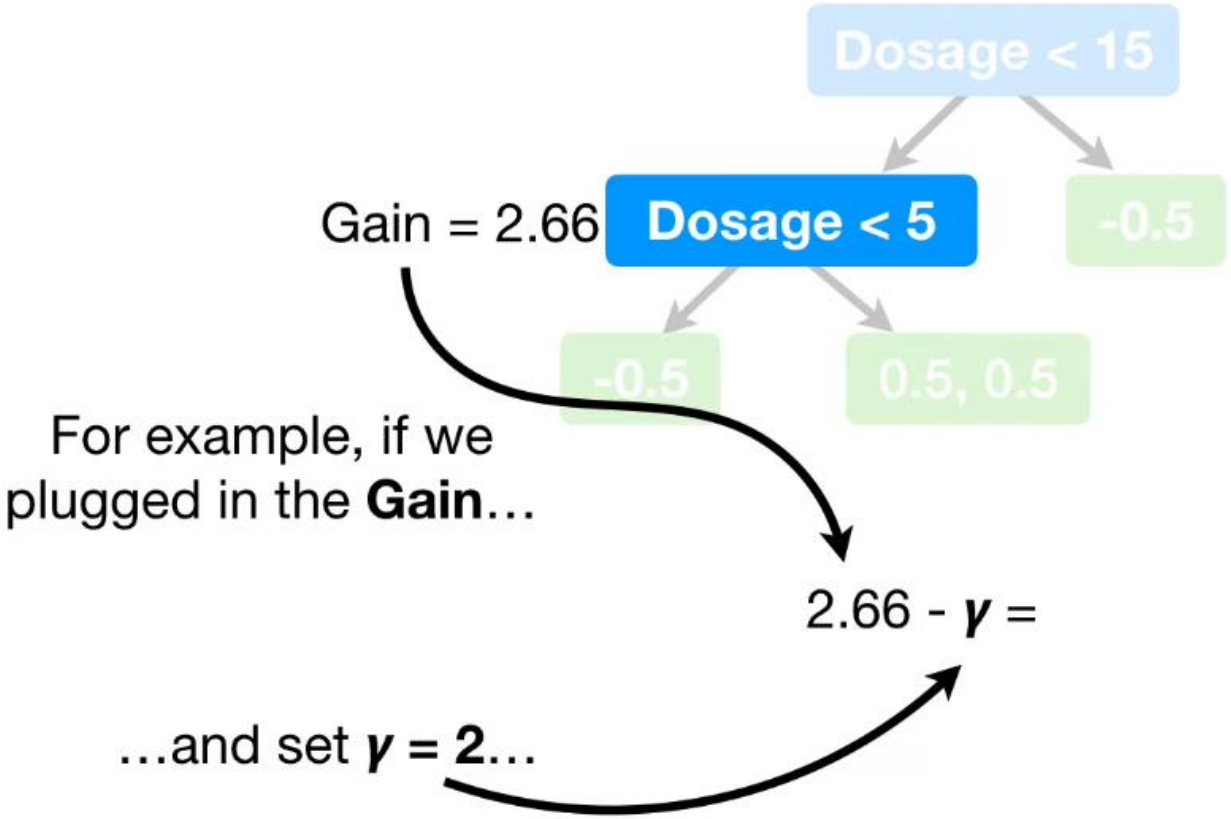




Predicted Drug Effectiveness

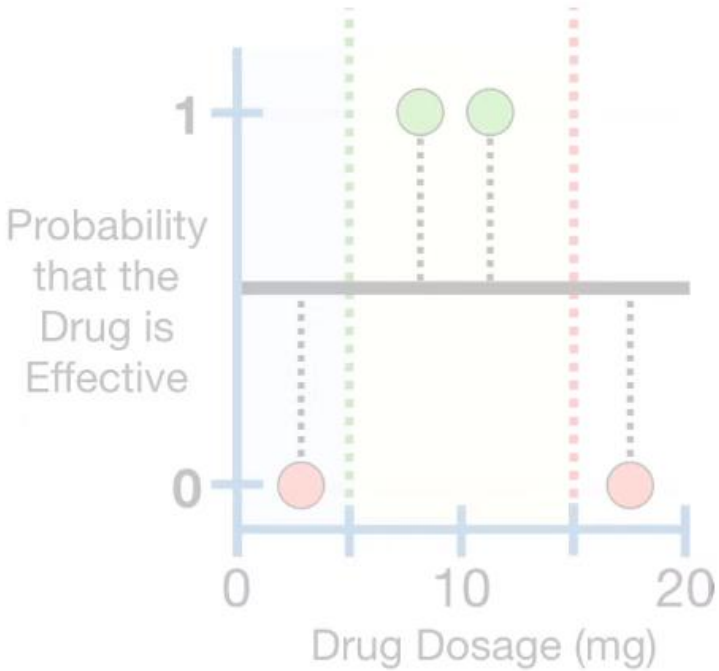
0.5

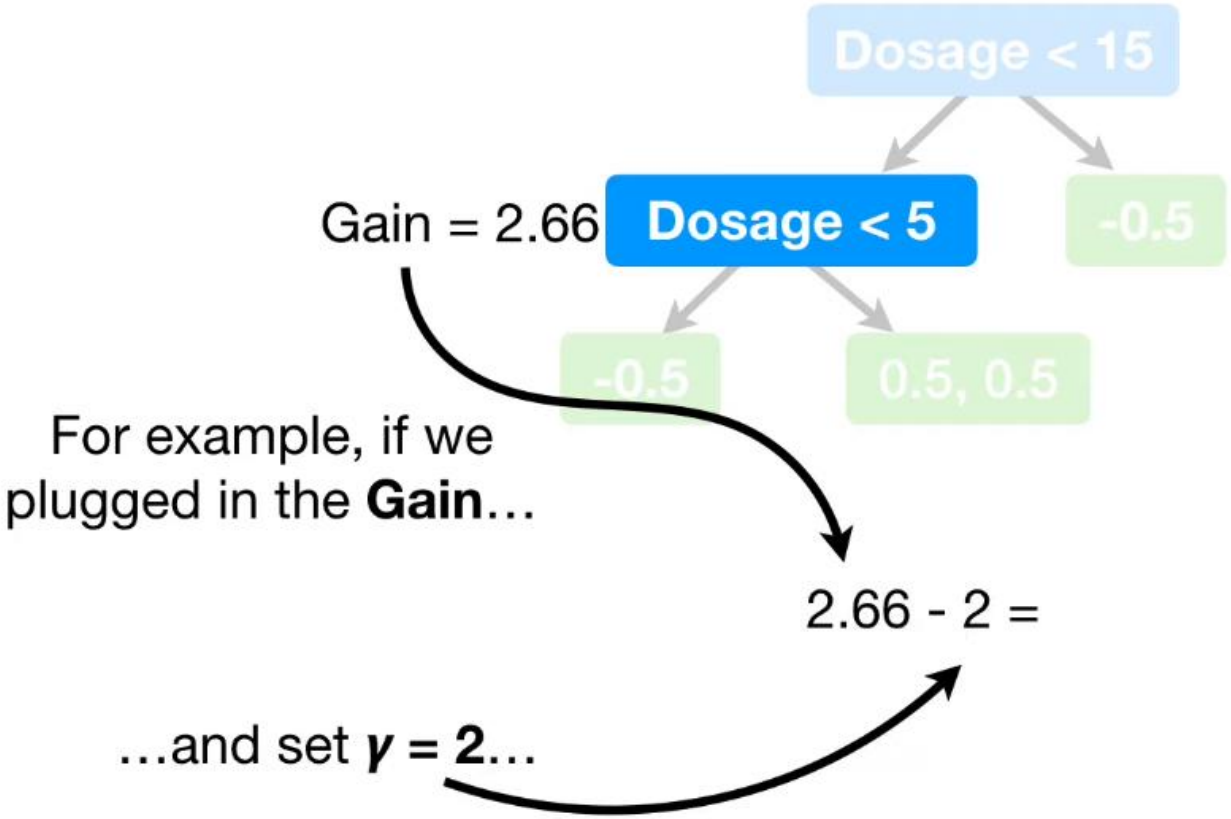




Predicted Drug Effectiveness

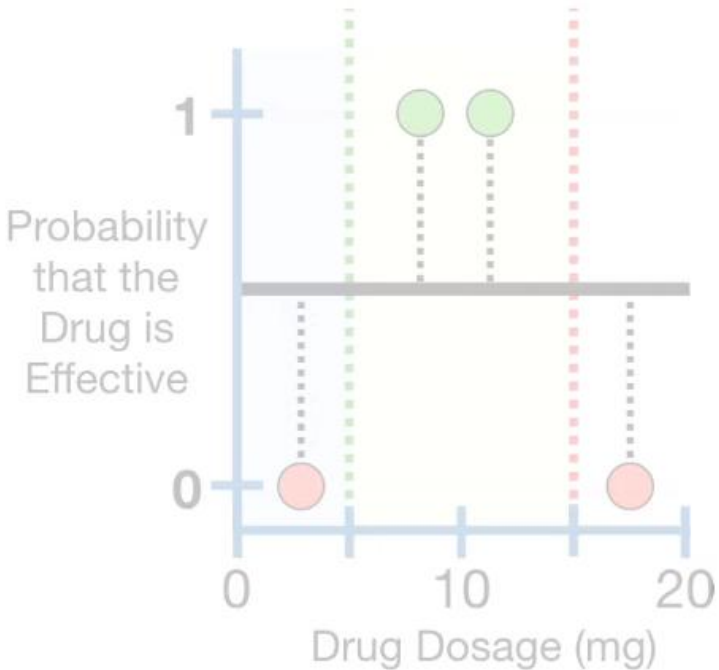
0.5

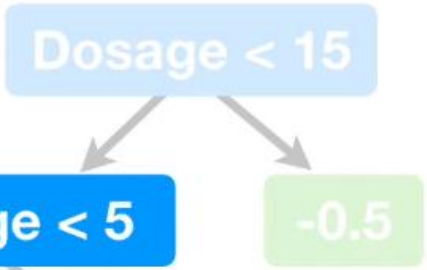




Predicted Drug Effectiveness

0.5





Gain = 2.66

For example, if we plugged in the **Gain**…

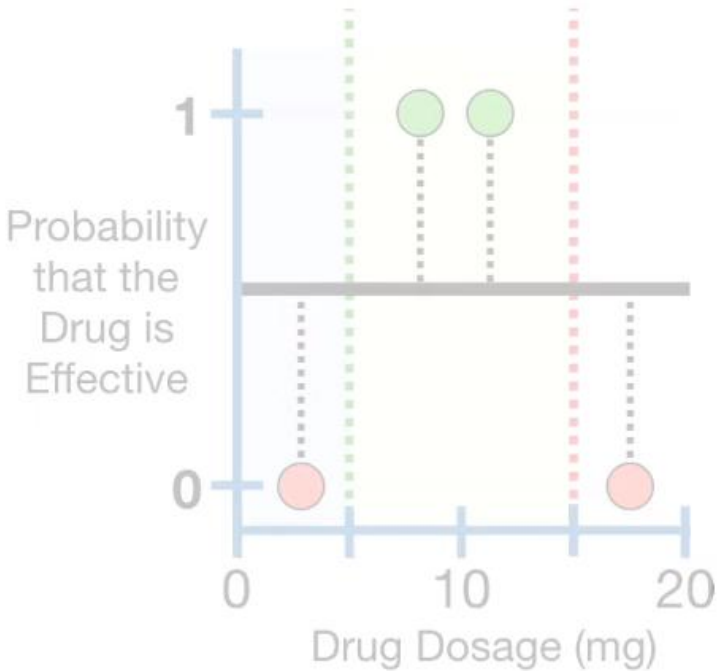
2.66 - 2 = 0.66

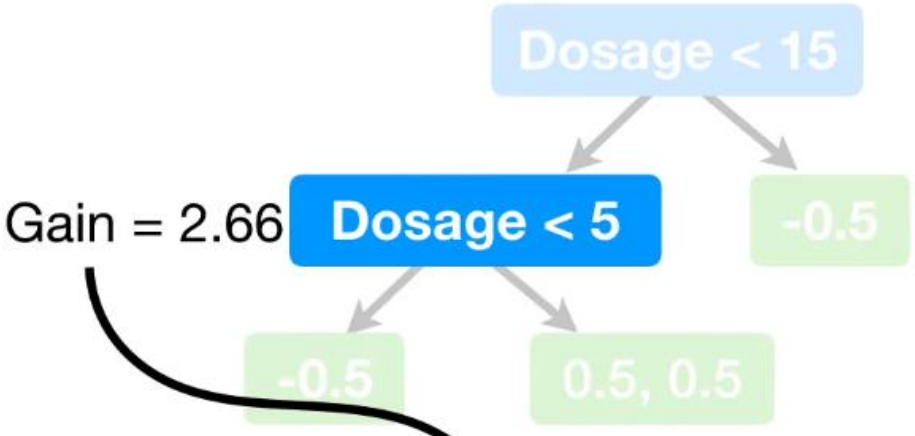
…and set ***y* = 2**…

…then we would not prune because the difference is a ***positive*** number.

Predicted Drug Effectiveness

0.5



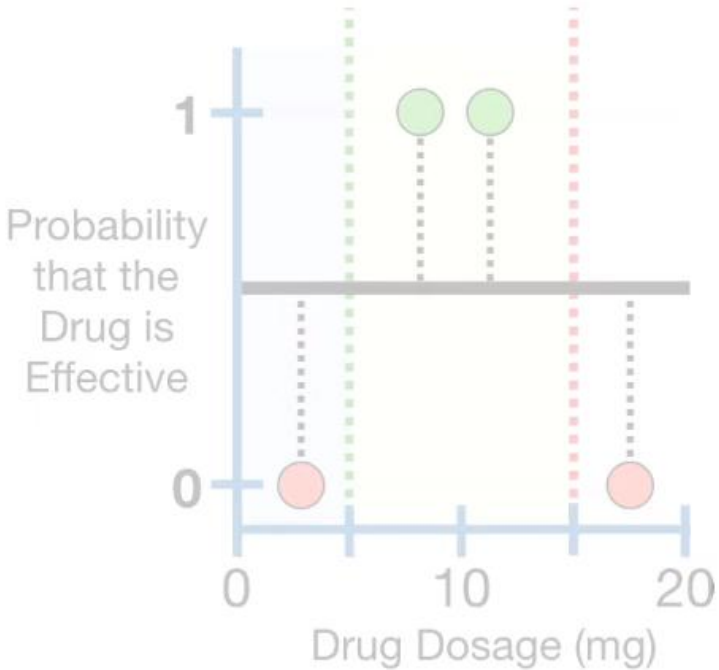


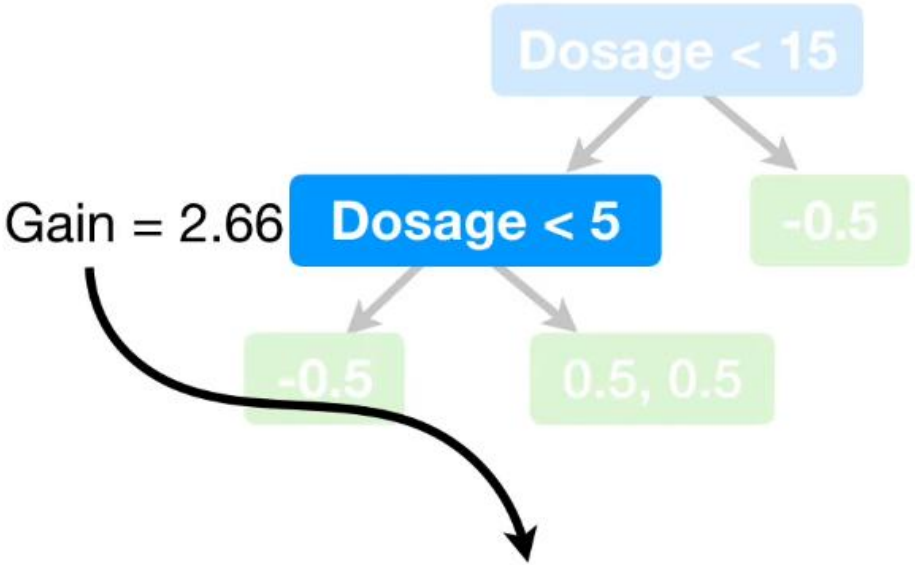
2.66 - γ =

In contrast, if we set $\gamma = 3$…

Predicted Drug Effectiveness

0.5



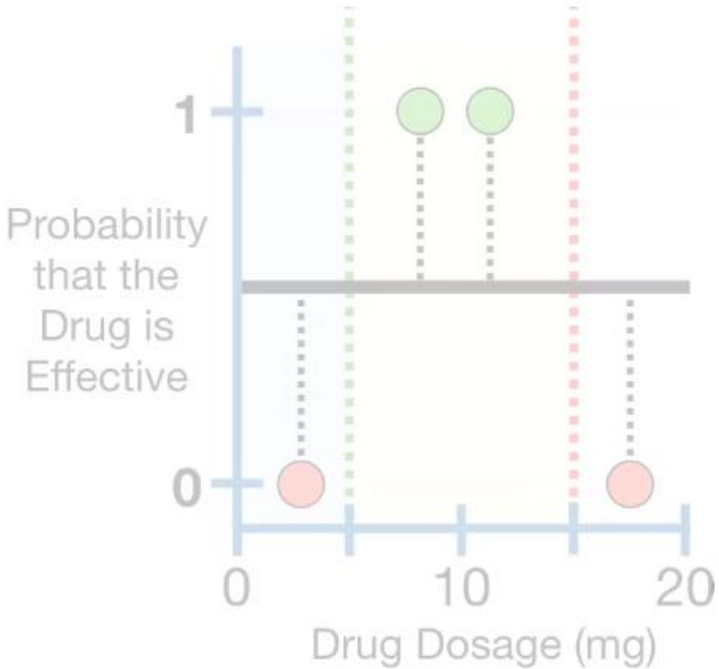


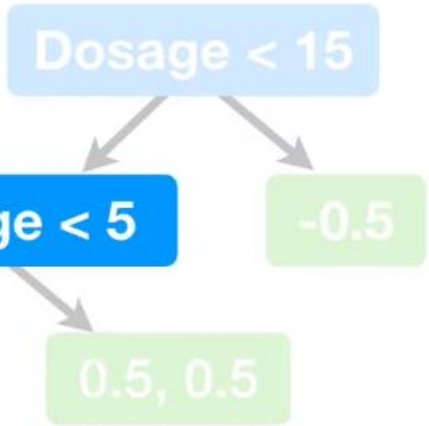
2.66 - 3 =

In contrast, if we set $\gamma = 3$...

Predicted Drug Effectiveness

0.5





2.66 - 3 = -0.34

In contrast, if we set ***γ* = 3**…

…then we would prune because the difference is a ***negative*** number.

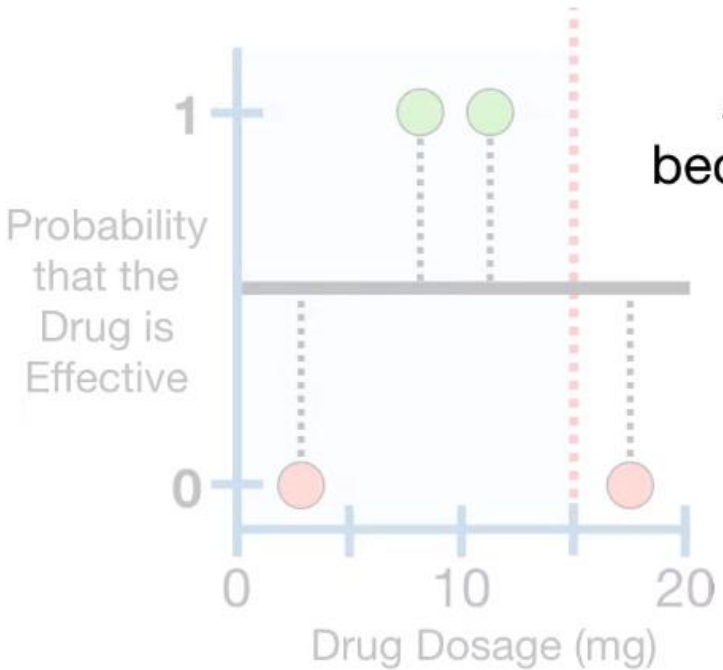
Predicted Drug Effectiveness

0.5

Gain = 1.33 **Dosage < 15**

-0.5, 0.5, 0.5

-0.5



And we would also prune this branch, because…

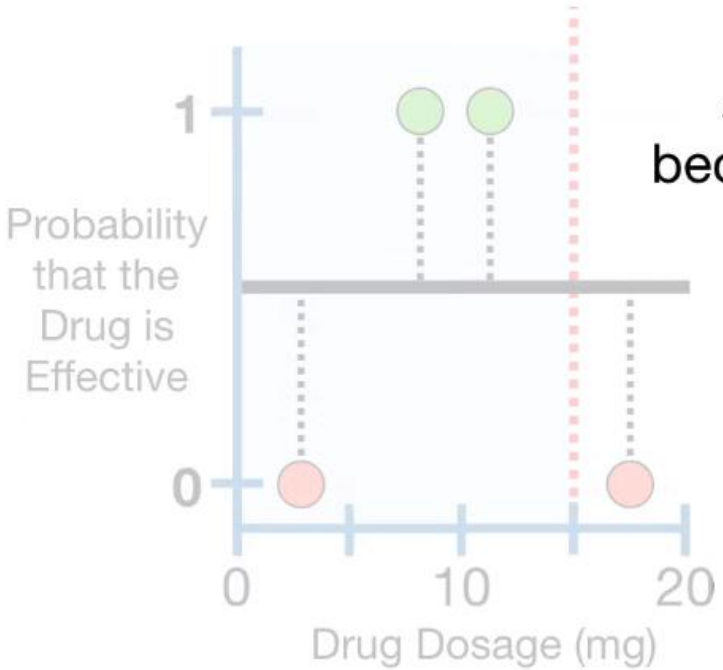
Predicted Drug Effectiveness

0.5

Gain = 1.33 **Dosage < 15**

-0.5, 0.5, 0.5

-0.5



And we would also prune this branch, because… **1.33 - 3 = a negative number**…

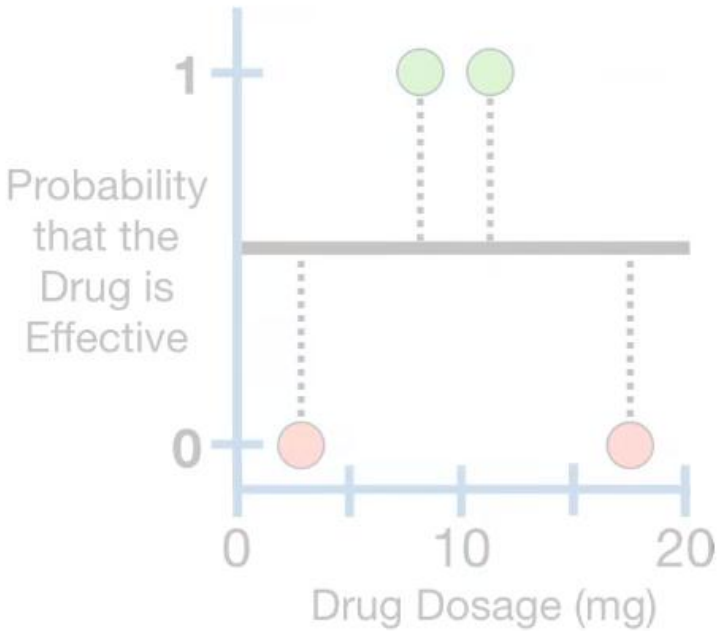
Gain

***γ* (gamma)**

Predicted Drug Effectiveness

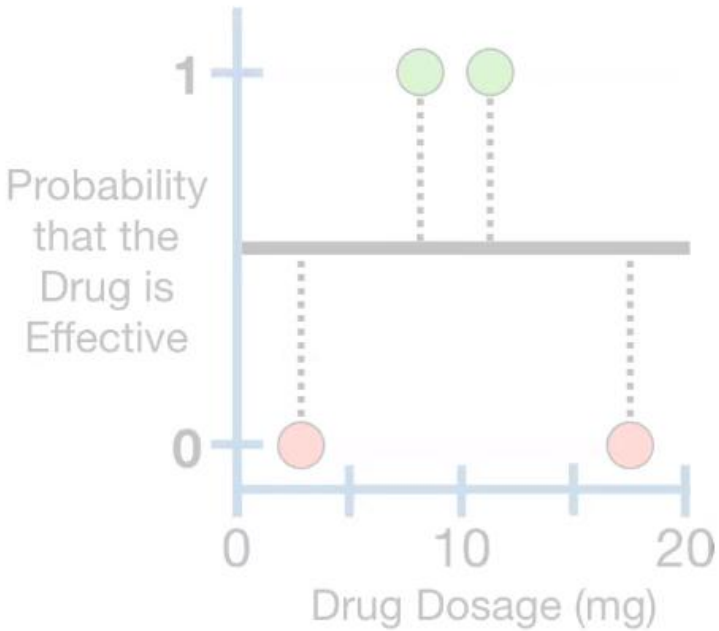
0.5

…and all we would be left with is the original prediction.



Predicted Drug Effectiveness

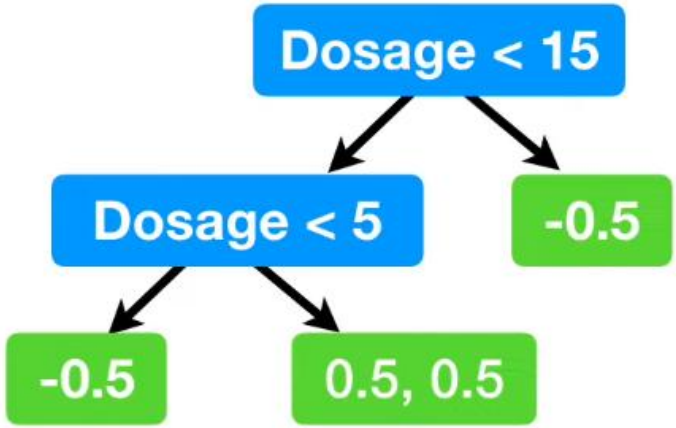
0.5



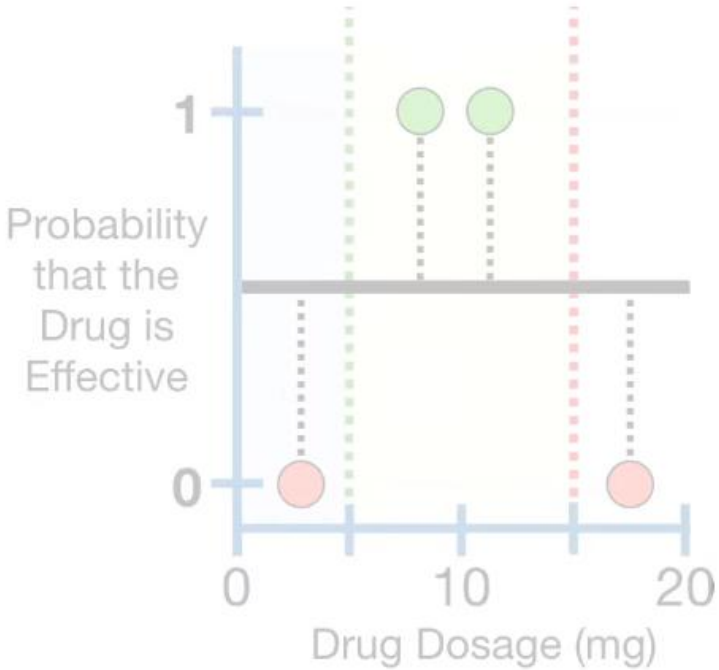
Small Bam.

Predicted Drug Effectiveness

0.5

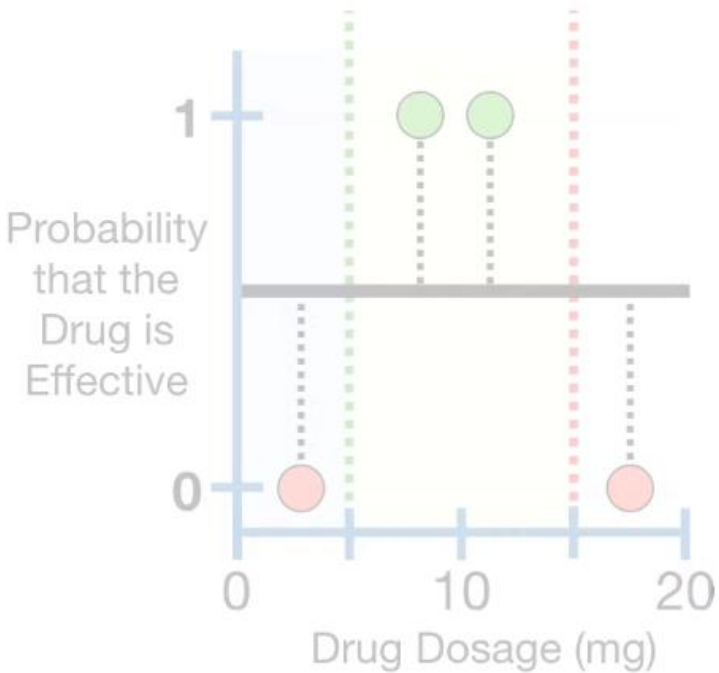


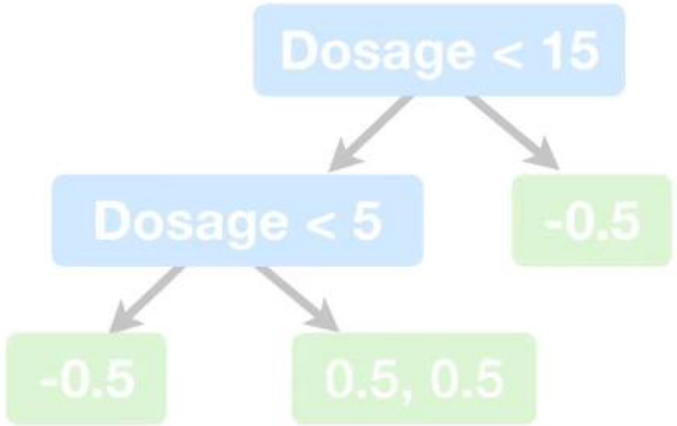
Now, going back to the original tree…



Predicted Drug Effectiveness

0.5



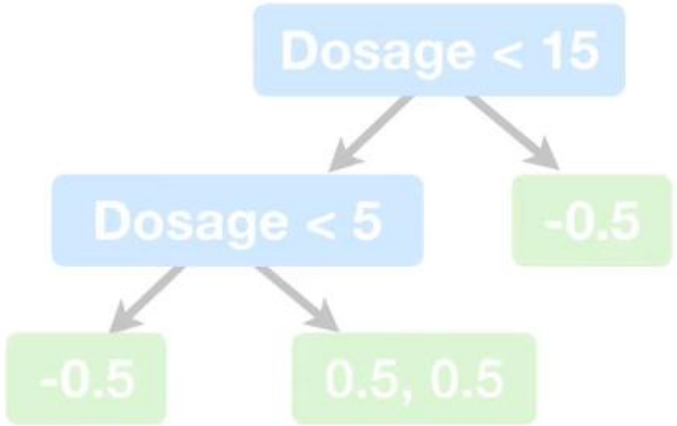


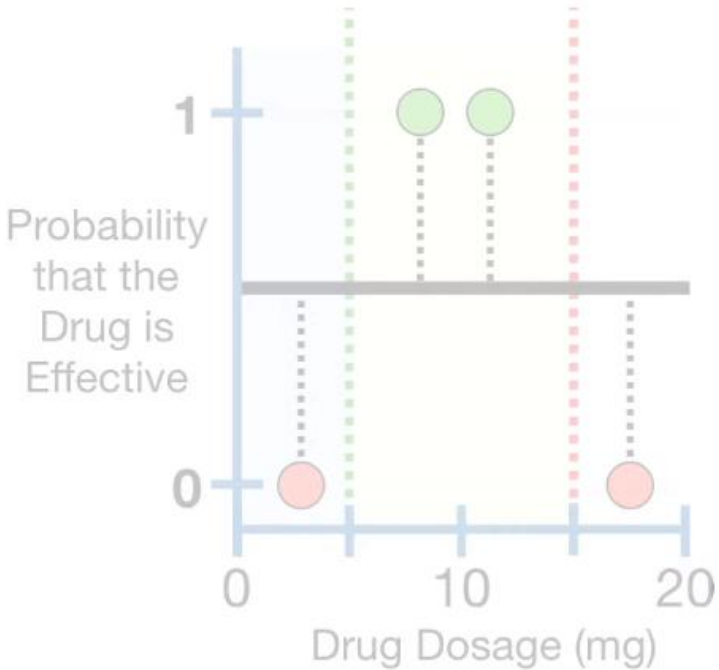
…remember from **Part 1**, that **λ (lambda)**, the **Regularization Parameter**, reduces the **Similarity Scores**…

$$\text{Similarity} = \frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5





…and that lower **Similarity Scores** result in lower values for **Gain**.

$$\text{Similarity} = \frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

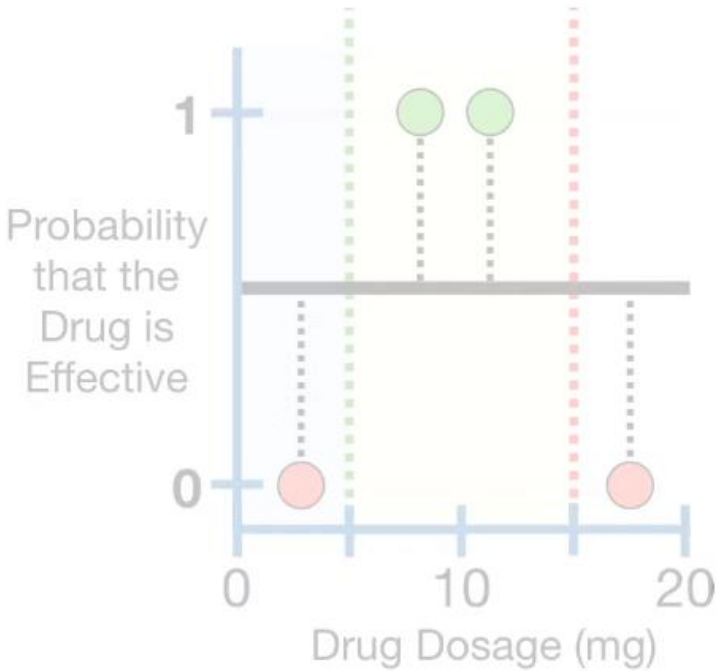
Gain = 0.34 **Dosage < 15**

Gain = 0.72 **Dosage < 5**

-0.5

0.5, 0.5

For example, if we set **λ (lambda) = 1**, then we will get these lower values for **Gain**…



$$\text{Similarity} = \frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

Predicted Drug Effectiveness

0.5

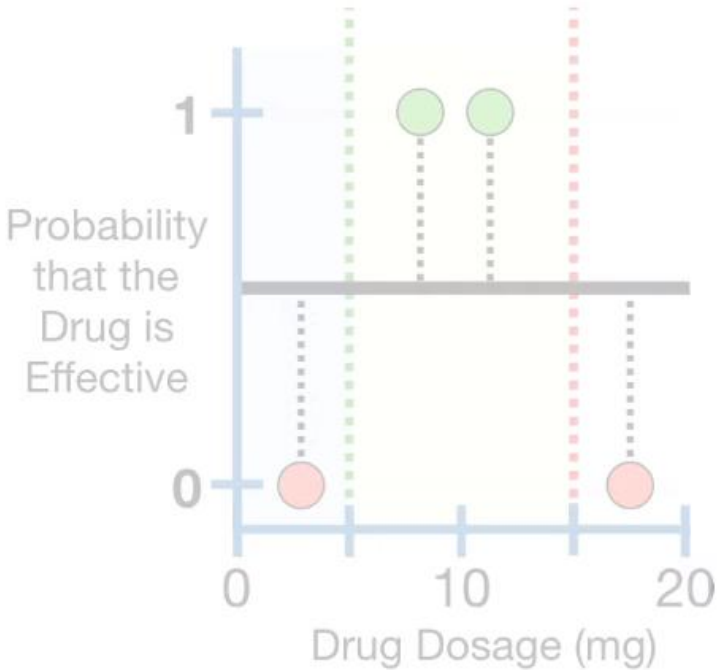
Gain = 0.34 **Dosage < 15**

Gain = 0.72 **Dosage < 5**

-0.5

0.5, 0.5

For example, if we set **λ (lambda) = 1**, then we will get these lower values for **Gain**…



$$\text{Similarity} = \frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

Predicted Drug Effectiveness

0.5

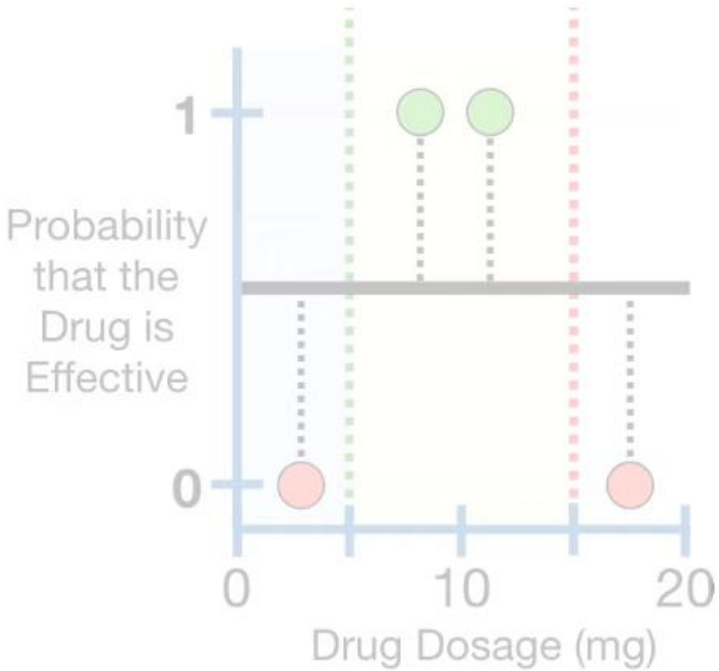
Gain = 0.34 **Dosage < 15**

Gain = 0.72 **Dosage < 5**

-0.5

-0.5

0.5, 0.5



…and that means a lower value for ***γ* (gamma)** will result in a negative difference and cause us to prune branches.

Gain - ***γ*** = { If positive, then do not prune.

If negative, then prune.

Predicted Drug Effectiveness

0.5

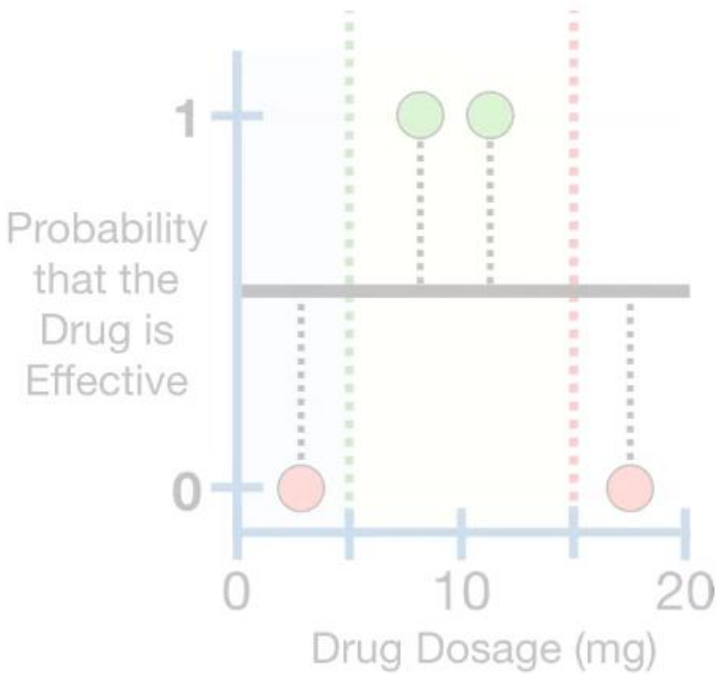
Gain = 0.34 **Dosage < 15**

Gain = 0.72 **Dosage < 5**

-0.5

-0.5

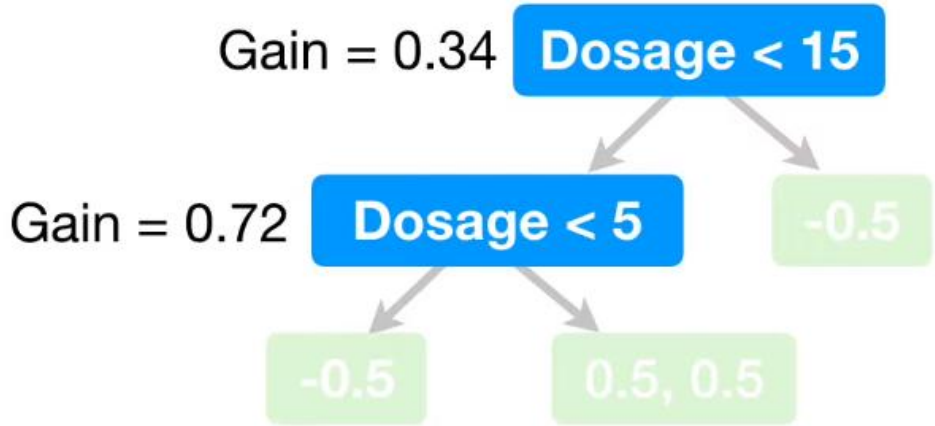
0.5, 0.5

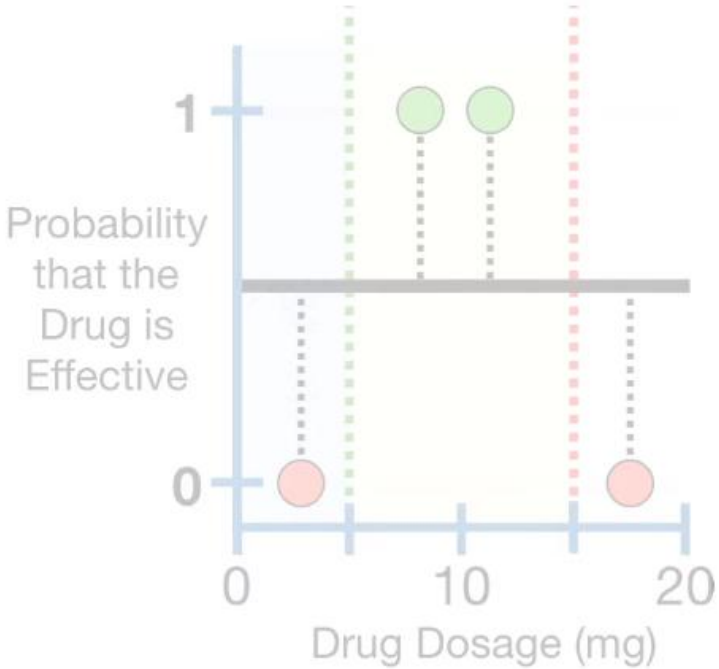


In other words, values for **λ (lambda)** greater than **0** reduce the sensitivity of the tree to individual observations by pruning and combining them with other observations.

Predicted Drug Effectiveness

0.5

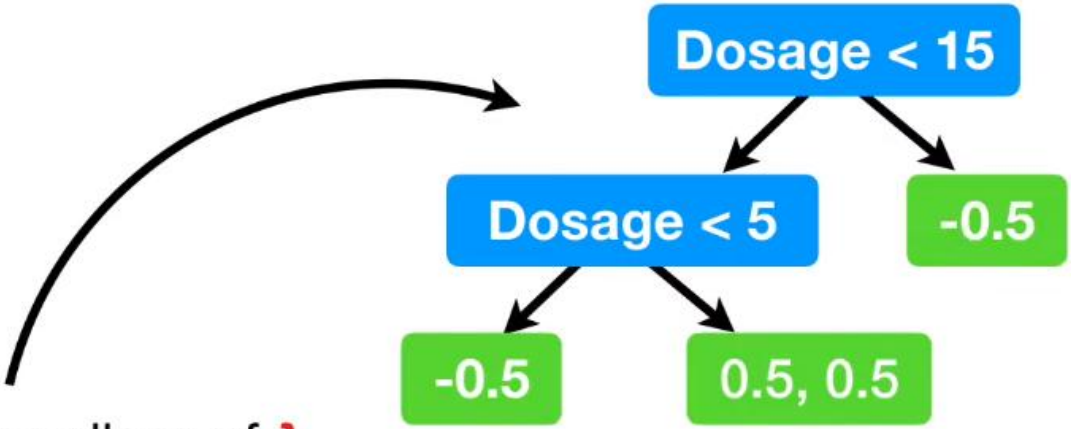




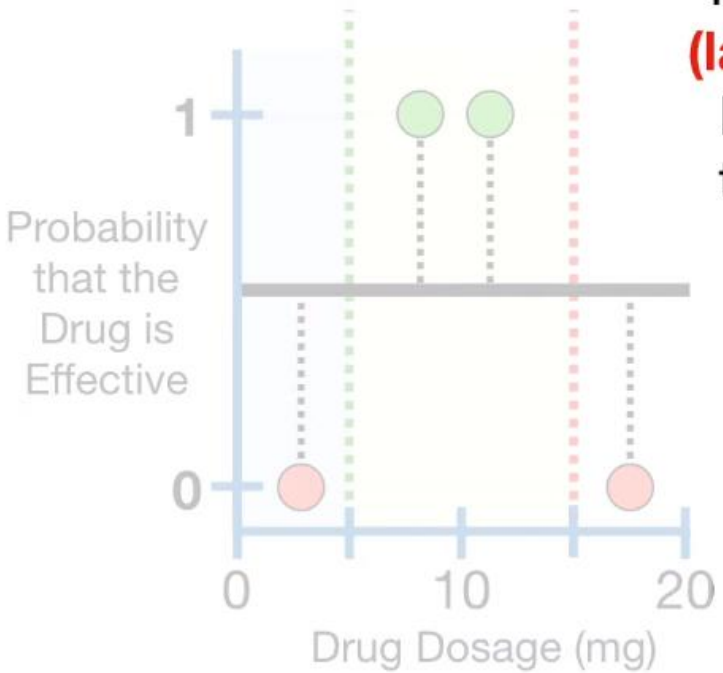
BAM!!!

Predicted Drug Effectiveness

0.5

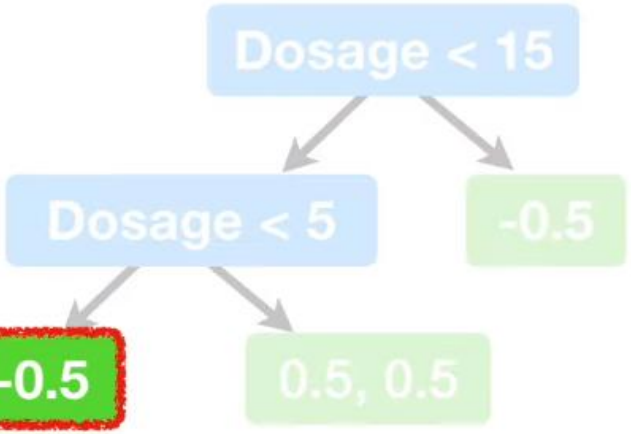


For now, regardless of **λ (lambda)** and **γ (gamma)**, let's assume that this is the tree we are working with…

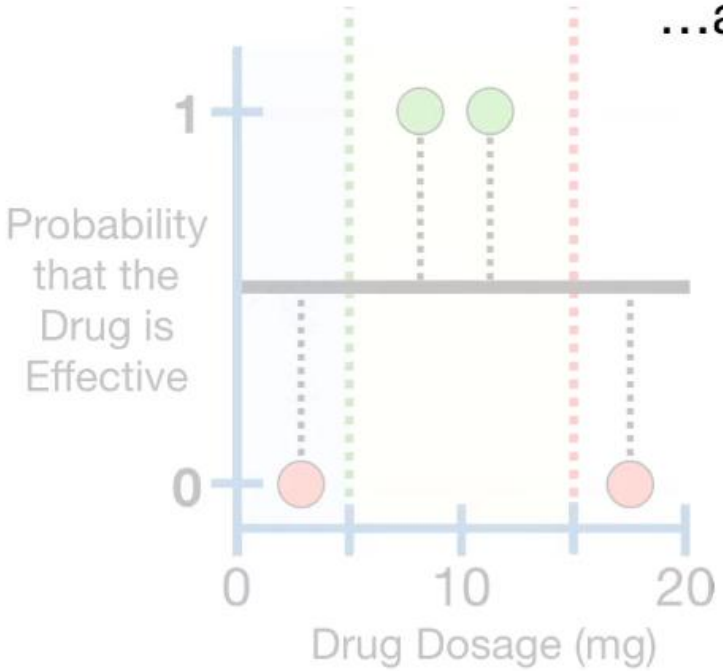


Predicted Drug Effectiveness

0.5

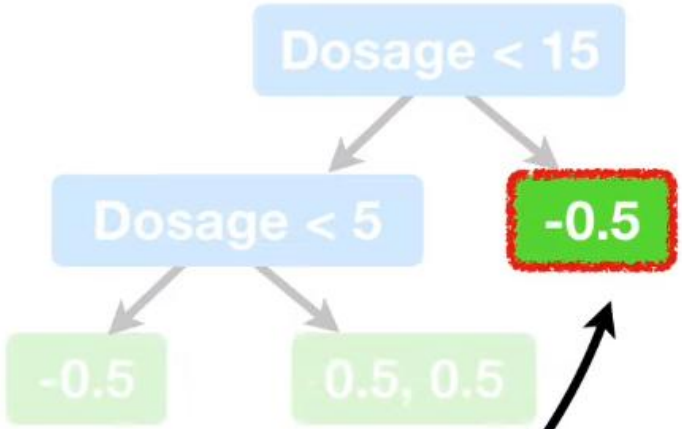


…and determine the **Output Values** for the leaves.

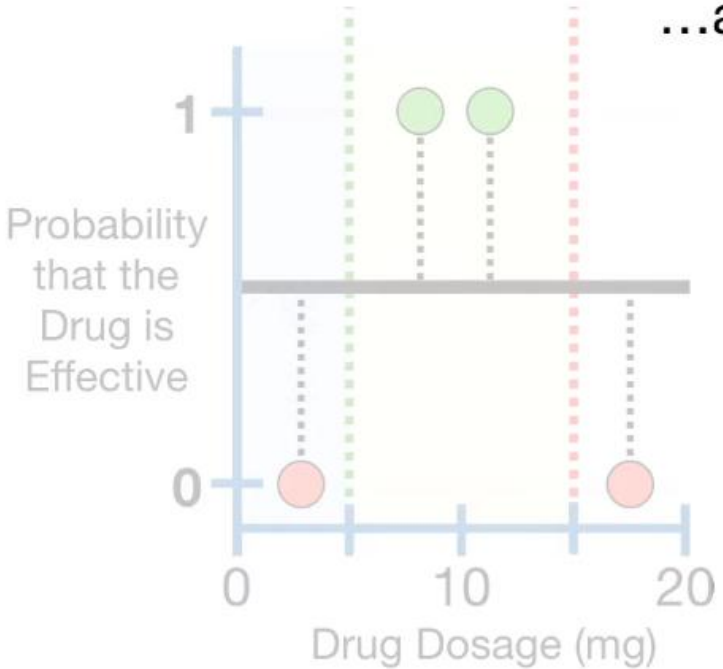


Predicted Drug Effectiveness

0.5

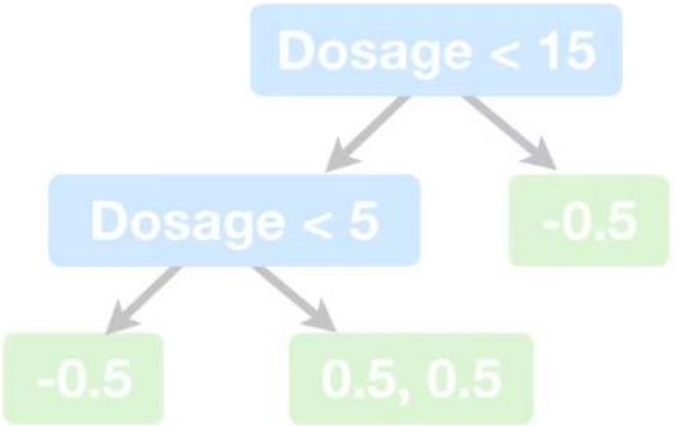


…and determine the **Output Values** for the leaves.

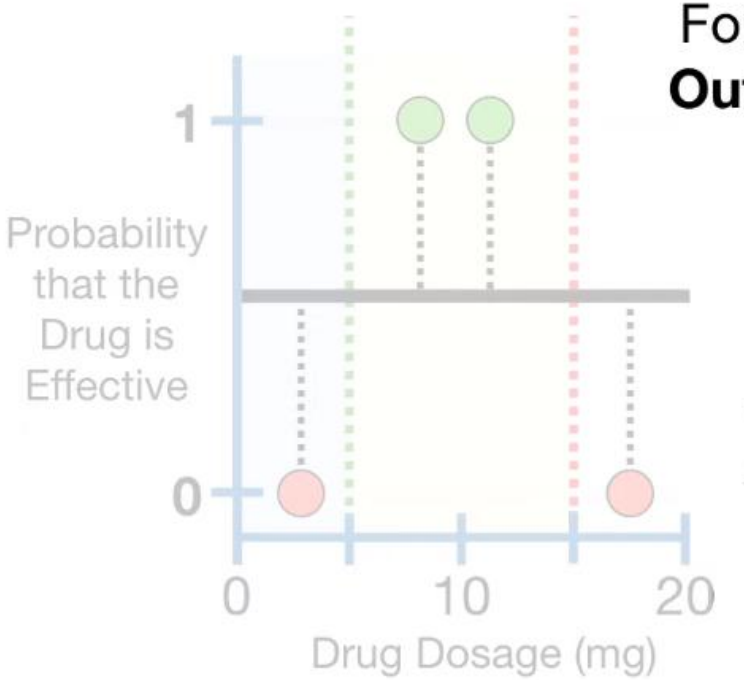


Predicted Drug Effectiveness

0.5



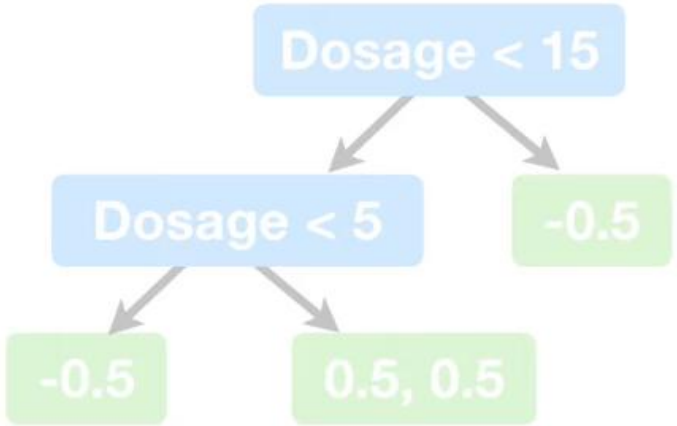
For **Classification**, the the **Output Value** for a leaf is…

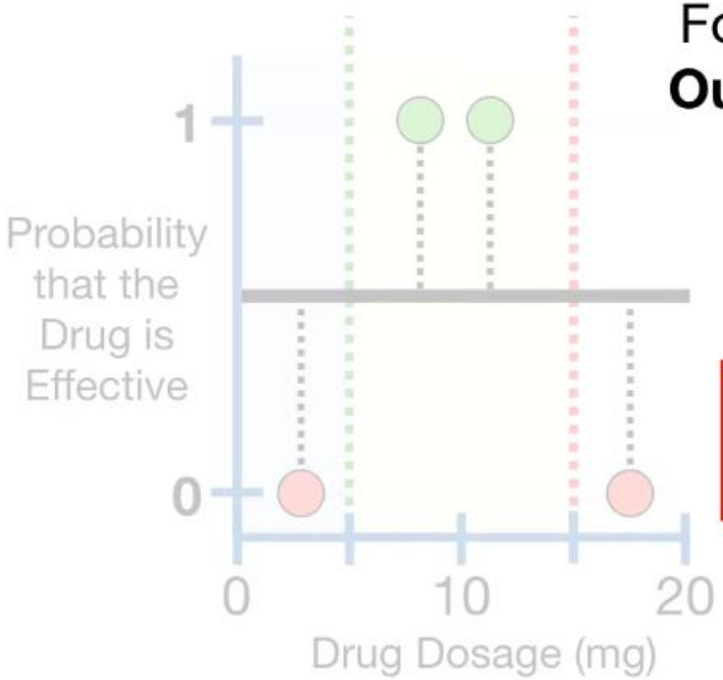


$$\frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5



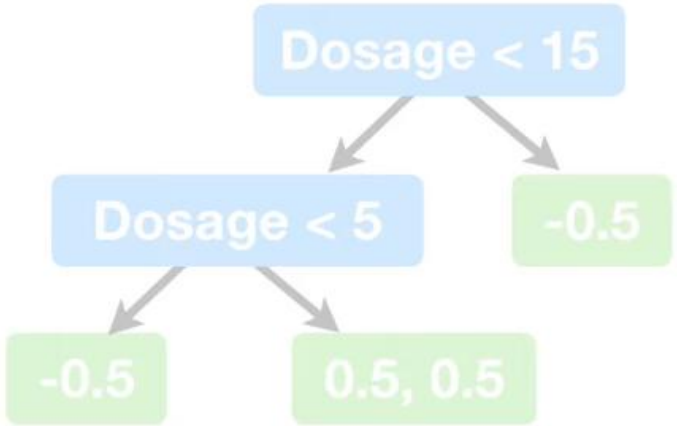


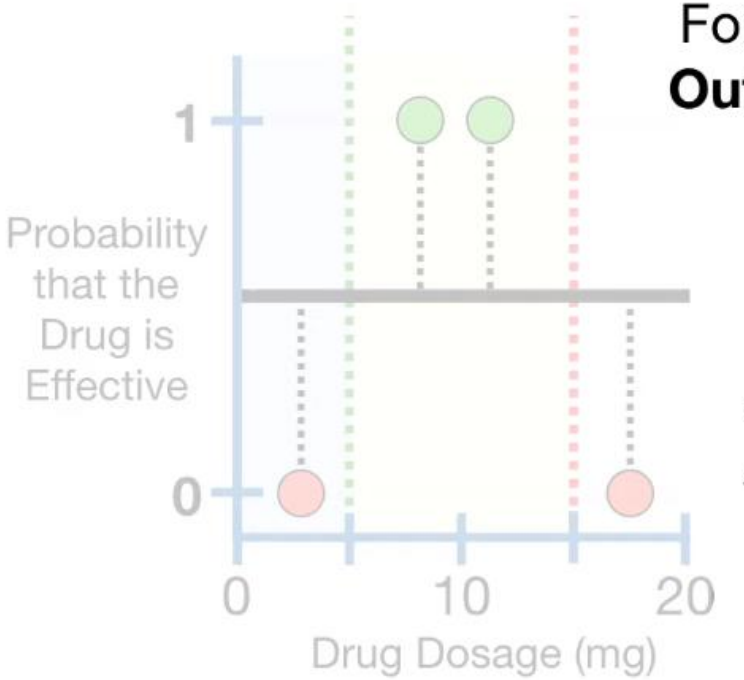
For **Classification**, the the **Output Value** for a leaf is…

$$\frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5



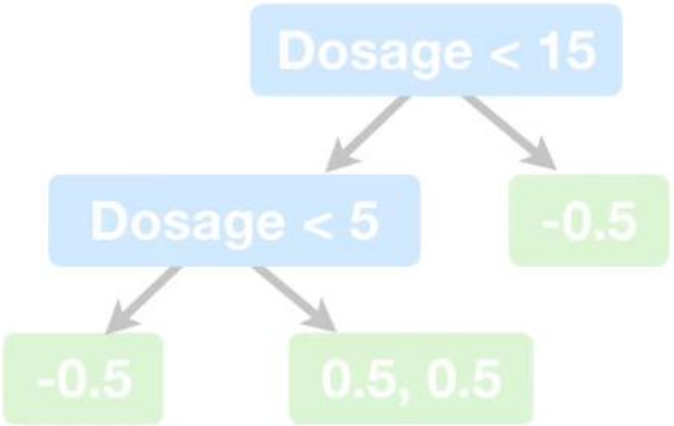


For **Classification**, the the **Output Value** for a leaf is…

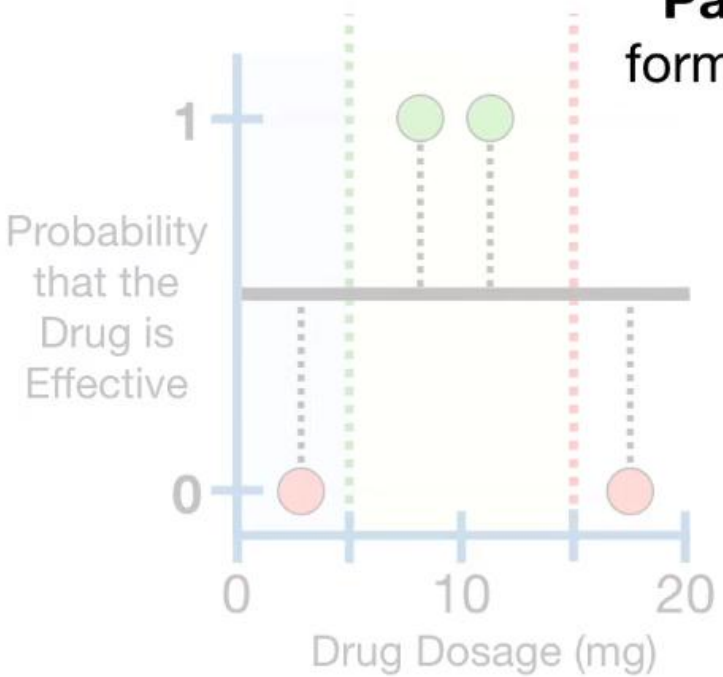
$$\frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

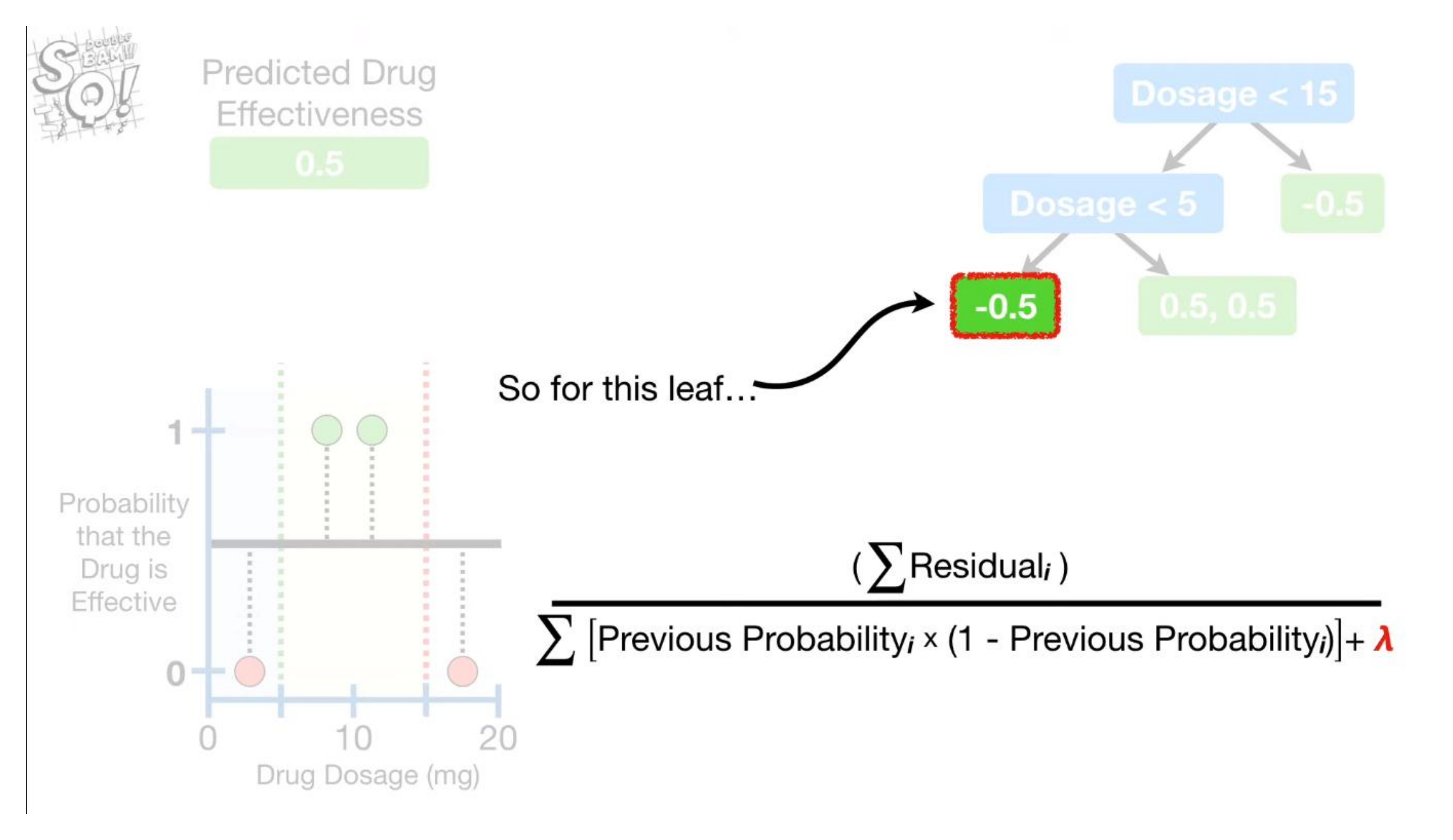
0.5



NOTE: With the exception of λ (lambda), the **Regularization Parameter**, this is the same formula we used for unextreme **Gradient Boost**.

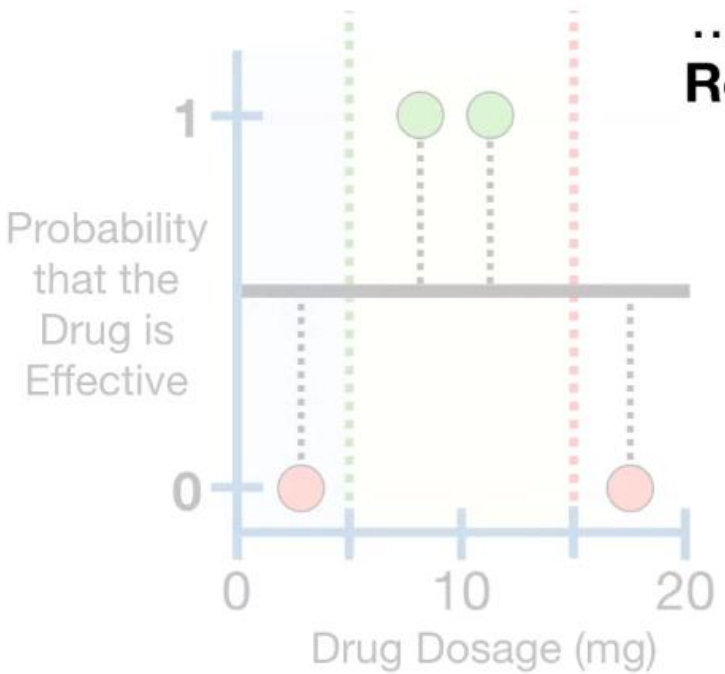


$$\frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

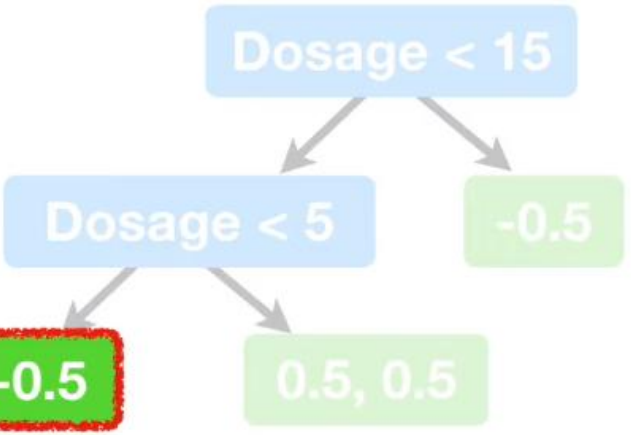

Predicted Drug Effectiveness
0.5
Dosage < 15
Dosage < 5
-0.5
-0.5
0.5, 0.5
So for this leaf…
1
0
Probability that the Drug is Effective
0
10
20
Drug Dosage (mg)
$\frac{(\sum \text{Residual}_i)}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$

Predicted Drug Effectiveness

0.5



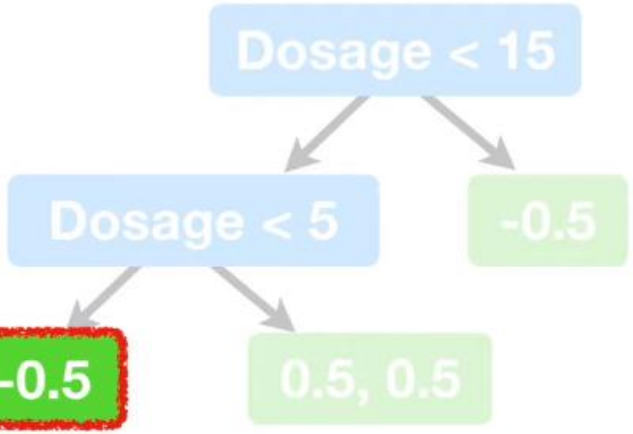
…we plug in the **Residual**, **-0.5**…



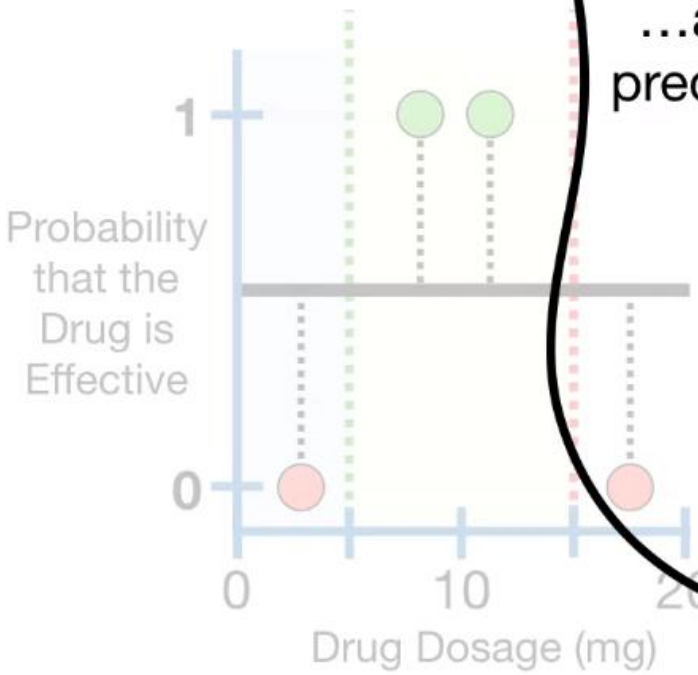
$$\frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5



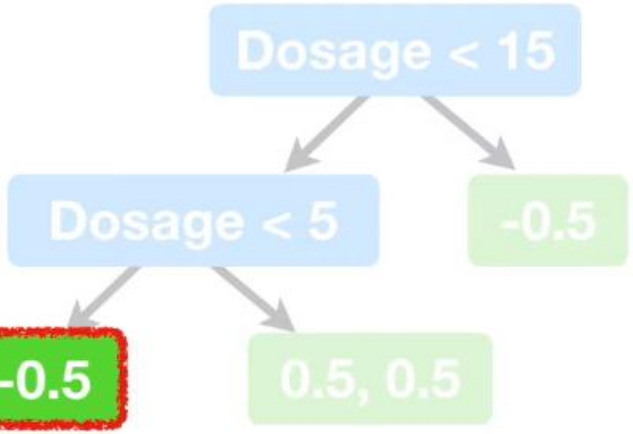
…and the previously predicted probability…



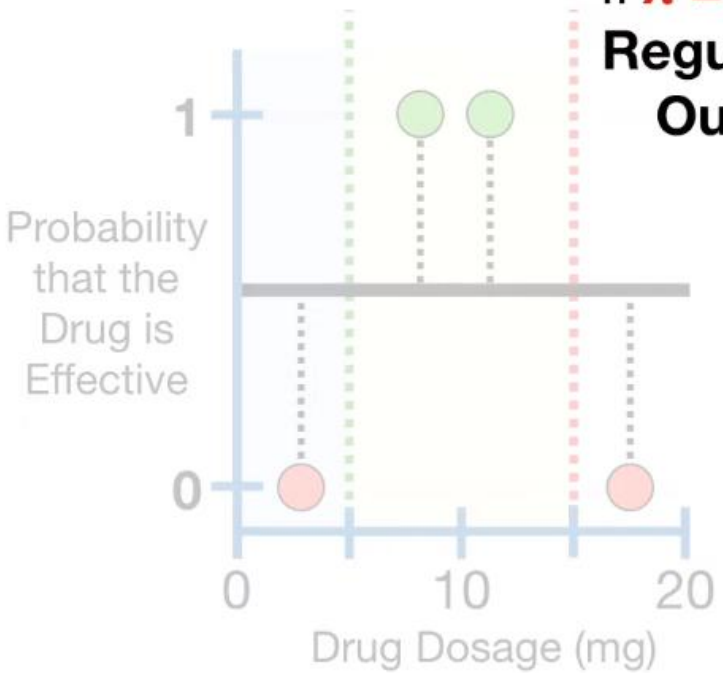
$$\frac{-0.5}{0.5 \times (1 - 0.5) + \lambda}$$

Predicted Drug Effectiveness

0.5



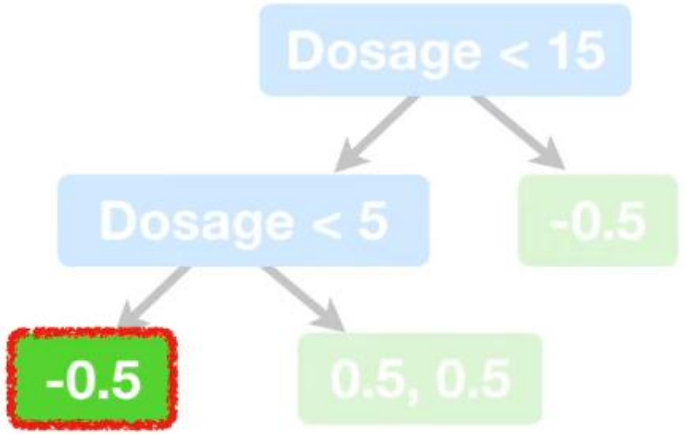
If $\lambda = 0$, then there is no **Regularization** and the **Output Value = -2**.

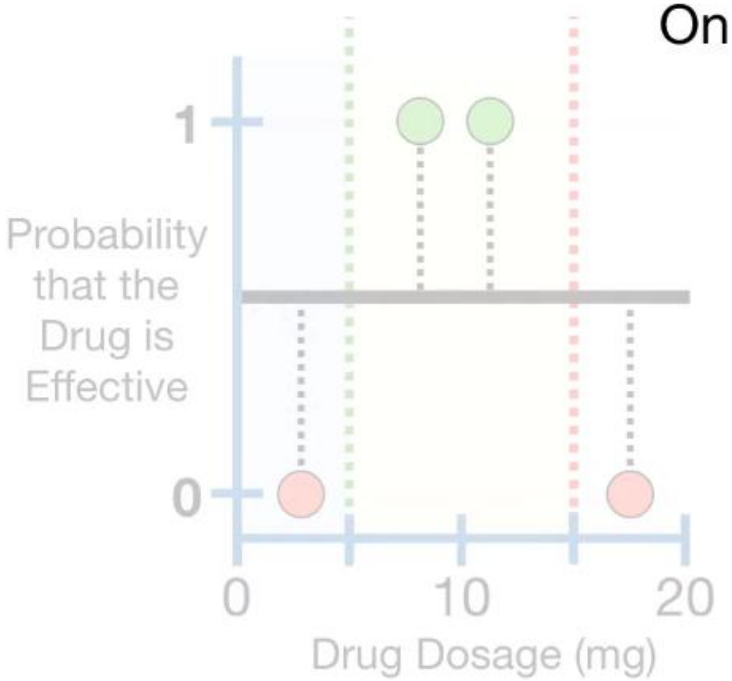


$$\frac{-0.5}{0.5 \times (1 - 0.5) + 0} = -2$$

Predicted Drug Effectiveness

0.5



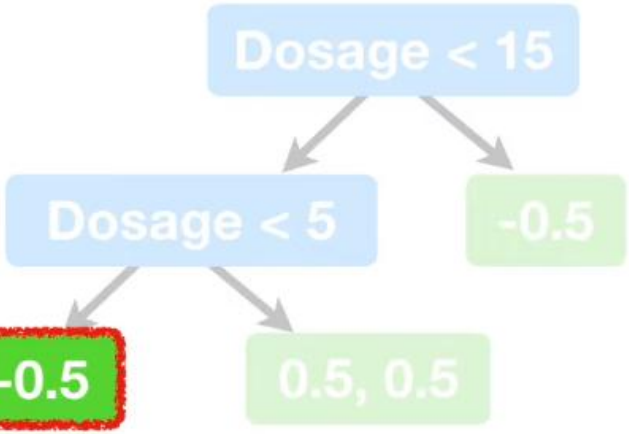


On the other hand, if $\lambda = 1$...

$$\frac{-0.5}{0.5 \times (1 - 0.5) + 1}$$

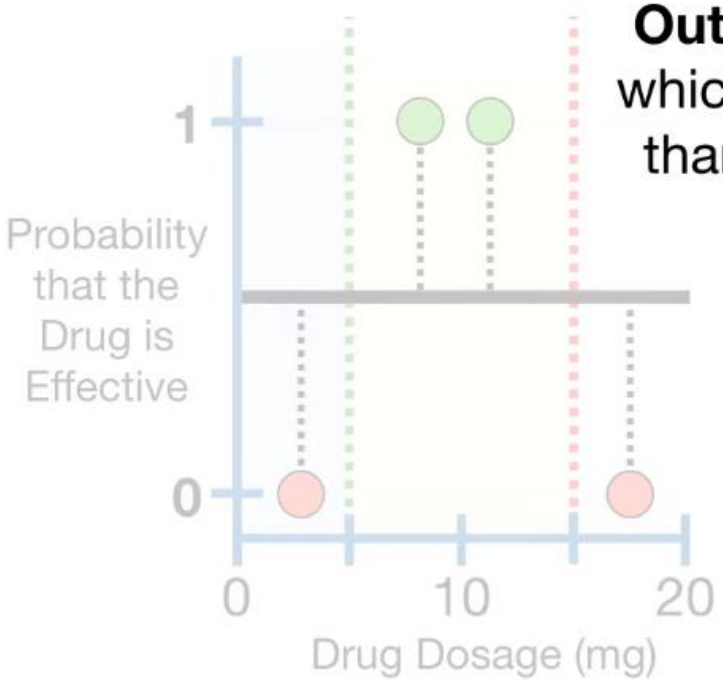
Predicted Drug Effectiveness

0.5



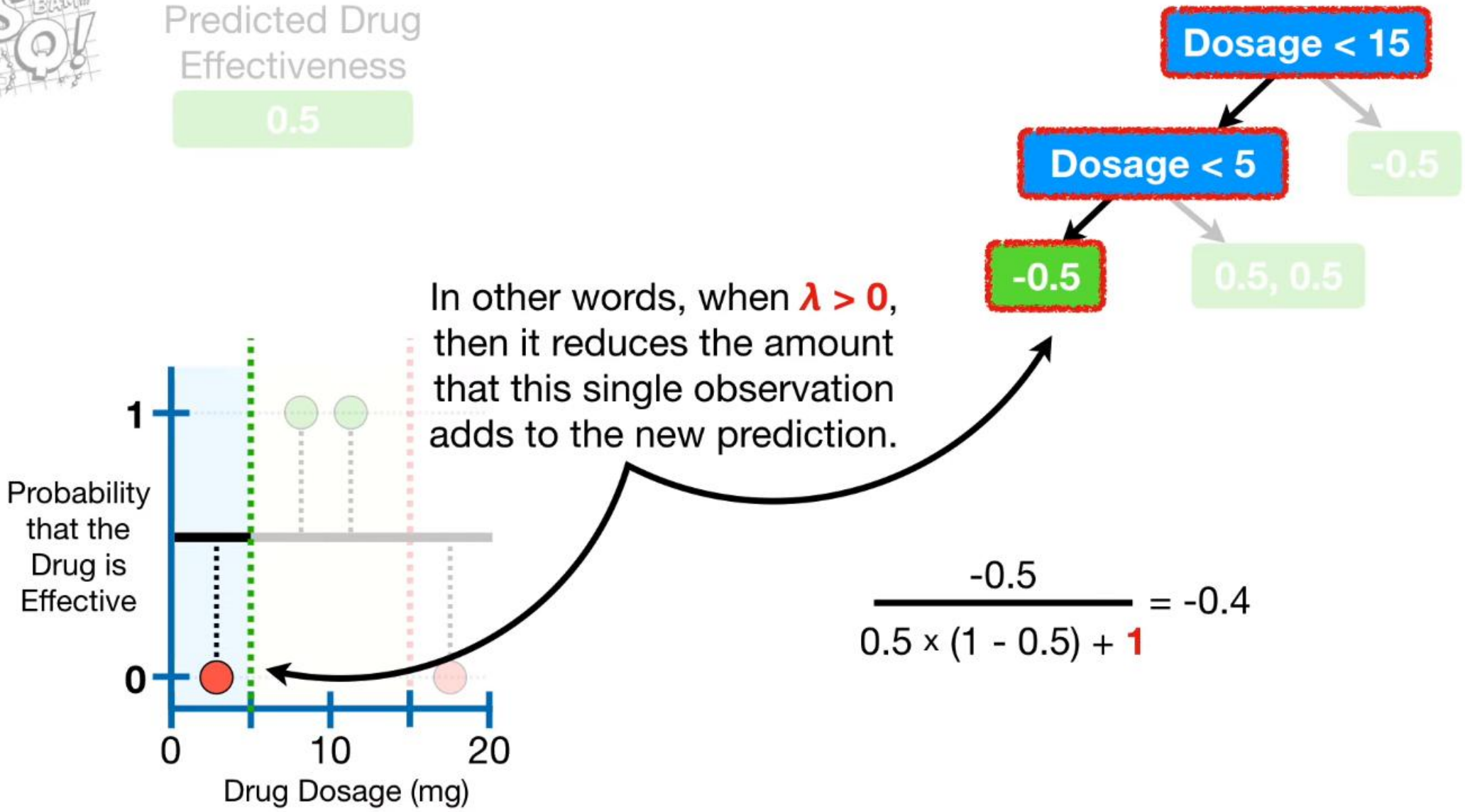
…then the **Output Value = -0.4**, which is closer to zero than **-2**, when $\lambda = 0$.

$$\frac{-0.5}{0.5 \times (1 - 0.5) + 1} = -0.4$$



Predicted Drug Effectiveness

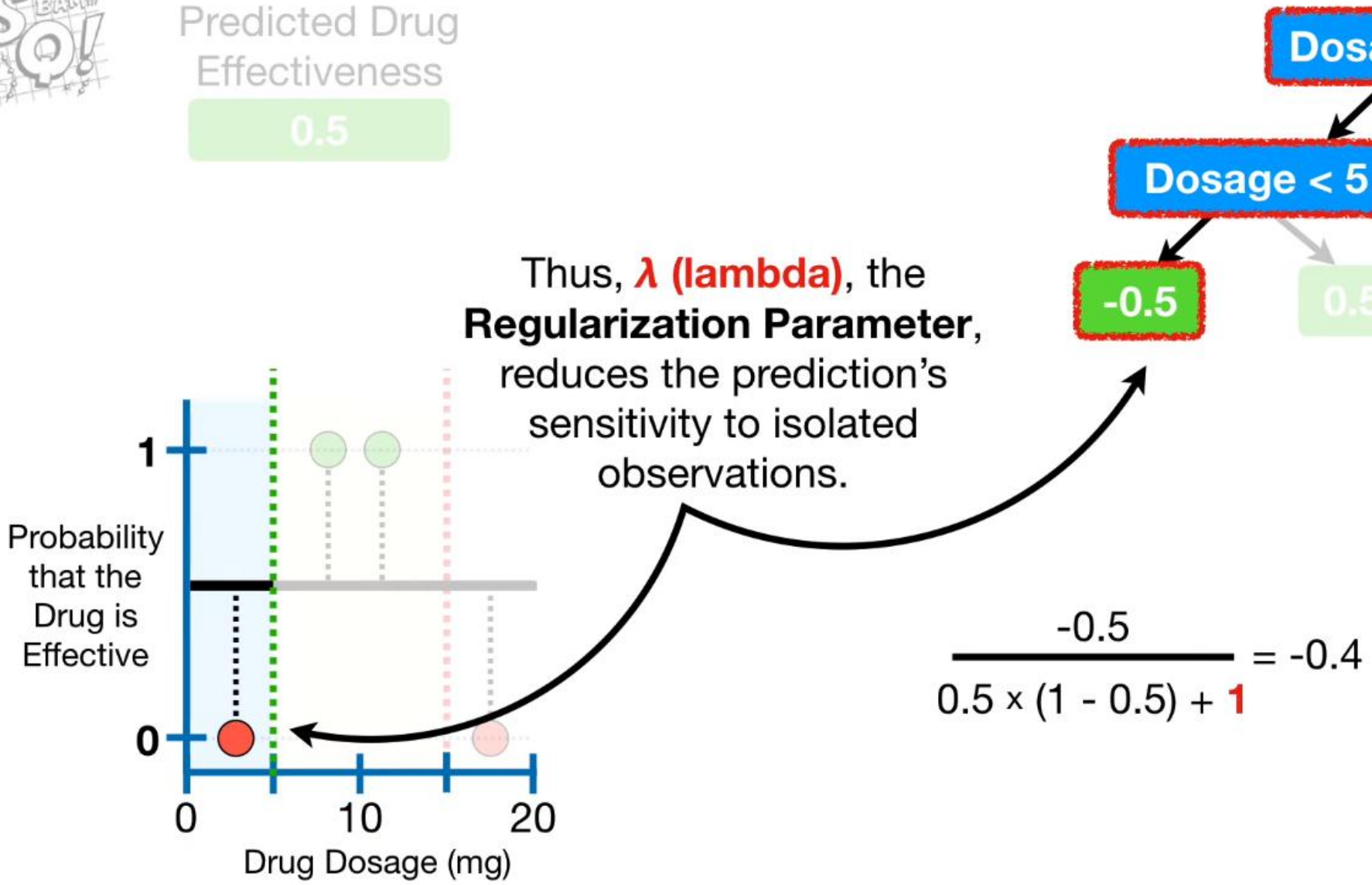
0.5



$$\frac{-0.5}{0.5 \times (1 - 0.5) + 1} = -0.4$$

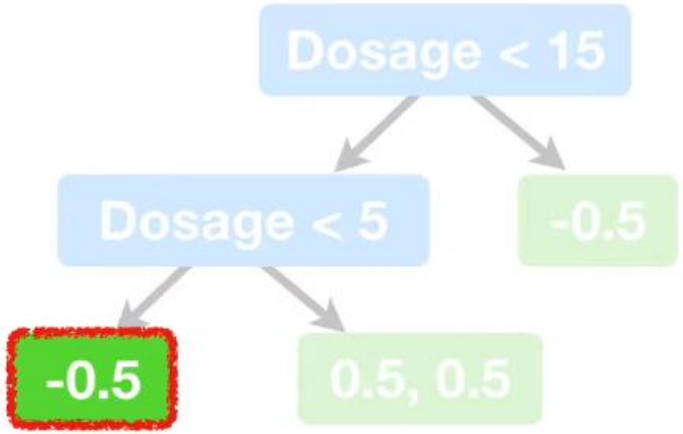
Predicted Drug Effectiveness

0.5

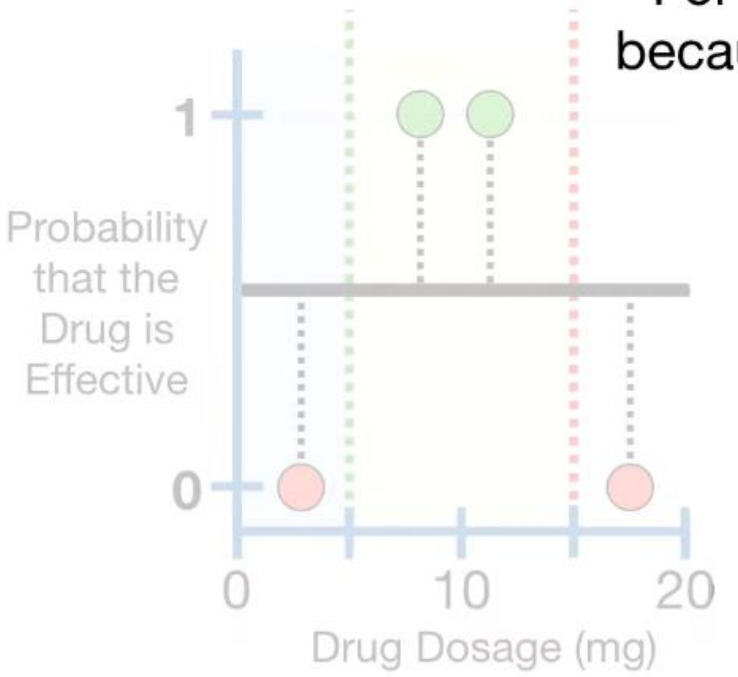


Predicted Drug Effectiveness

0.5



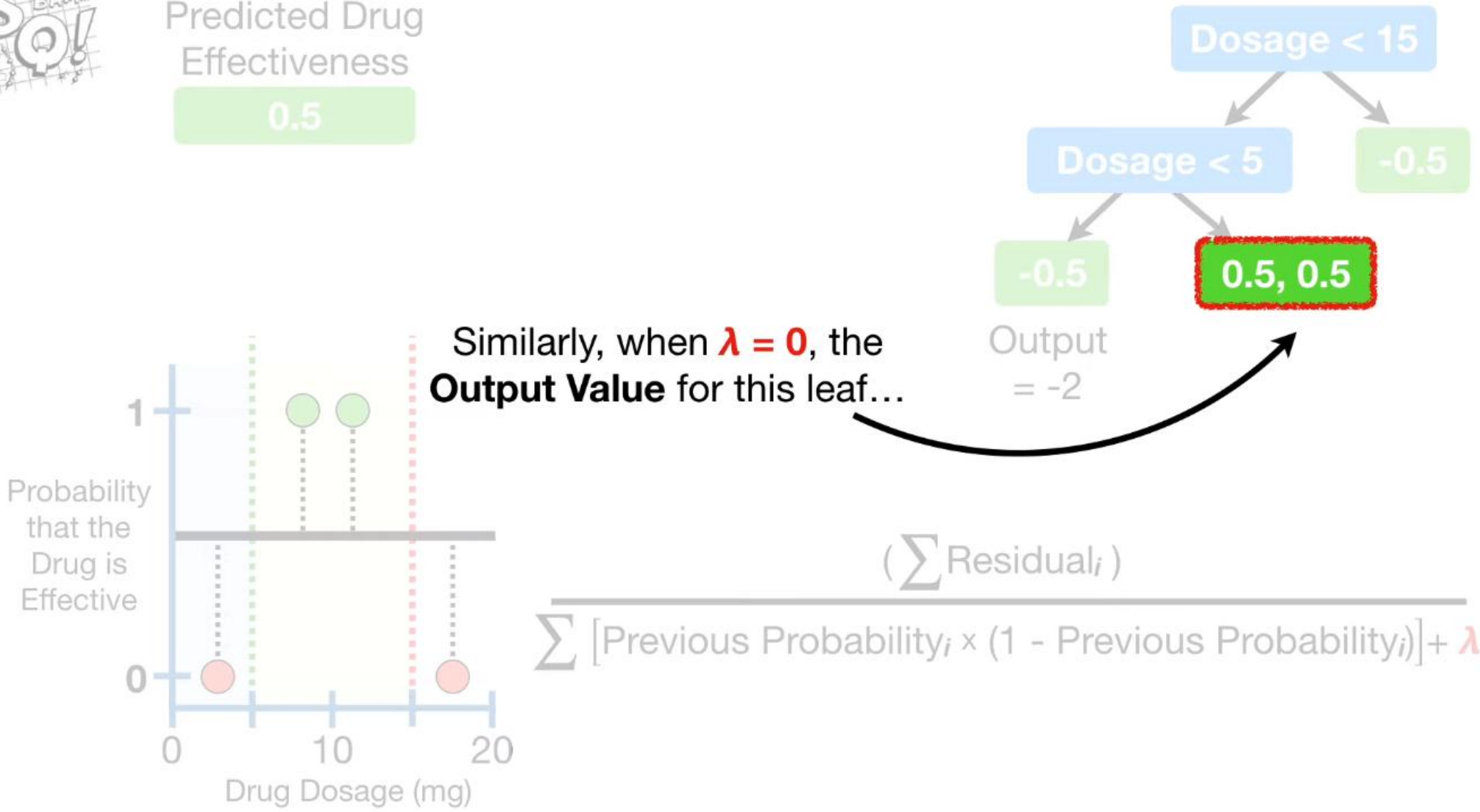
For now, we'll let $\lambda = 0$, because this is the default value…



$$\frac{-0.5}{0.5 \times (1 - 0.5) + 0} = -2$$

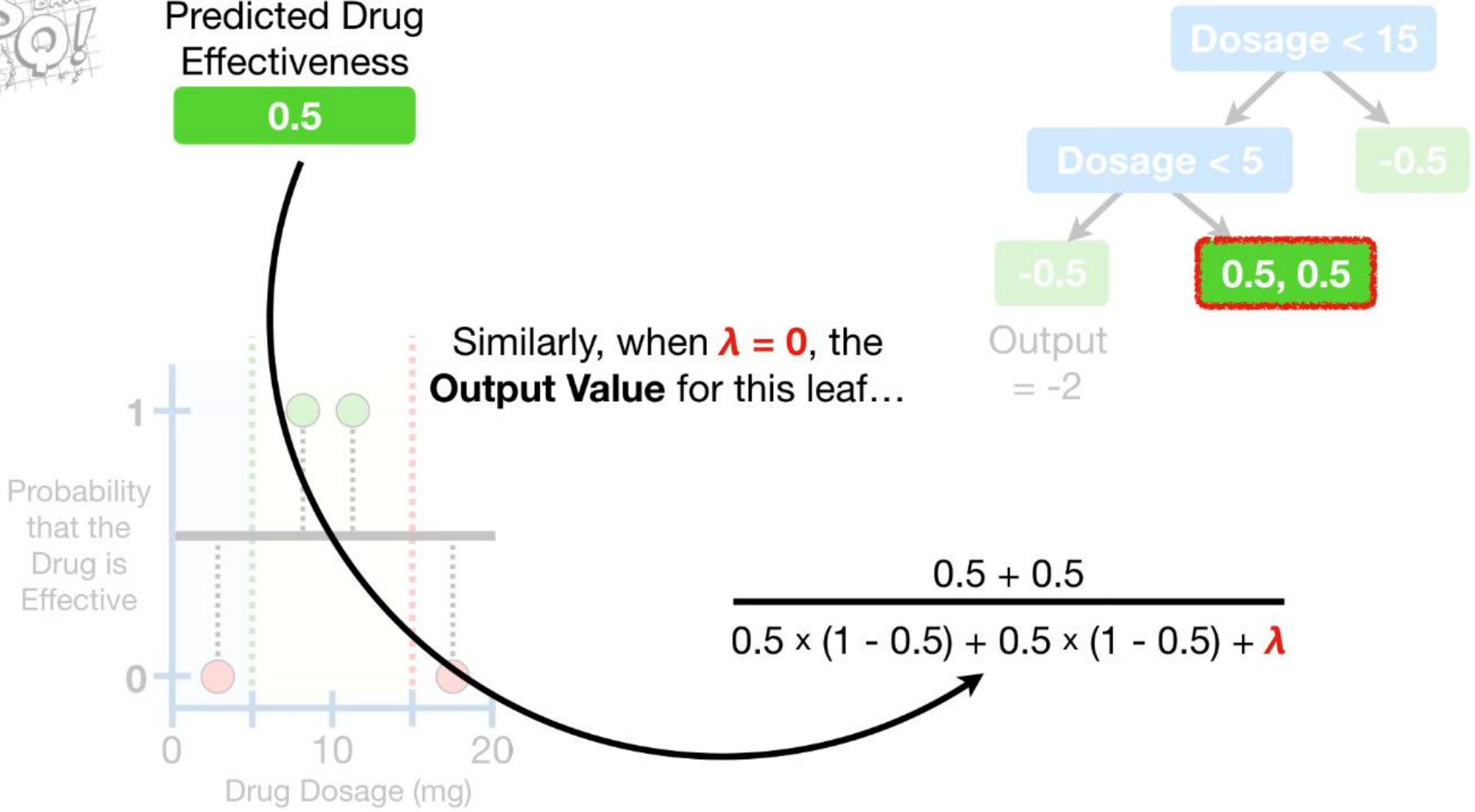
Predicted Drug Effectiveness

0.5



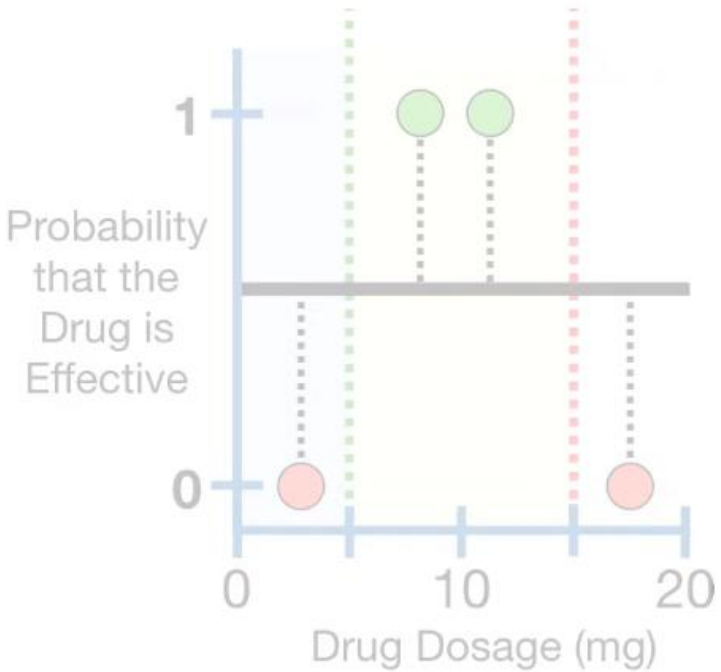
Predicted Drug Effectiveness

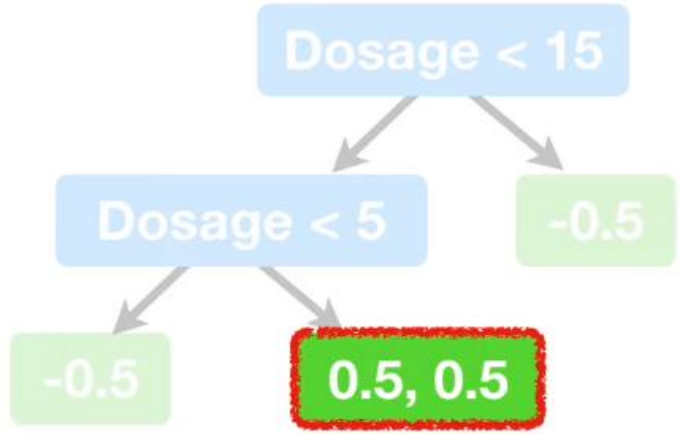
0.5



Predicted Drug Effectiveness

0.5





Output = -2

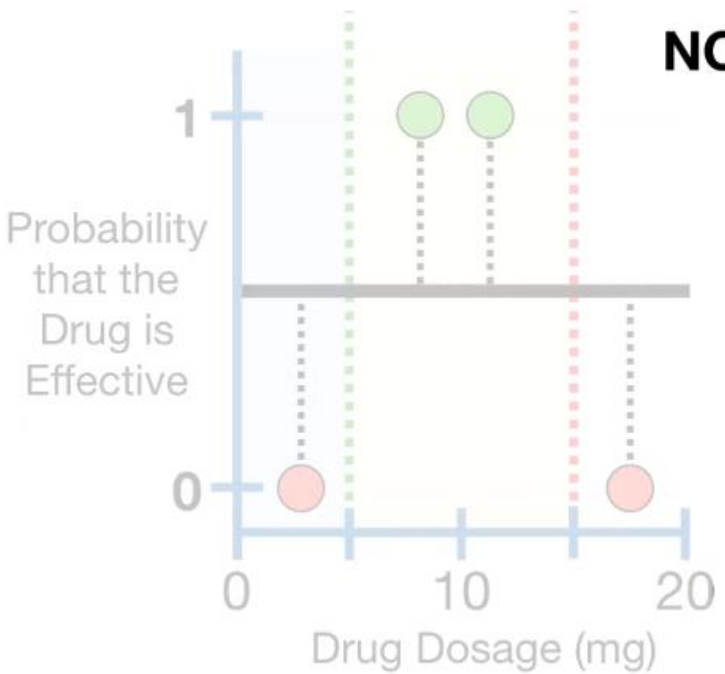
…is **2**.

$$\frac{0.5 + 0.5}{0.5 \times (1 - 0.5) + 0.5 \times (1 - 0.5) + \mathbf{0}} = 2$$

Predicted Drug Effectiveness

0.5

Output = -2

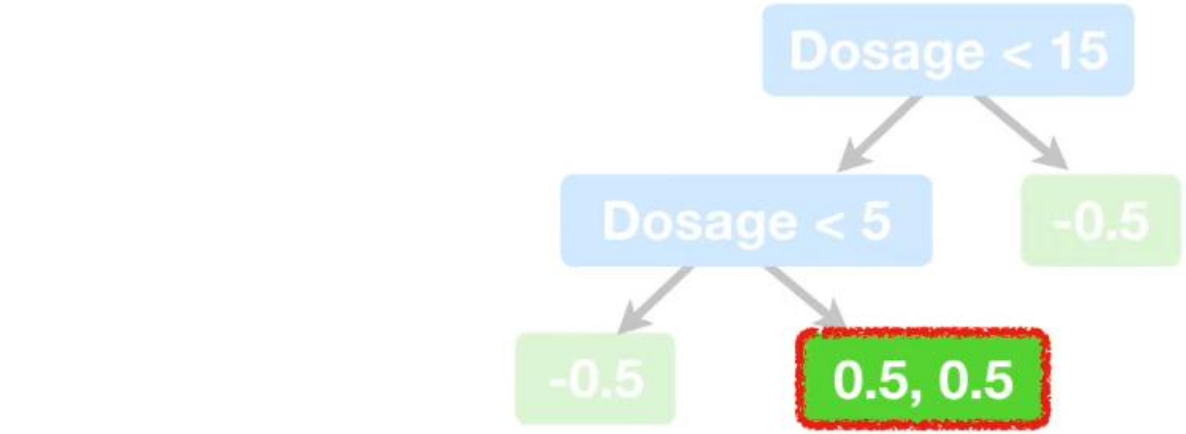


NOTE: If $\lambda = 1$, then…

$$\frac{0.5 + 0.5}{0.5 \times (1 - 0.5) + 0.5 \times (1 - 0.5) + 1}$$

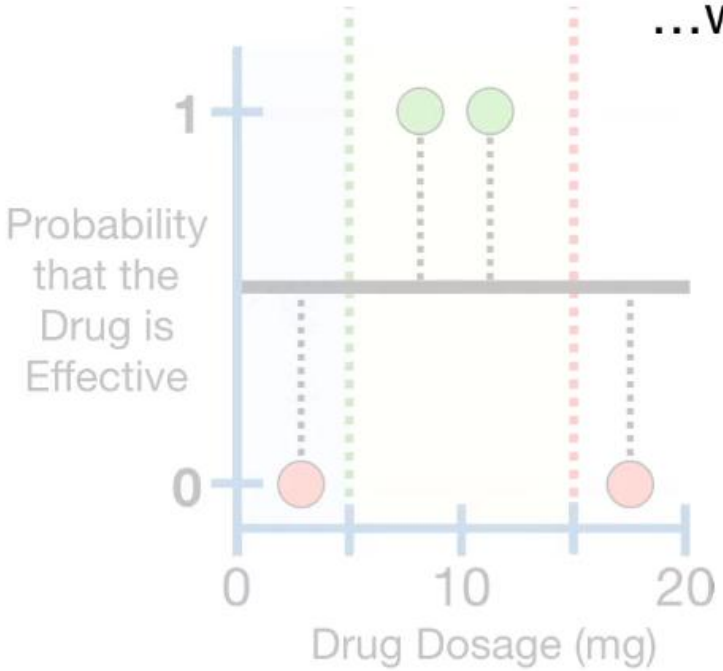
Predicted Drug Effectiveness

0.5



Output = -2

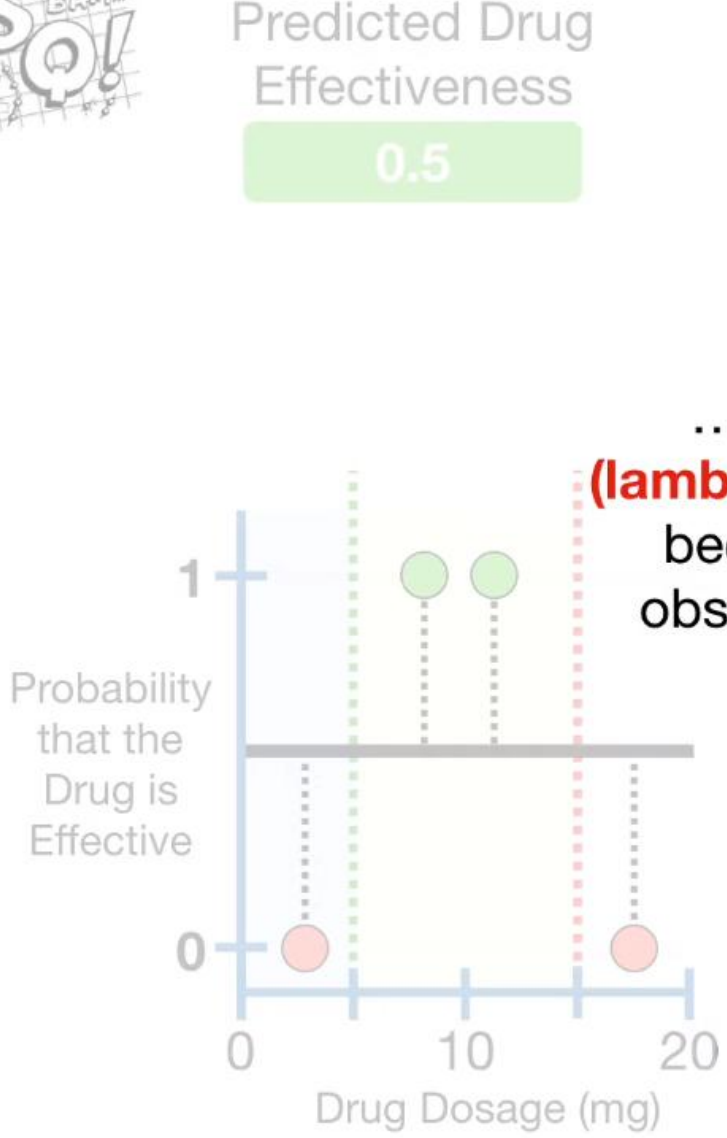
…we get **0.67**, which is closer to **0**…

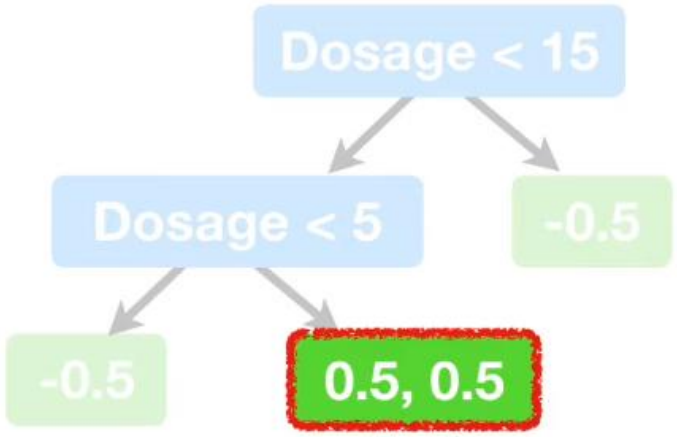


$$\frac{0.5 + 0.5}{0.5 \times (1 - 0.5) + 0.5 \times (1 - 0.5) + 1} = 0.67$$

Predicted Drug Effectiveness

0.5



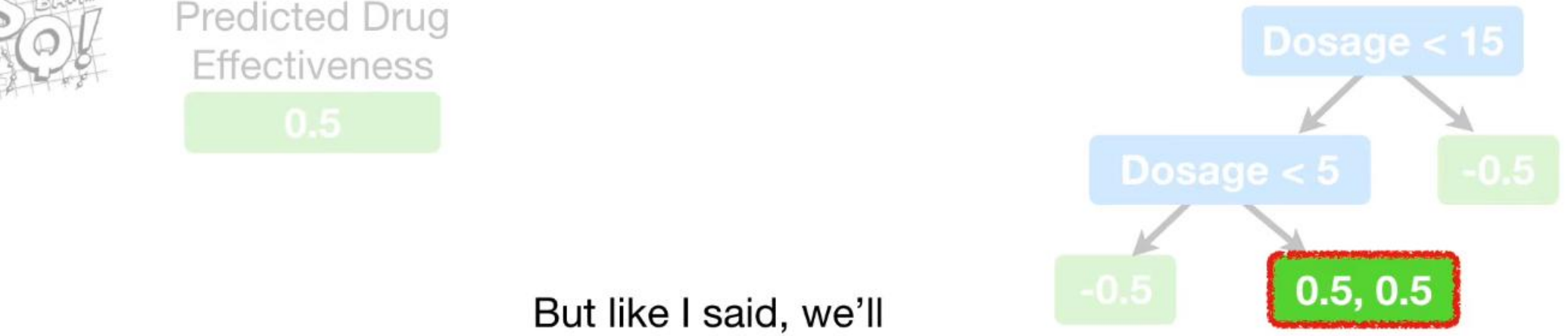


…but the effect of **λ (lambda)** is smaller this time because there are two observations in this leaf.

$$\frac{0.5 + 0.5}{0.5 \times (1 - 0.5) + 0.5 \times (1 - 0.5) + \mathbf{1}} = 0.67$$

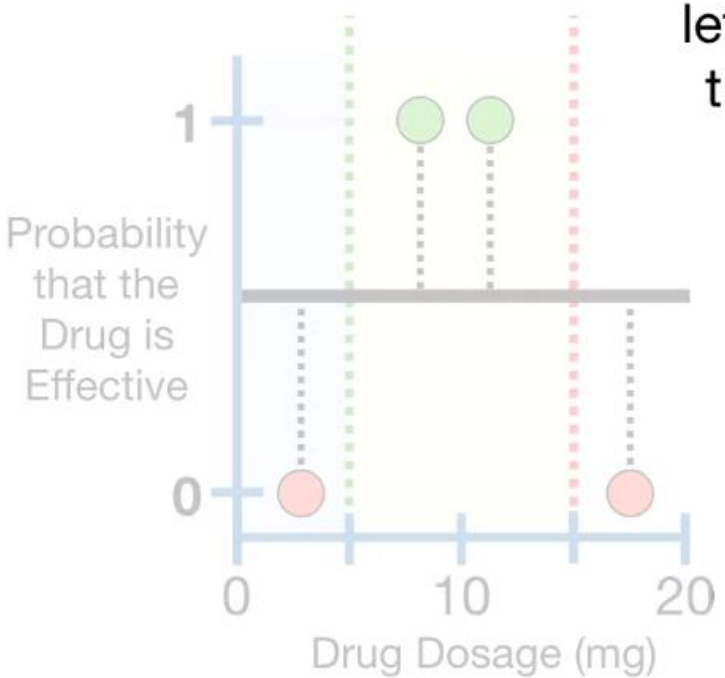
Predicted Drug Effectiveness

0.5



Output = -2

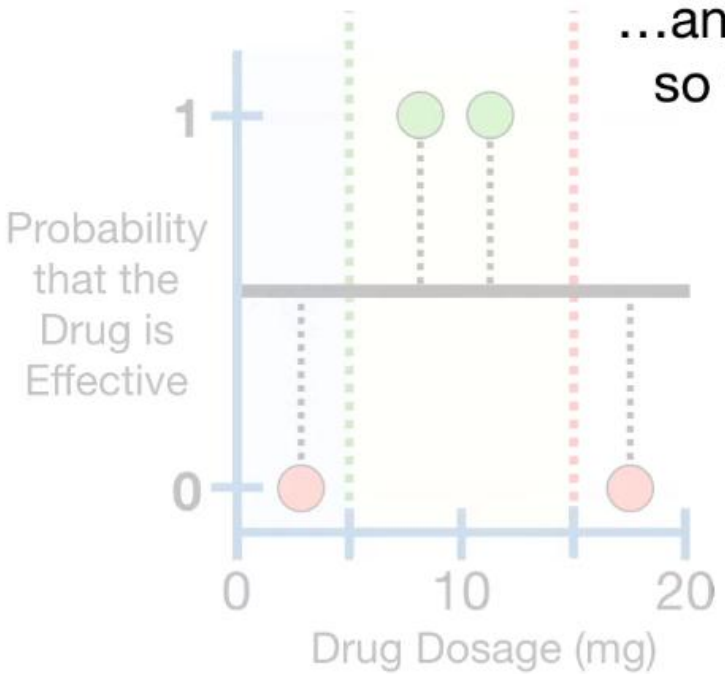
But like I said, we'll let $\lambda = 0$ since that is the default value…



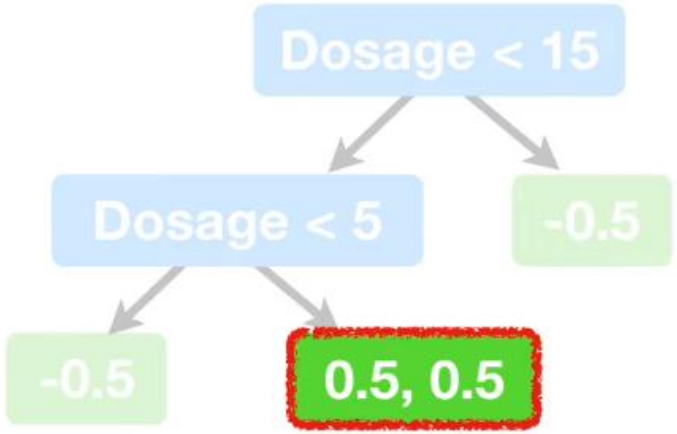
$$\frac{0.5 + 0.5}{0.5 \times (1 - 0.5) + 0.5 \times (1 - 0.5) + 0} = 2$$

Predicted Drug Effectiveness

0.5



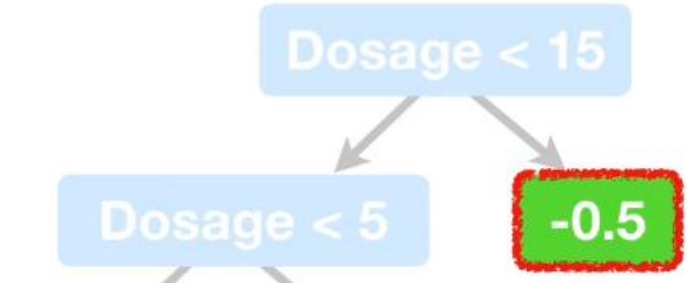
…and put **2** under the leaf so we will remember it.



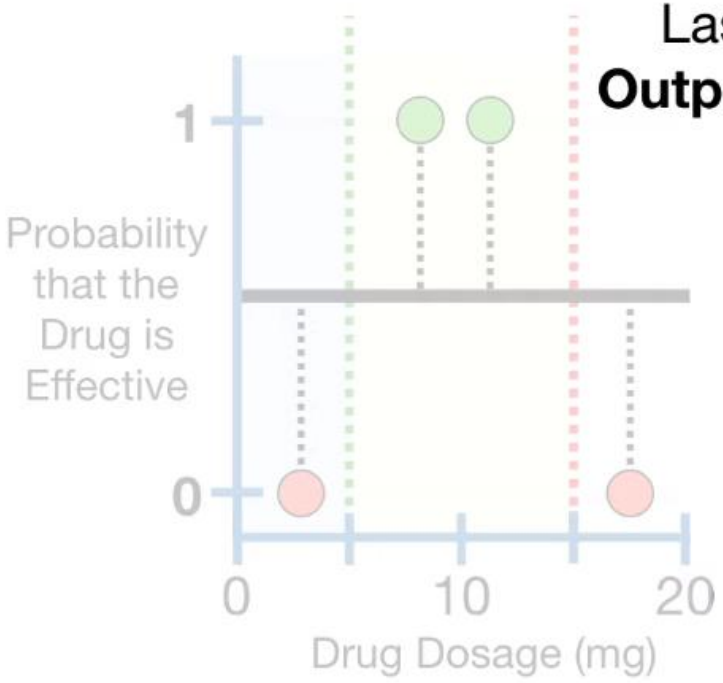
$$\frac{0.5 + 0.5}{0.5 \times (1 - 0.5) + 0.5 \times (1 - 0.5) + \mathbf{0}} = 2$$

Predicted Drug Effectiveness

0.5



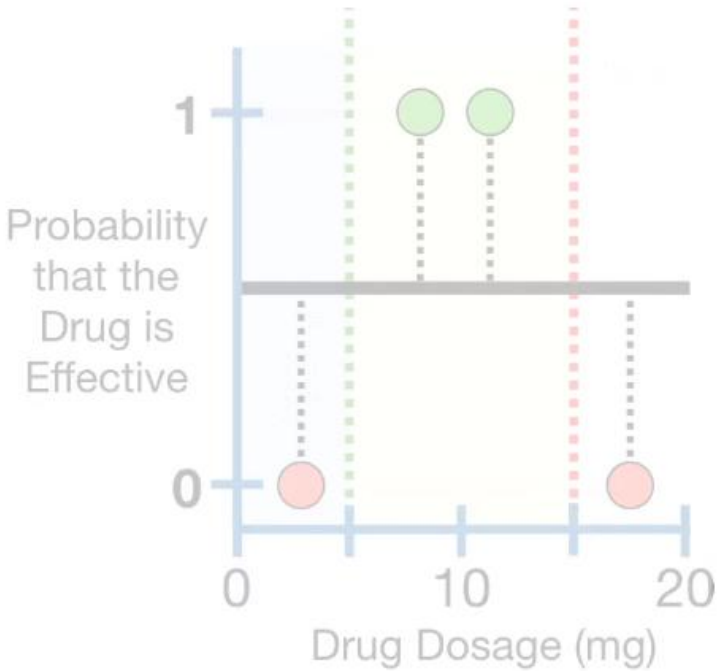
Lastly, when $\lambda = 0$, the **Output Value** for this leaf…



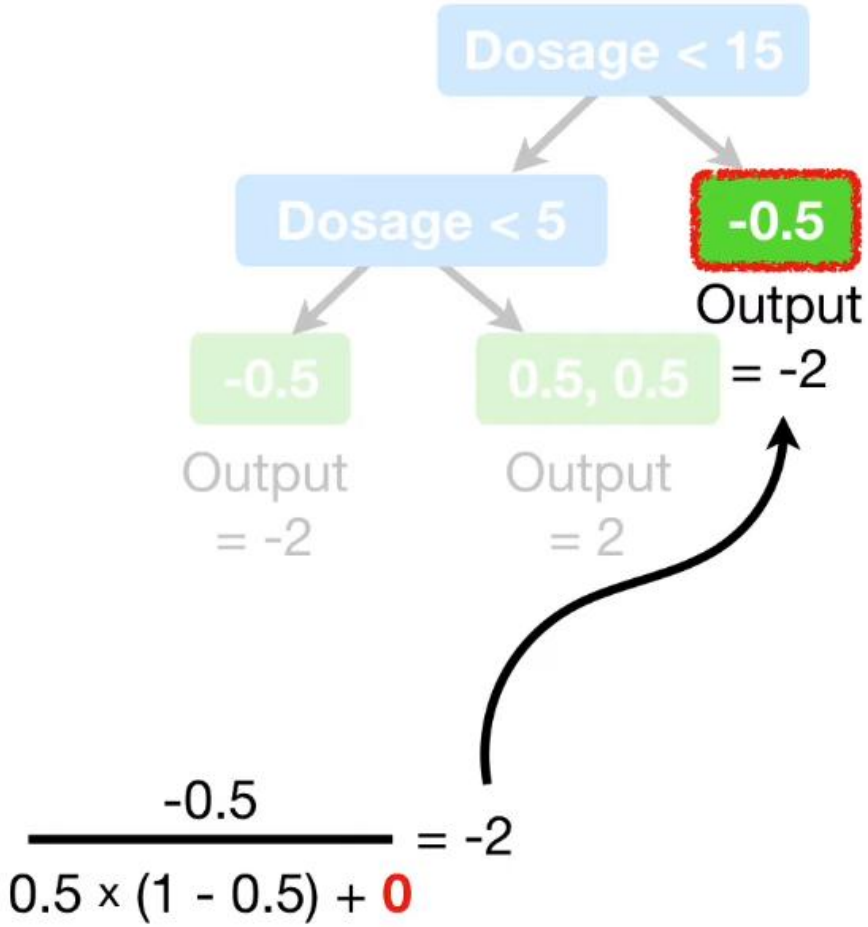
$$\frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Predicted Drug Effectiveness

0.5

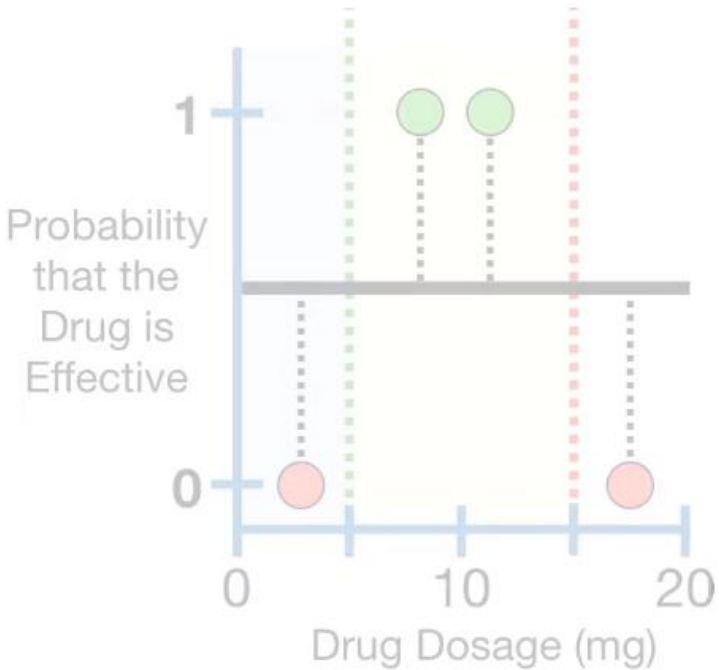


…is **-2**.

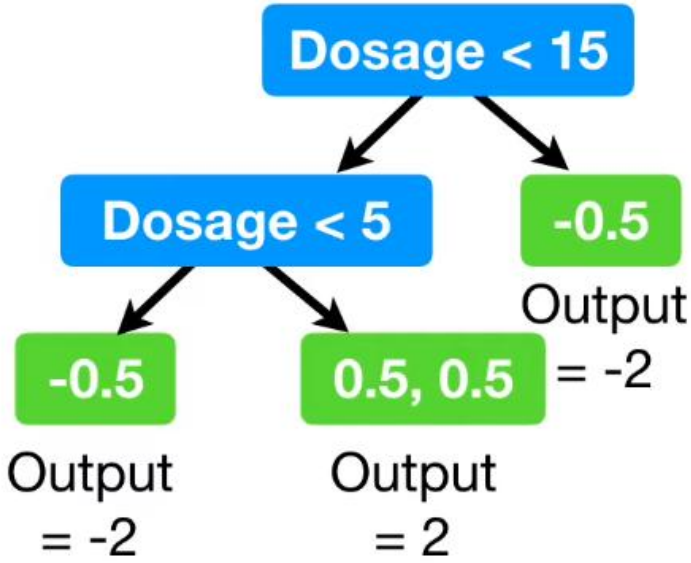


Predicted Drug Effectiveness

0.5



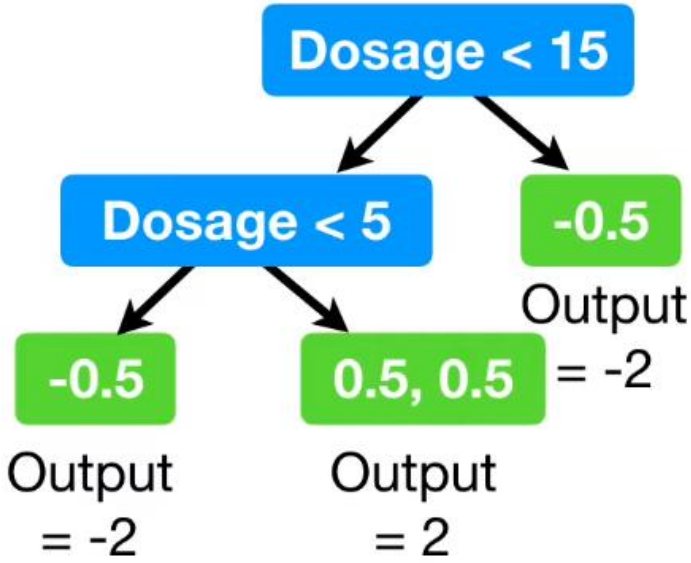
Hooray!!!
The first tree is complete!

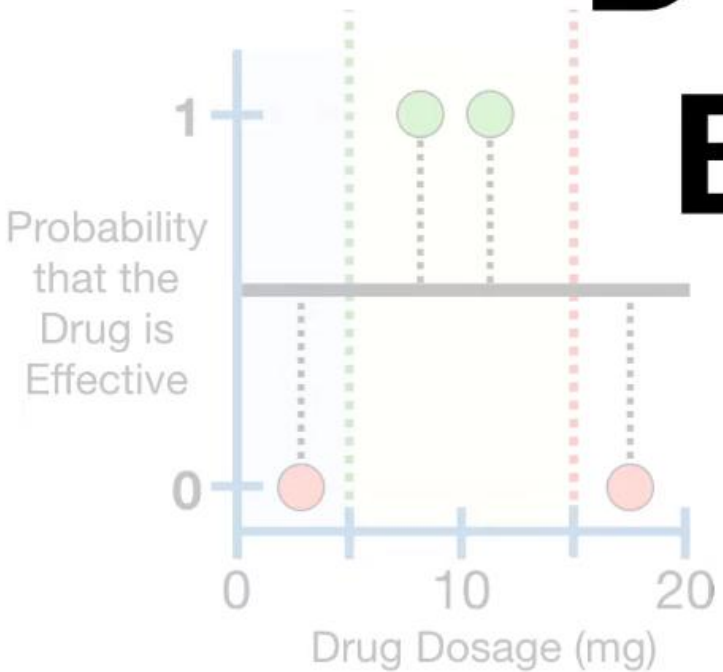


Predicted Drug Effectiveness

0.5

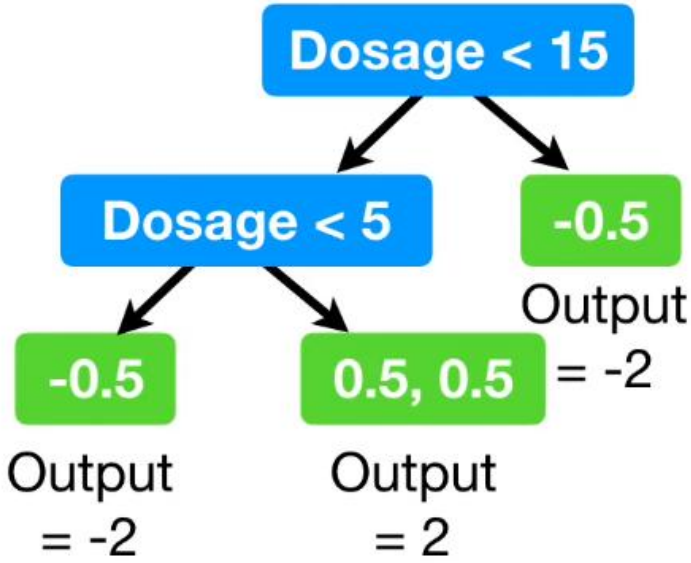
DOUBLE BAM!!!



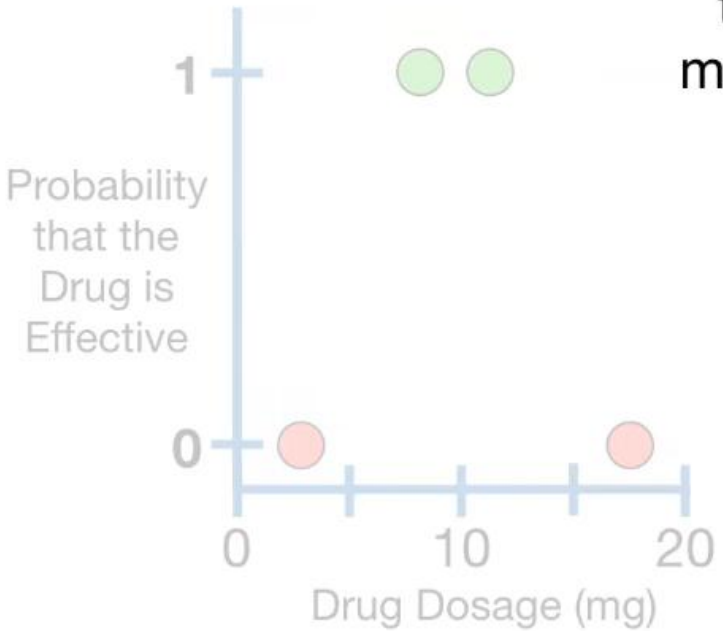


Predicted Drug Effectiveness

0.5



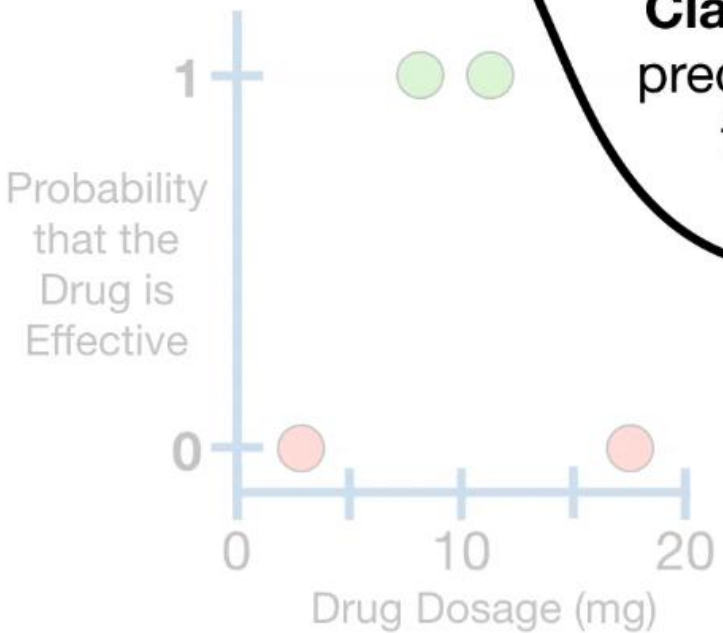
Now that we have built the first tree, we can make new **Predictions**.

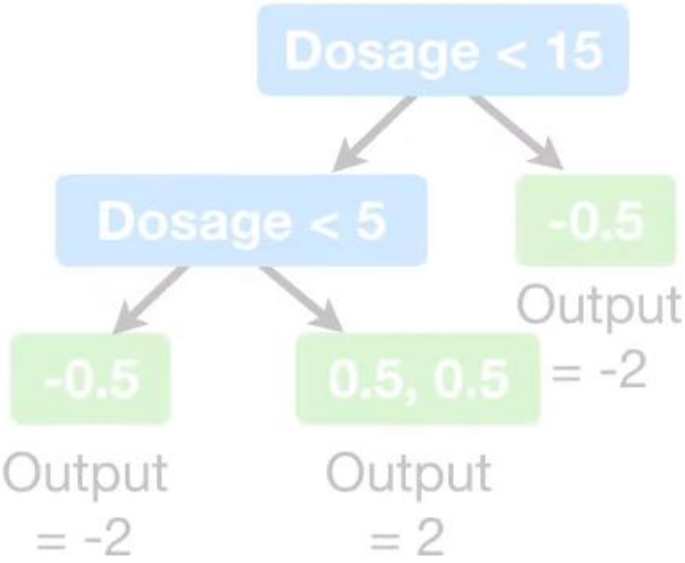


Predicted Drug Effectiveness

0.5

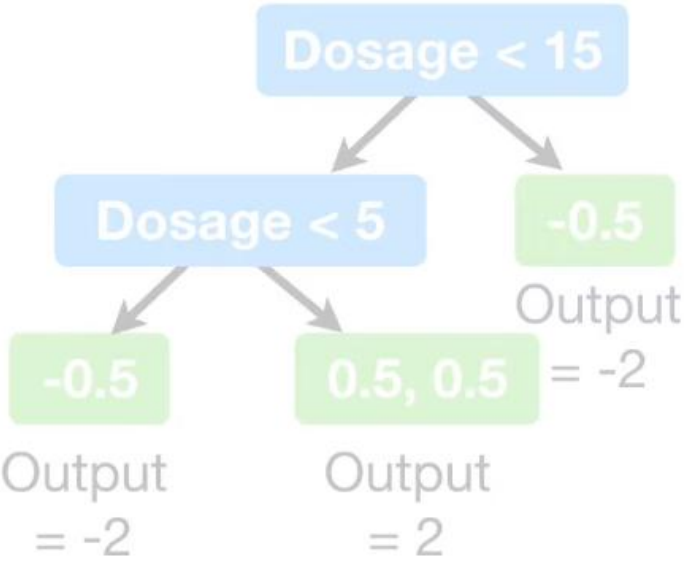
Just like other boosting methods, **XGBoost** for **Classification** makes new predictions by starting with the initial prediction.



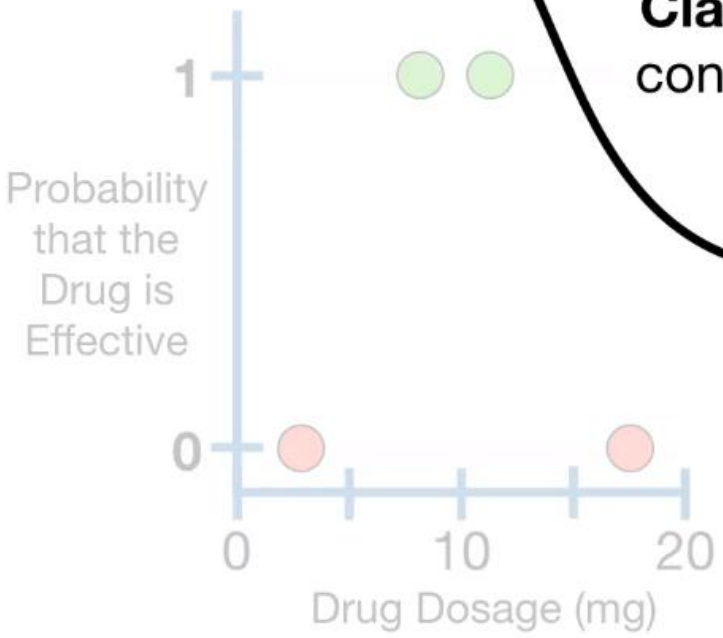


Predicted Drug Effectiveness

0.5



However, just like with unextreme **Gradient Boost** for **Classification**, we need to convert this probability to a **log(odds)** value.

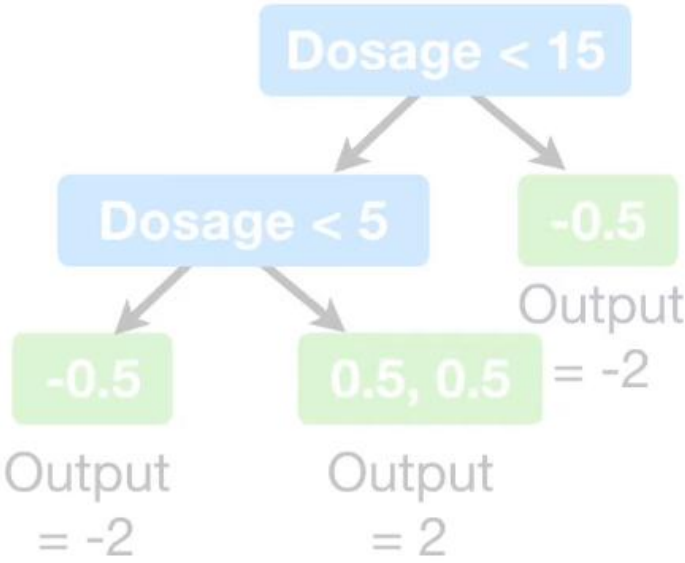


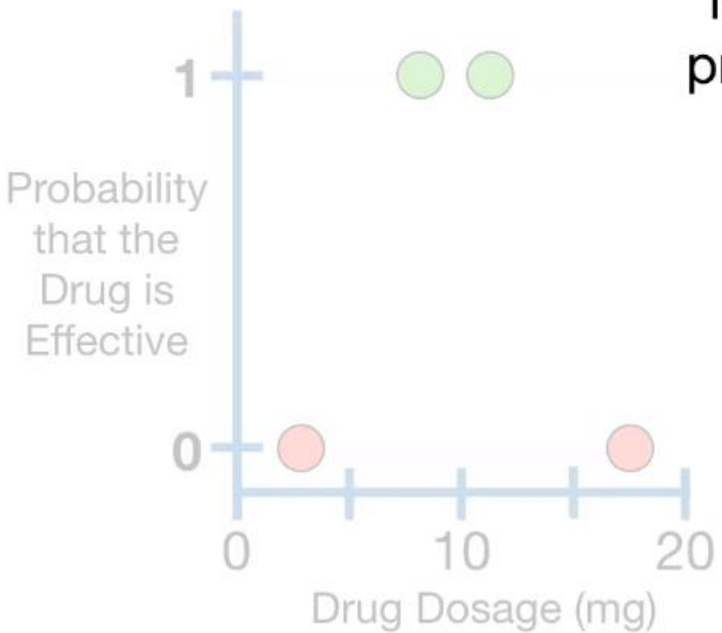
Predicted Drug Effectiveness

0.5

$$\frac{p}{1-p} = \text{odds}$$

So, since this is the formula that converts probabilities to **odds**…





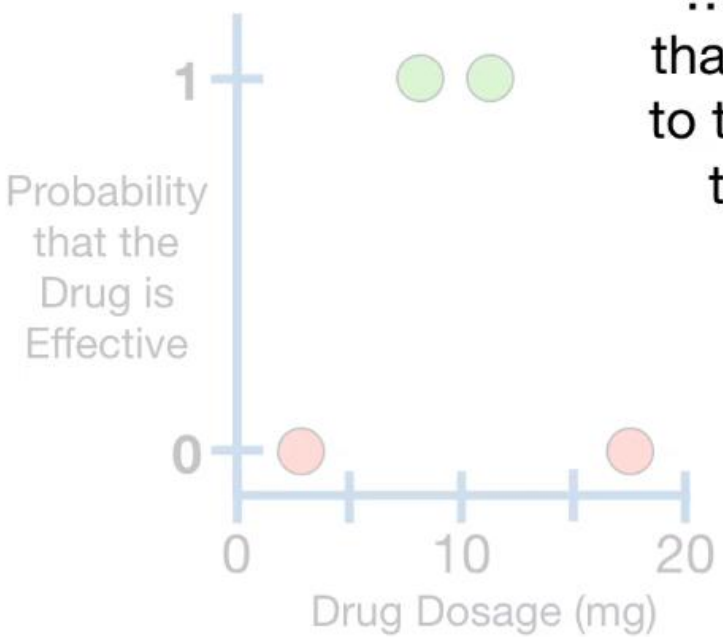
Predicted Drug Effectiveness

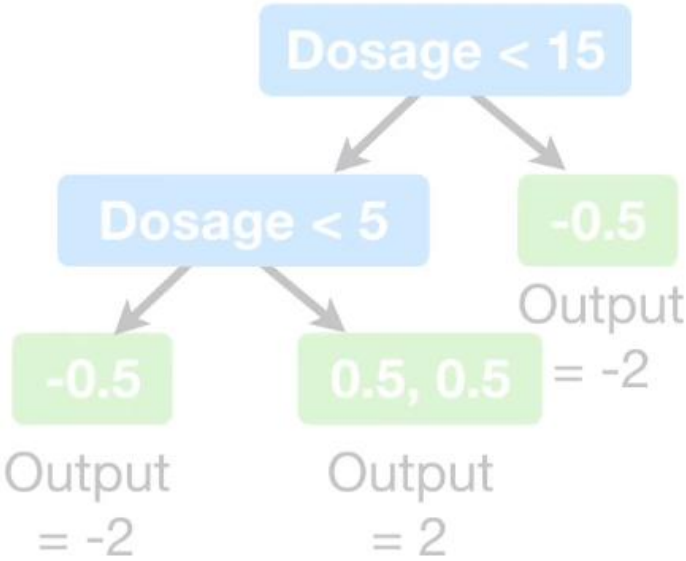
0.5

$$\frac{p}{1-p} = \text{odds}$$

$$\log\left(\frac{p}{1-p}\right) = \log(\text{odds})$$

…we can get a formula that converts probabilities to the **log(odds)** by taking the log of both sides.





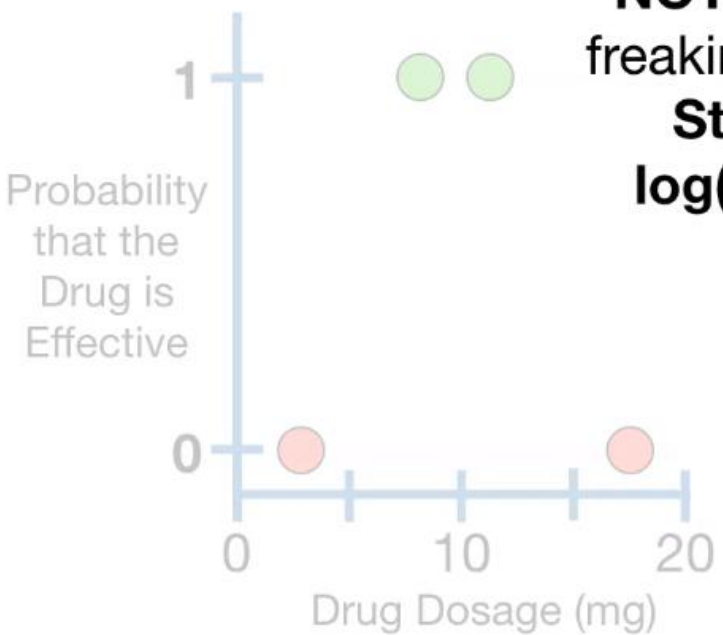
Predicted Drug Effectiveness

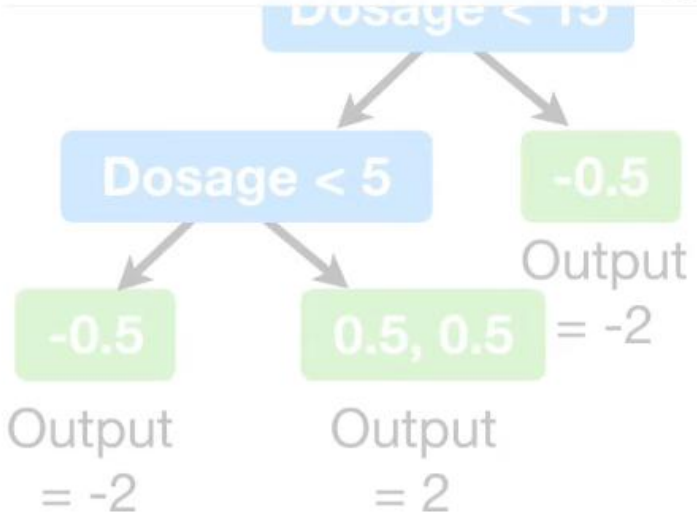
0.5

$$\frac{p}{1-p} = \text{odds}$$

$$\log\left(\frac{p}{1-p}\right) = \log(\text{odds})$$

NOTE: If these equations are freaking you out, just watch the **StatQuest** on **odds** and **log(odds)**. The link is in the description below.



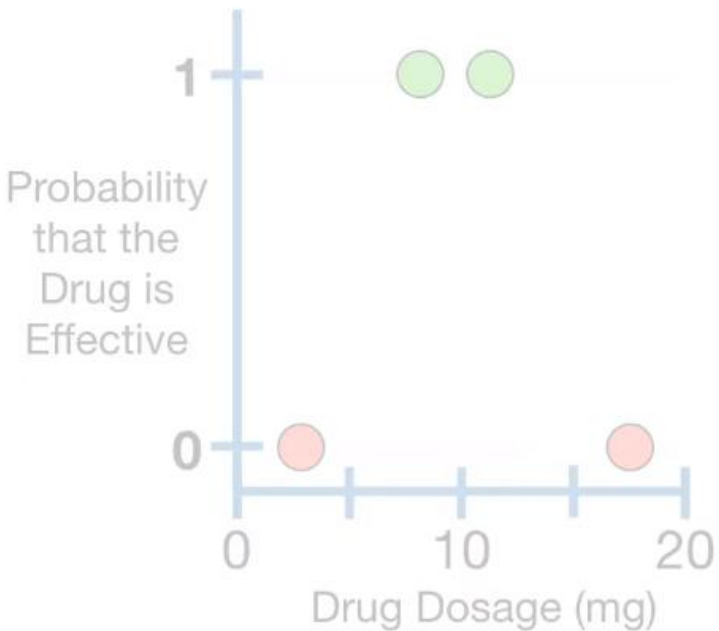


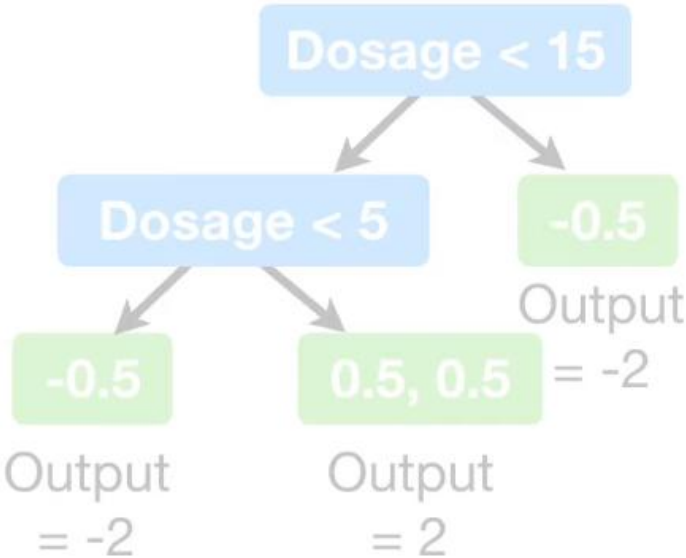
Predicted Drug Effectiveness

0.5

$$\log\left(\frac{0.5}{1 - 0.5}\right) = \log(\text{odds})$$

…do the math…

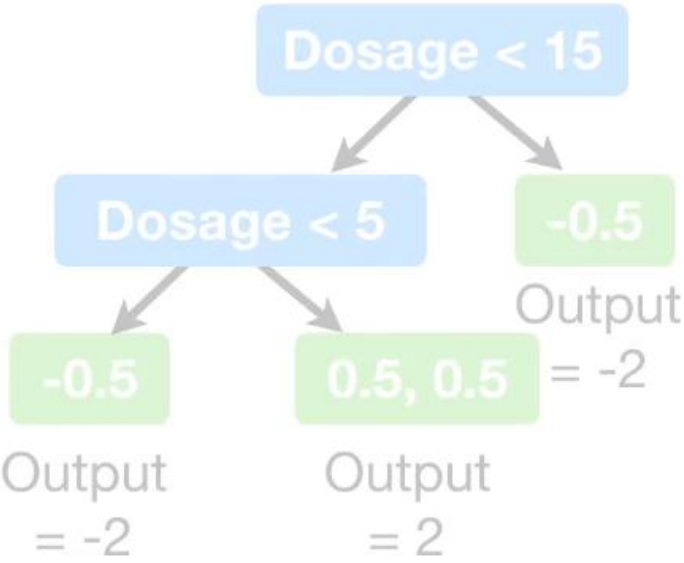




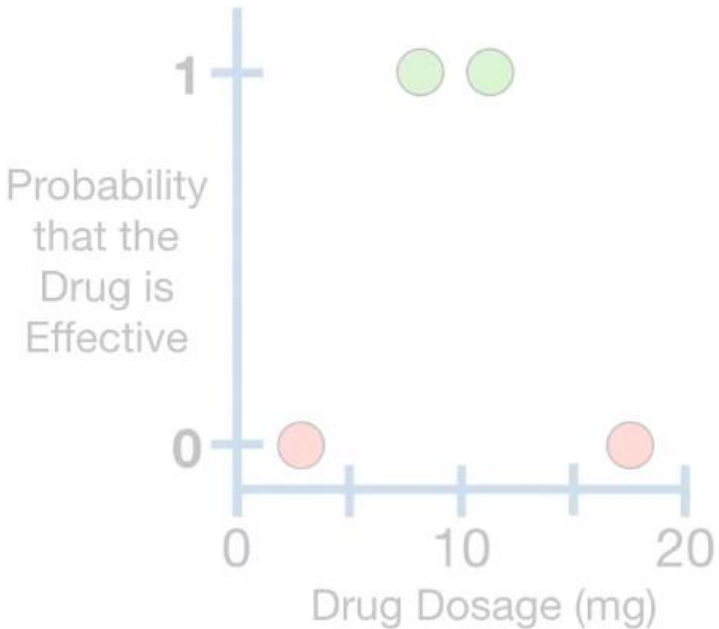
Predicted Drug Effectiveness

0.5

0 = log(odds)



…do the math…

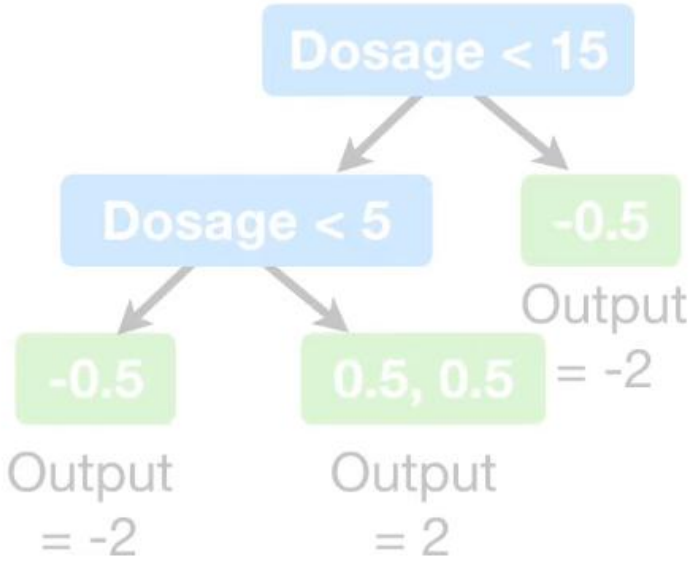


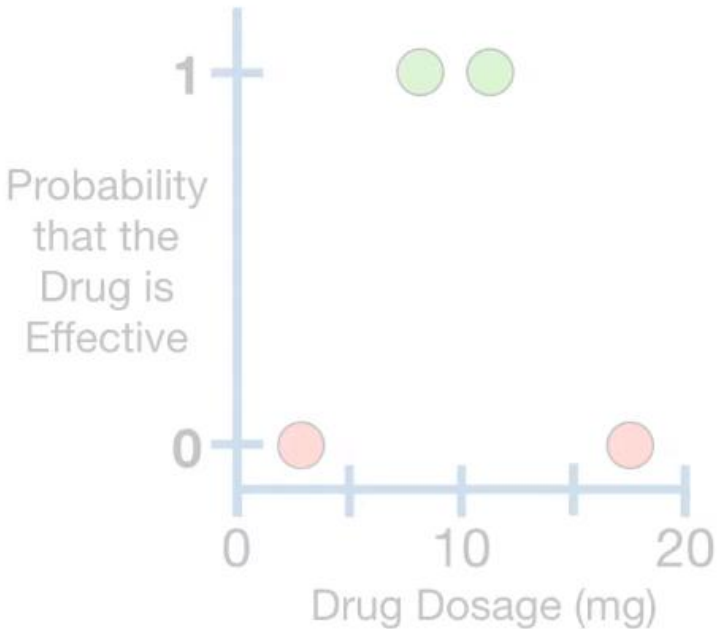
Predicted Drug Effectiveness

0.5

0 = log(odds)

…and we see that when ***p* = 0.5**, the **log(odds) = 0**.



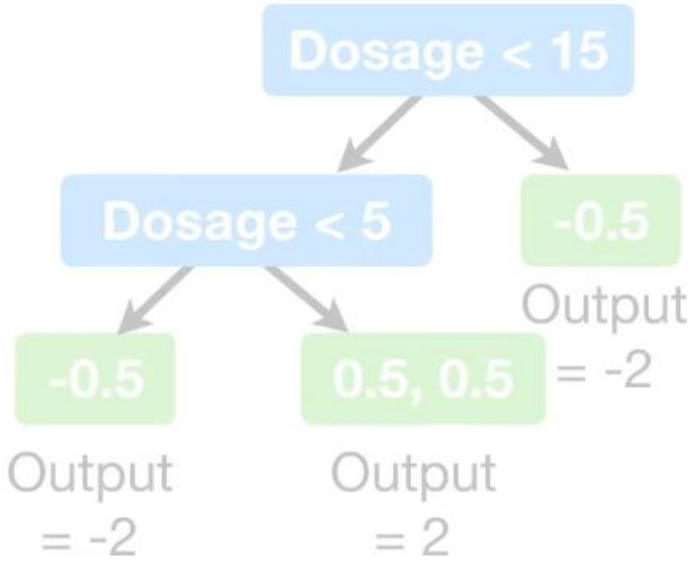


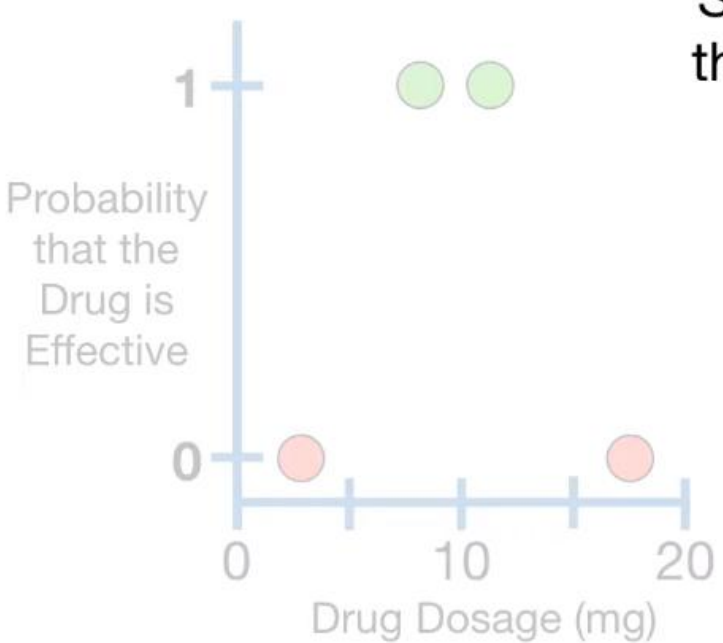
Predicted Drug Effectiveness

0.5

Output = log(odds) = 0

So let's put that under the initial prediction so we don't forget.





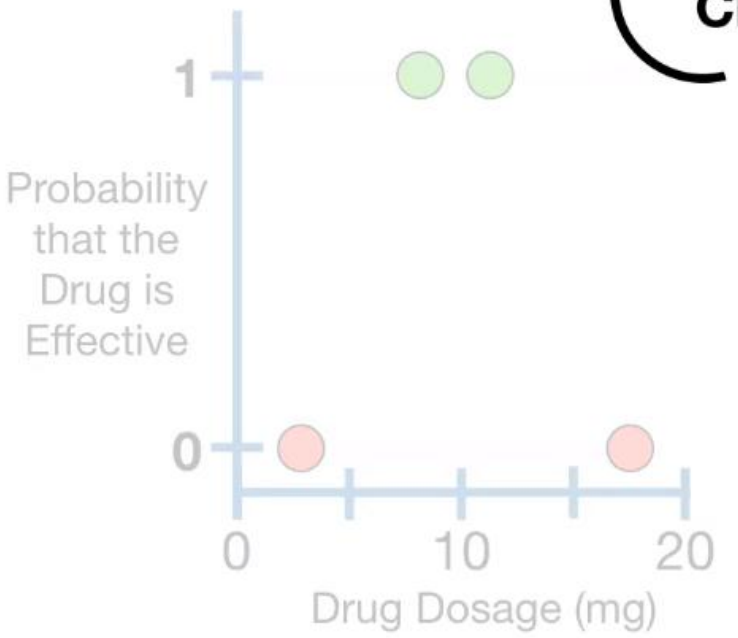
Predicted Drug Effectiveness

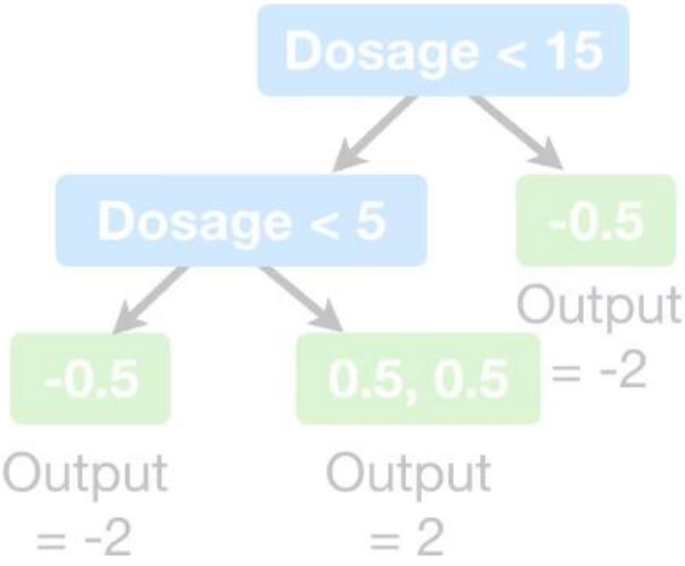
0.5

+

Output = log(odds) = 0

Now, just like unextreme **Gradient Boost** for **Classification**, we add the **log(odds)** of the initial prediction…



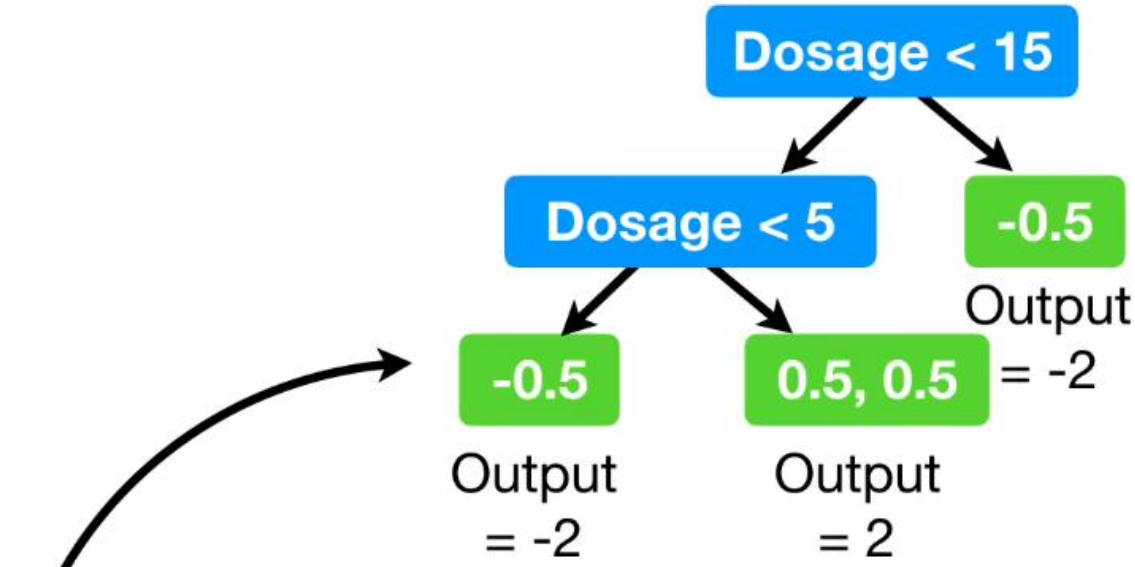


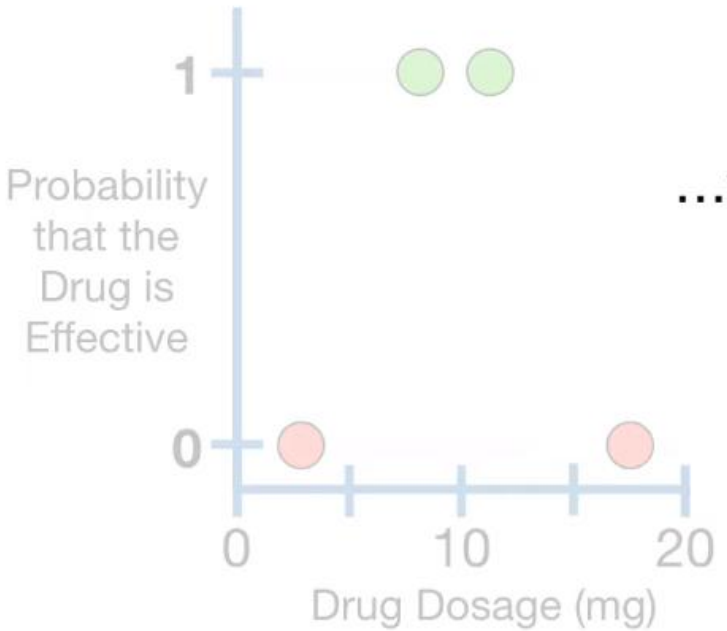
Predicted Drug Effectiveness

0.5

+

Output = log(odds) = 0





…to the output of the **Tree**, scaled by a **Learning Rate**.

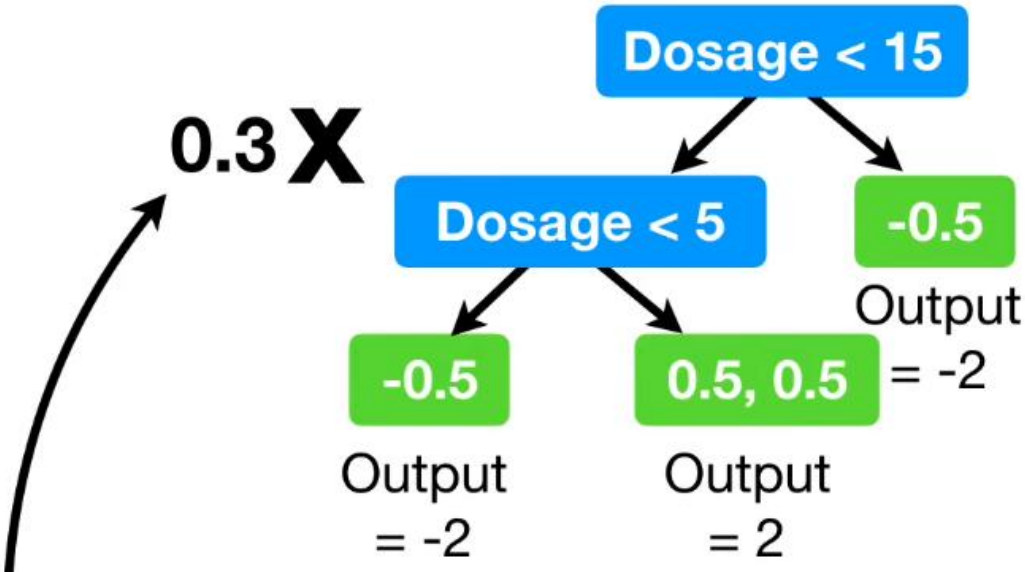
Predicted Drug Effectiveness

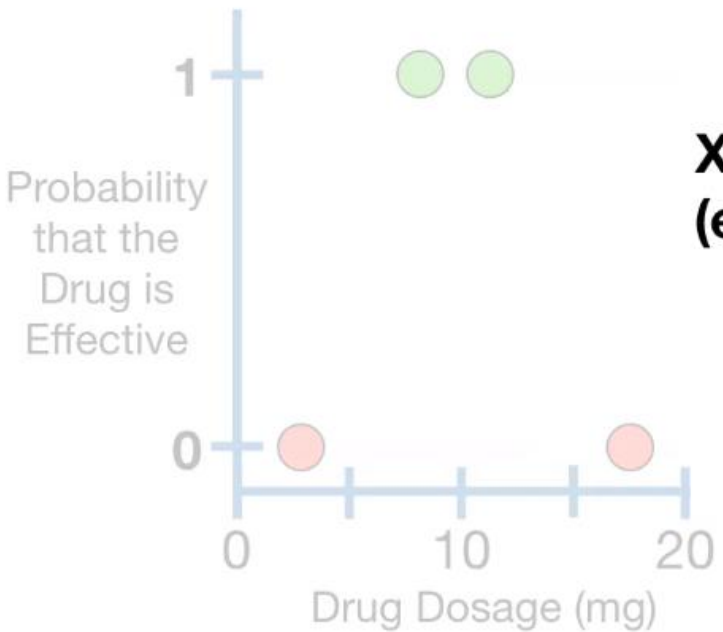
0.5

+

Output = log(odds) = 0

0.3 X





XGBoost calls the **Learning Rate ε (eta)** and the default value is **0.3**, so that's what we'll use.

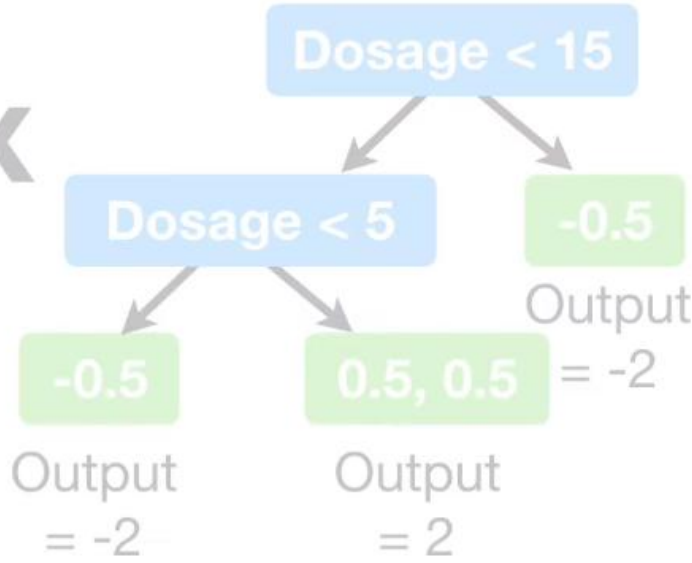
Predicted Drug Effectiveness

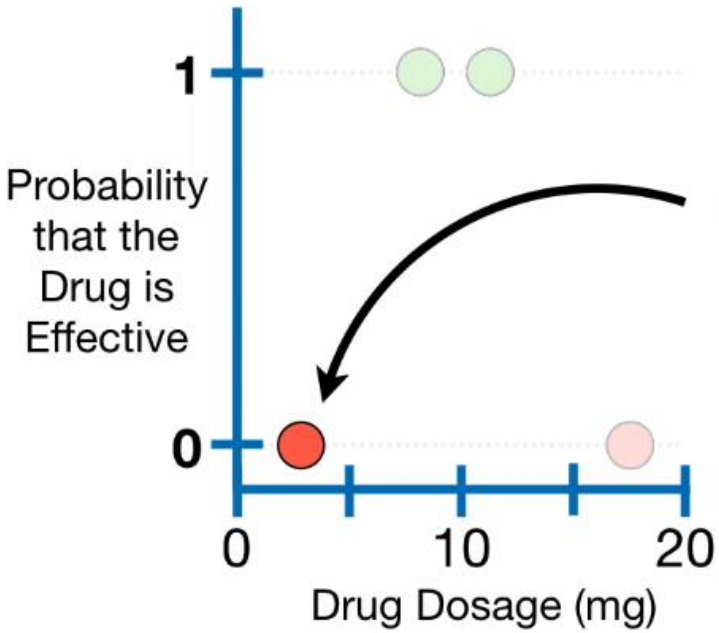
0.5

+

Output = log(odds) = 0

0.3 X





Thus, the new **Predicted** value for this observation, with **Dosage = 2**…

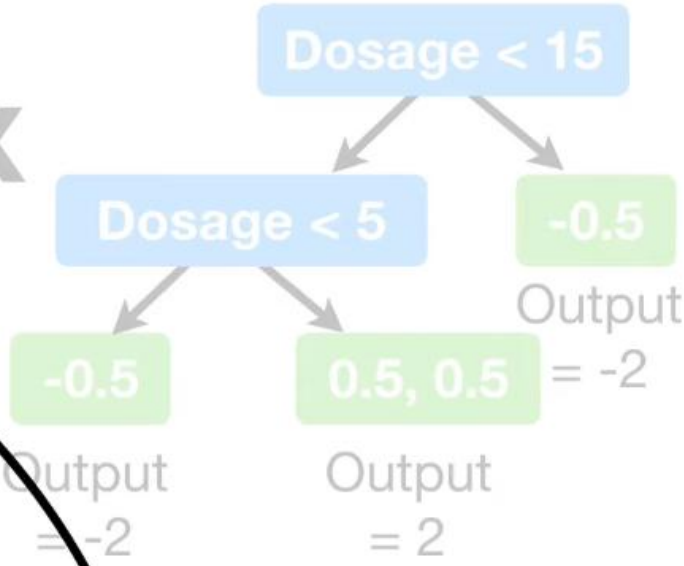
Predicted Drug Effectiveness

0.5

+

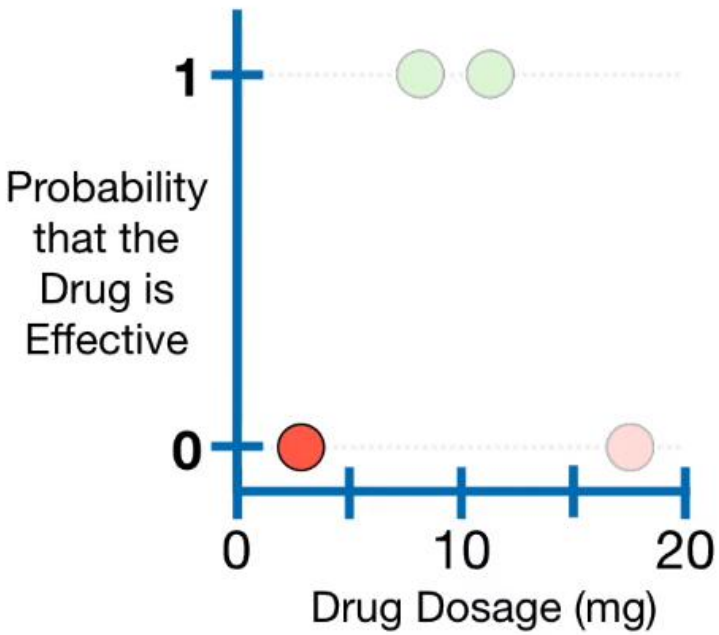
Output = log(odds) = 0

0.3 X



…is the **log(odds)** of original prediction, **0**…

log(odds) Prediction = 0



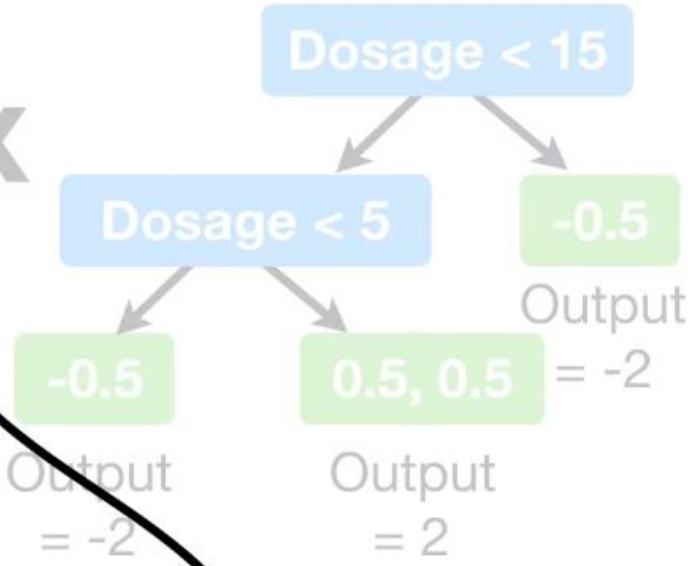
Predicted Drug Effectiveness

0.5

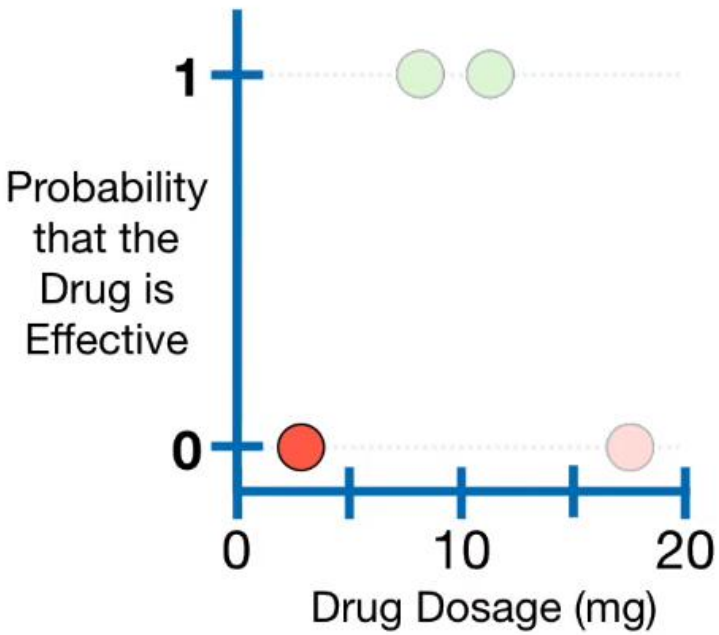
+

Output = log(odds) = 0

0.3 X



…plus the **Learning Rate** (***ε*, eta**), **0.3**…



log(odds) Prediction = 0 + (0.3

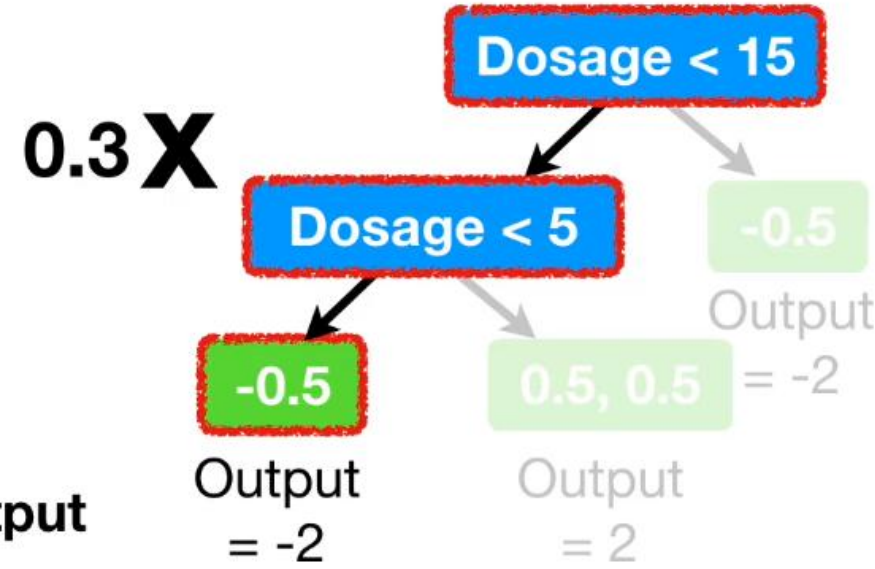
Predicted Drug Effectiveness

0.5

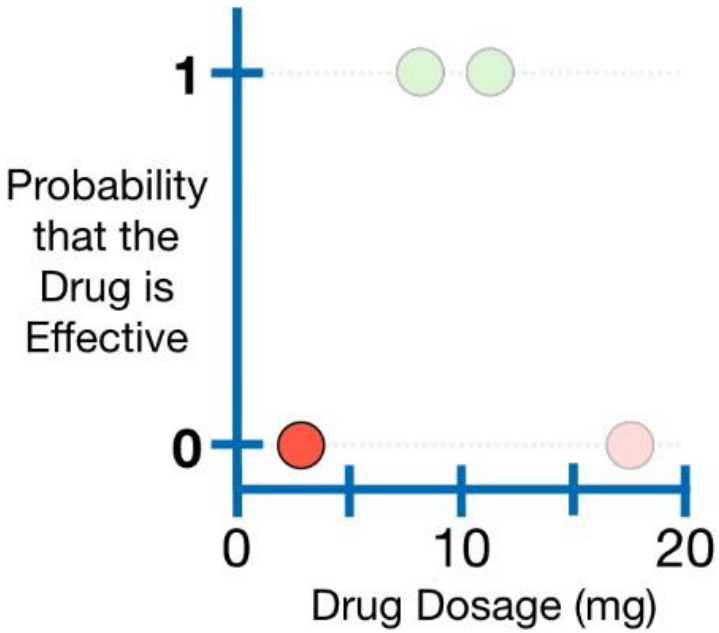
+

Output = log(odds) = 0

0.3 X



…times the **Output Value**, **-2**…



log(odds) Prediction = 0 + (0.3 x -2)

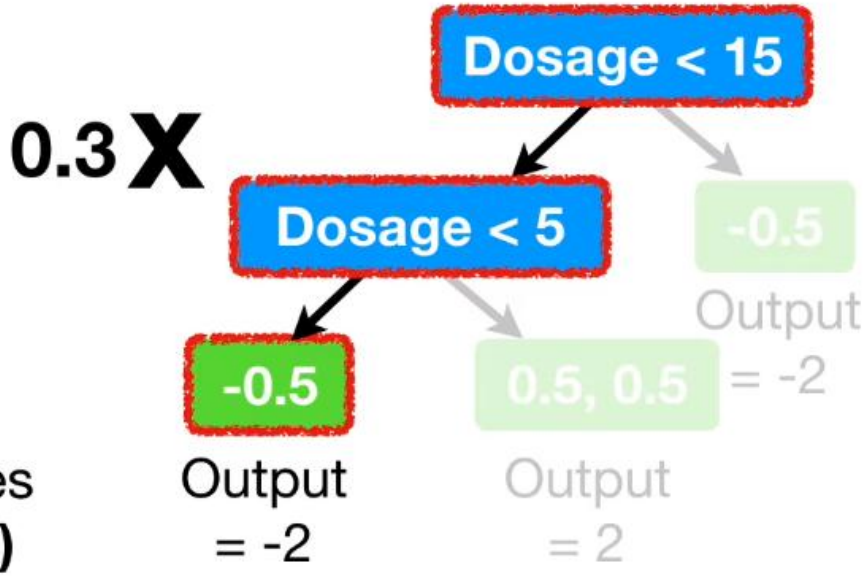
Predicted Drug Effectiveness

0.5

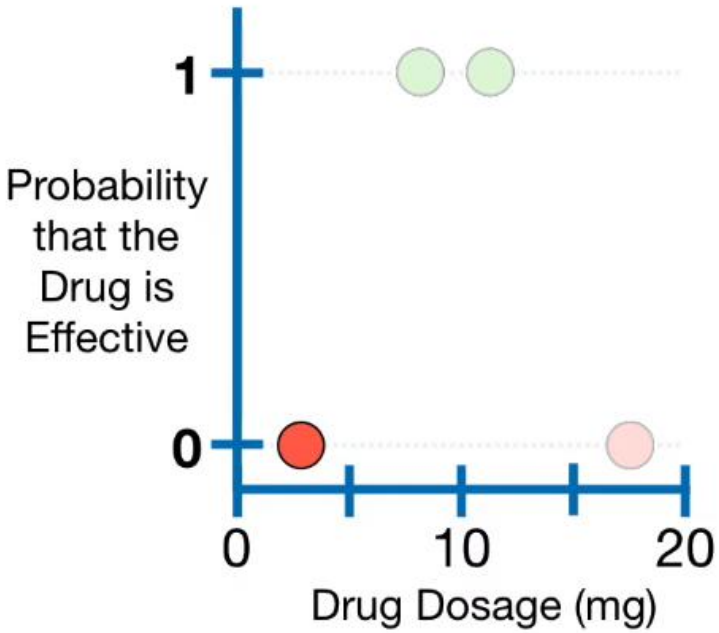
+

Output = log(odds) = 0

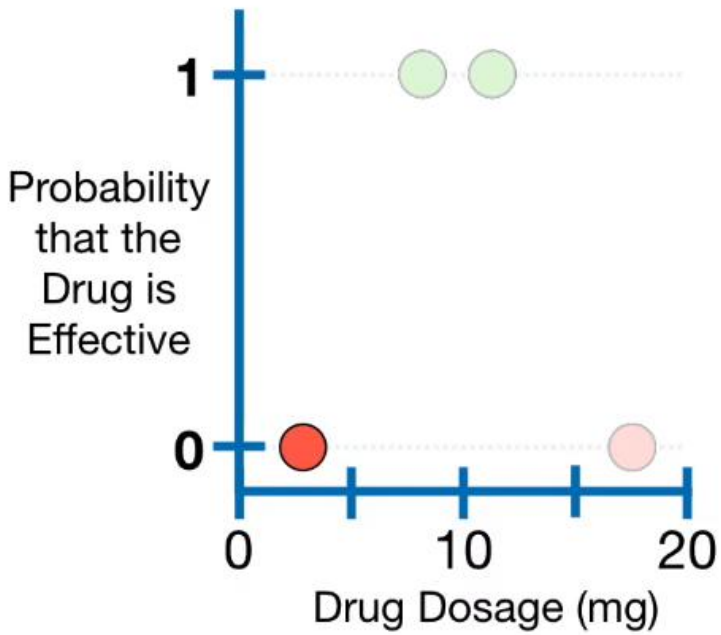
0.3 X



…and that gives us a **log(odds)** value = **-0.6**.



log(odds) Prediction = 0 + (0.3 x -2) = -0.6

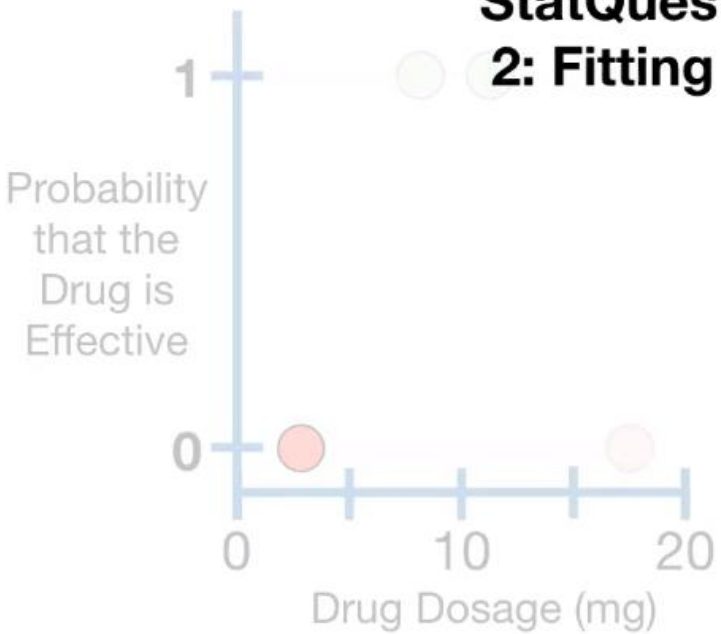


To convert a **log(odds)** value into a probability, we plug it into the **Logistic Function**.

$$\text{Probability} = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}$$

log(odds) Prediction = 0 + (0.3 x -2) = -0.6

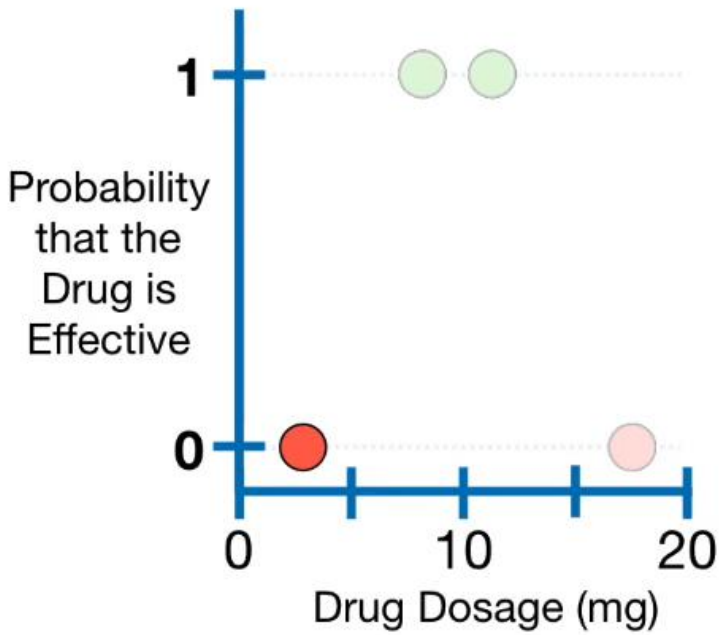
NOTE: If the **Logistic Function** makes you feel a little uncomfortable, check out the **StatQuest**, **Logistic Regression Details Part 2: Fitting a Line With Maximum Likelihood.**



$$\text{Probability} = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}$$

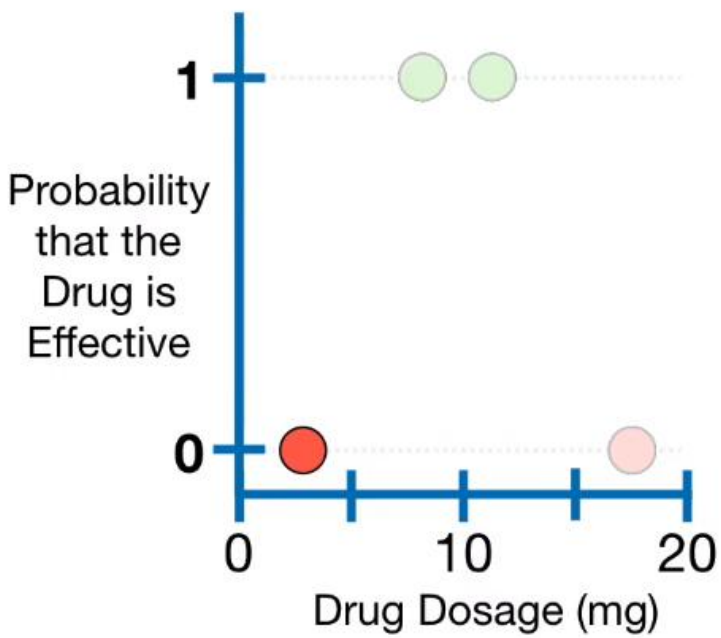
log(odds) Prediction = 0 + (0.3 × -2) = -0.6

Assuming we're cool with this equation, let's plug in the **log(odds)**.



$$\text{Probability} = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}$$

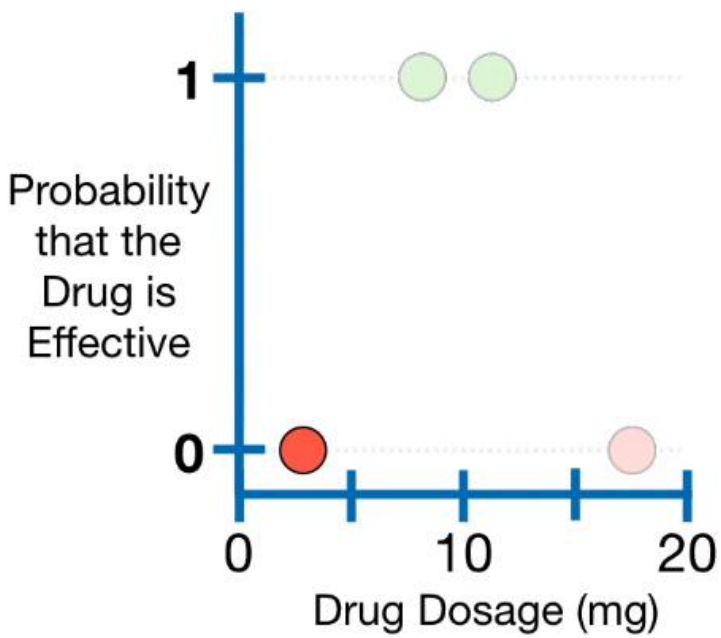
$$\log(\text{odds}) \text{ Prediction} = 0 + (0.3 \times -2) = -0.6$$



…do the math…

$$\text{Probability} = \frac{e^{-0.6}}{1 + e^{-0.6}}$$

log(odds) Prediction = 0 + (0.3 x -2) = -0.6



…and the the new predicted probability is **0.35**.

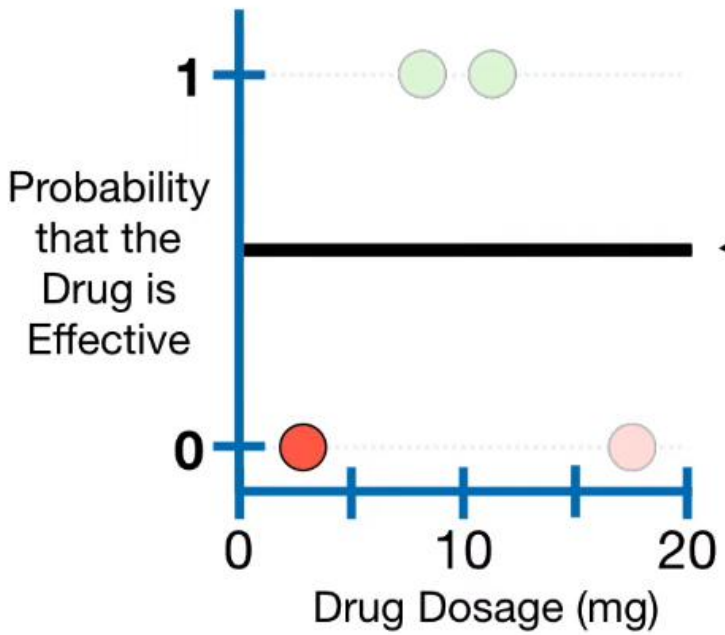
$$\text{Probability} = \frac{e^{-0.6}}{1 + e^{-0.6}} = 0.35$$

log(odds) Prediction = 0 + (0.3 x -2) = -0.6

Predicted Drug Effectiveness

0.5

Output = log(odds) = 0



Remember, the original **Prediction** was **0.5**…

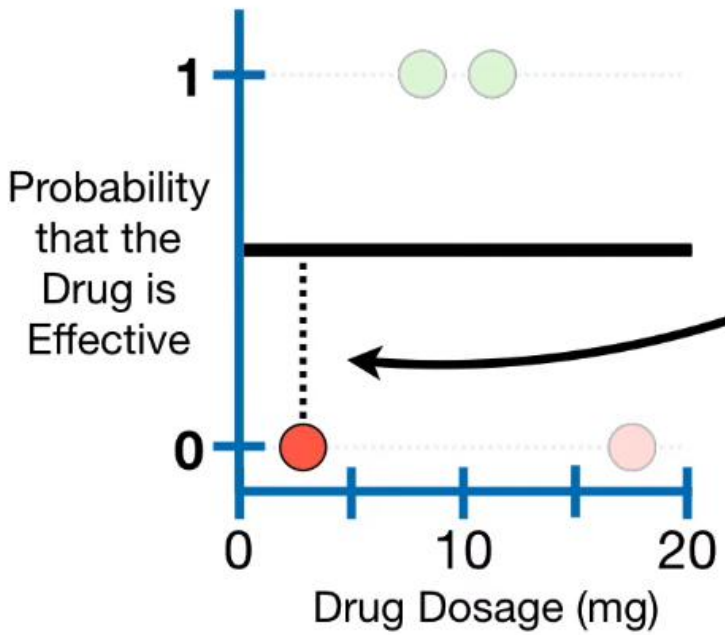
$$\text{Probability} = \frac{e^{-0.6}}{1 + e^{-0.6}} = 0.35$$

log(odds) Prediction = 0 + (0.3 x -2) = -0.6

Predicted Drug Effectiveness

0.5

Output = log(odds) = 0



…and this was the original **Residual**.

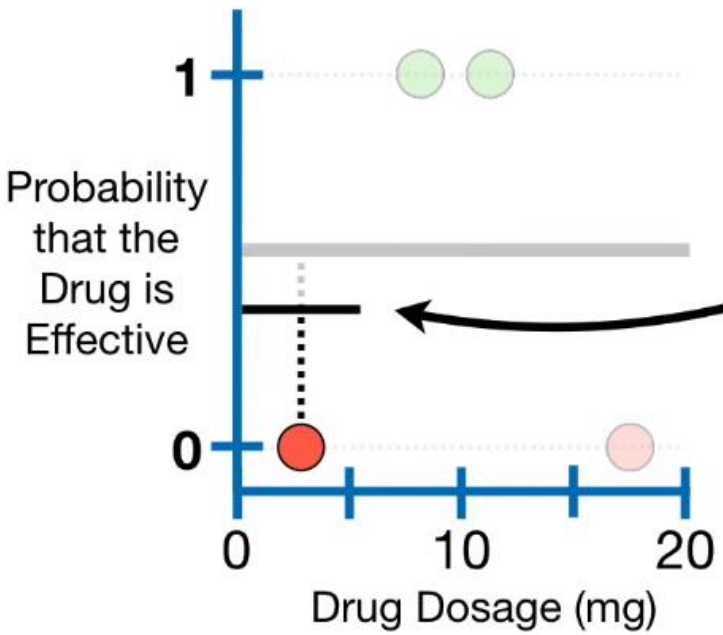
$$\text{Probability} = \frac{e^{-0.6}}{1 + e^{-0.6}} = 0.35$$

log(odds) Prediction = 0 + (0.3 x -2) = -0.6

Predicted Drug Effectiveness

0.5

Output = log(odds) = 0



Now the new predicted probability is **0.35**…

$$\text{Probability} = \frac{e^{-0.6}}{1 + e^{-0.6}} = 0.35$$

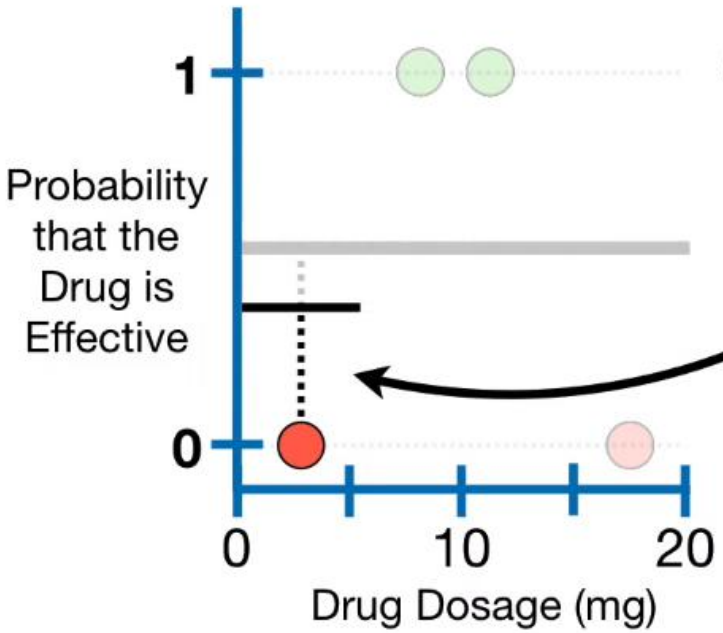
log(odds) Prediction = 0 + (0.3 x -2) = -0.6

Predicted Drug Effectiveness

0.5

Output = log(odds) = 0

…and the new **Residual** is smaller than before, so we have taken a small step in the right direction!!!



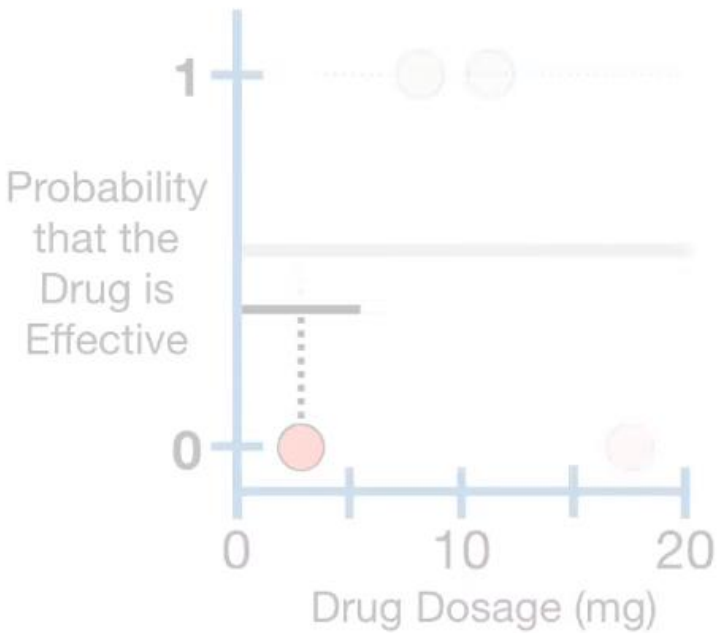
$$\text{Probability} = \frac{e^{-0.6}}{1 + e^{-0.6}} = 0.35$$

log(odds) Prediction = 0 + (0.3 x -2) = -0.6

Predicted Drug Effectiveness

0.5

Output = log(odds) = 0



NOTE: You may be wondering why we even bothered adding the **log(odds)** of the initial prediction since it is **0**.

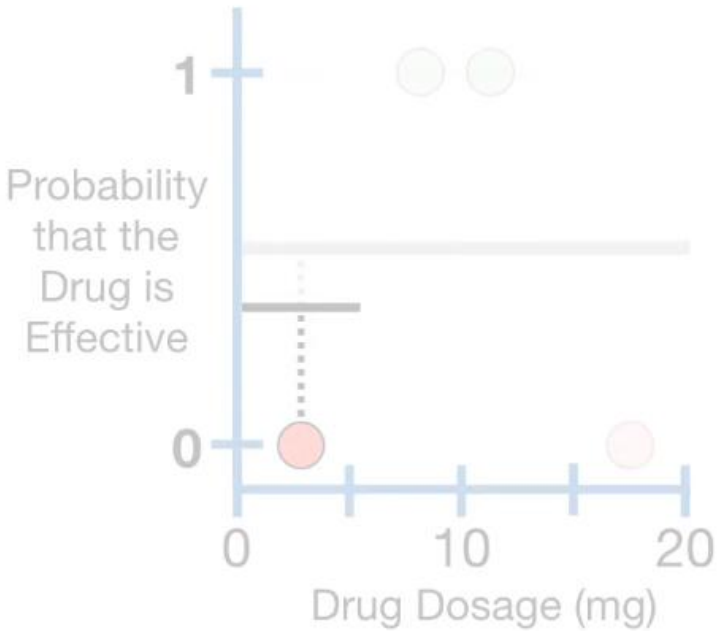
log(odds) Prediction = 0 + (0.3 x -2) = -0.6

Predicted Drug Effectiveness

0.5

Output = log(odds) = 0

This is always the case if you use the default value, **0.5**, for the initial prediction.

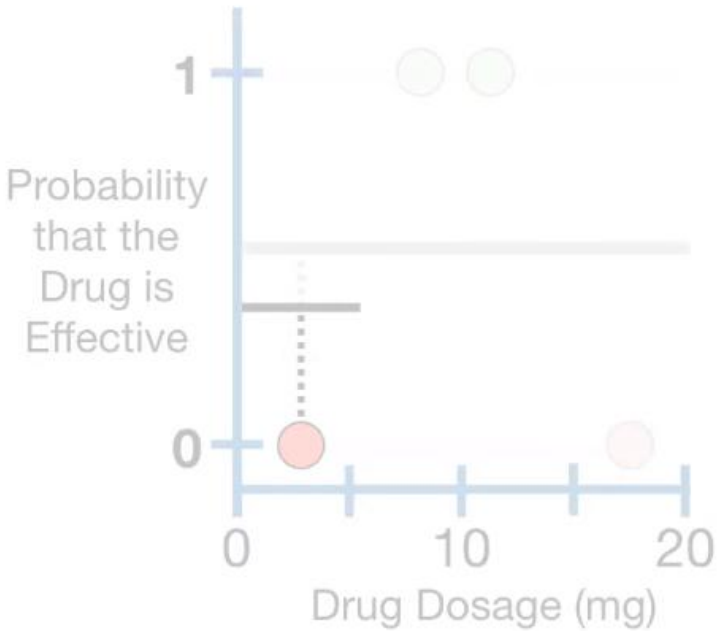


log(odds) Prediction = 0 + (0.3 x -2) = -0.6

Predicted Drug Effectiveness

0.5

Output = log(odds) = 0



However, you can change the initial prediction to any probability, and any value other than **0.5** will give you something more interesting to add.

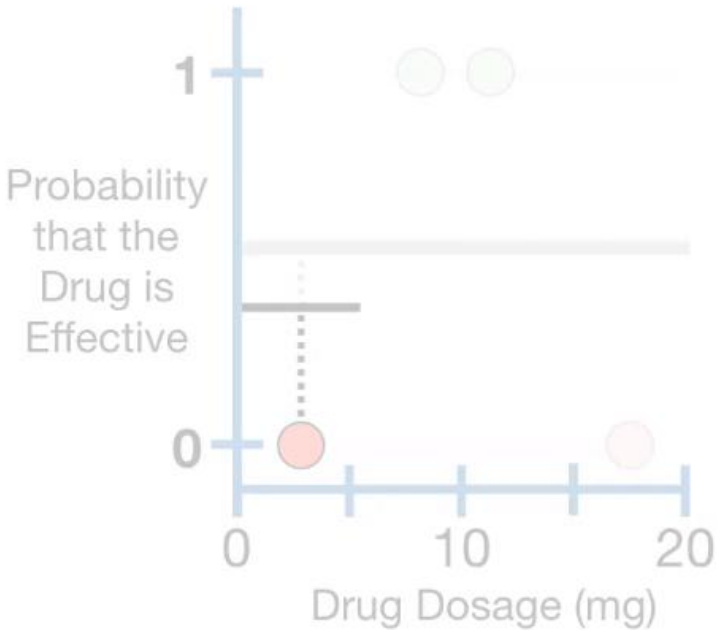
log(odds) Prediction = 0 + (0.3 × -2) = -0.6

Predicted Drug Effectiveness

0.75

Output = log(odds) = 1.1

For example, if **75%** of the observations in the **Training Data** said that the drug was effective, we might set the initial prediction to **0.75** and now the initial **log(odds) = 1.1**…

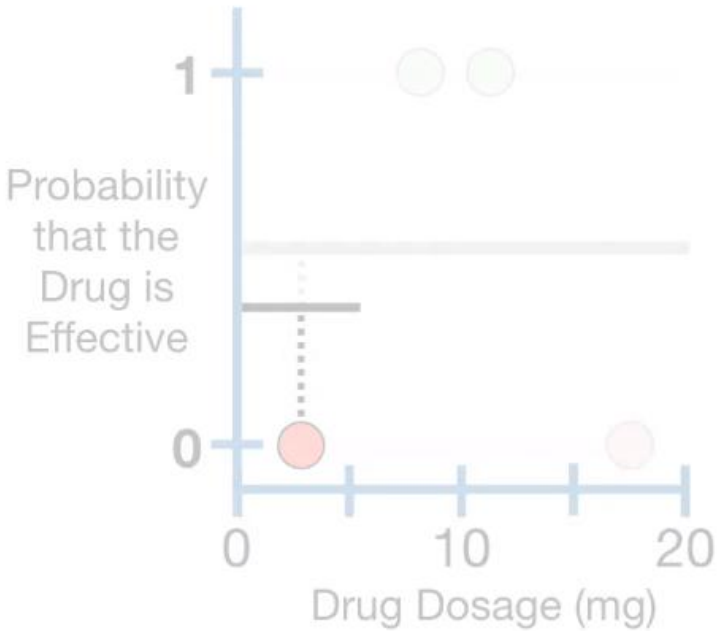


log(odds) Prediction = 0 + (0.3 × -2) = -0.6

Predicted Drug Effectiveness

0.75

Output = log(odds) = 1.1



…so we would plug **1.1** into the equation instead of **0**.

log(odds) Prediction = 0 + (0.3 x -2) = -0.6

Predicted Drug Effectiveness

0.75

Output = log(odds) = 1.1

…so we would plug **1.1** into the equation instead of **0**.

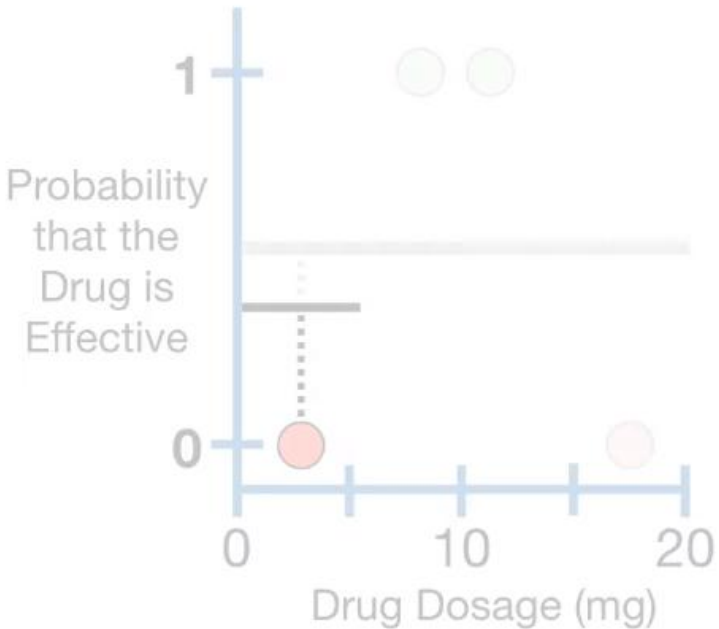
log(odds) Prediction = 1.1 + (0.3 × -2) = 0.5

Predicted Drug Effectiveness

0.5

Output = log(odds) = 0

But since the default initial prediction is **0.5**, we'll use **0** for the initial **log(odds)** in the remaining examples.

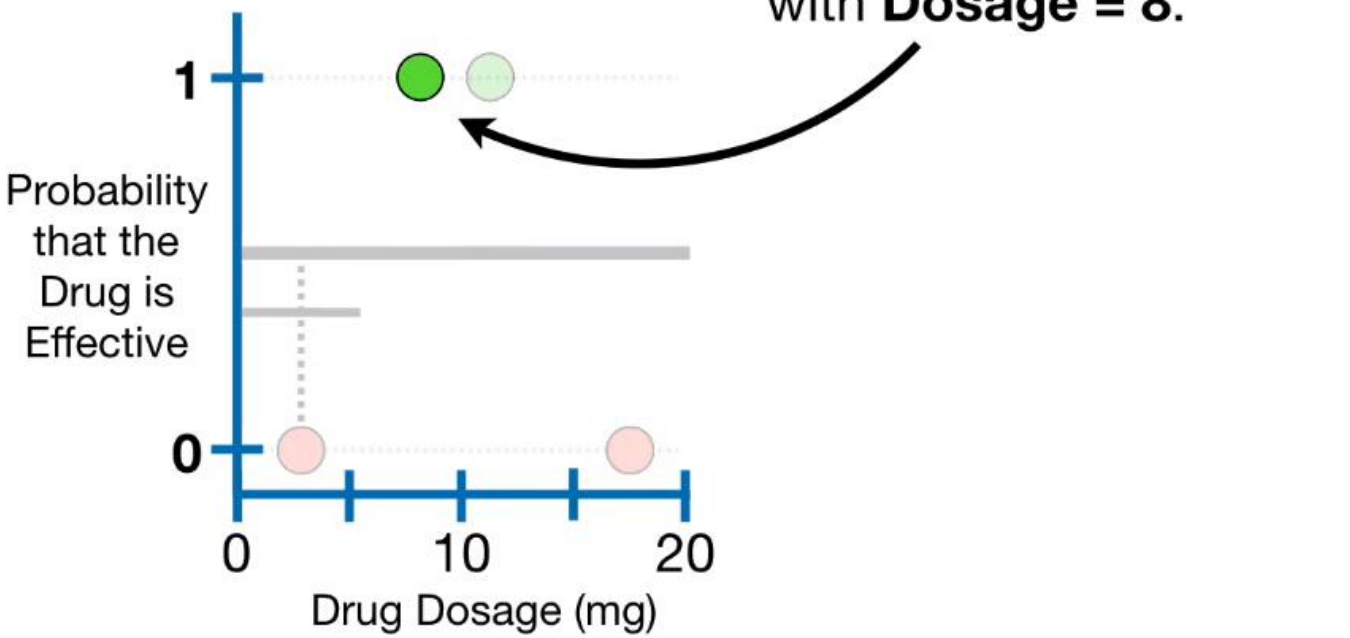


Predicted Drug Effectiveness

0.5

Output = log(odds) = 0

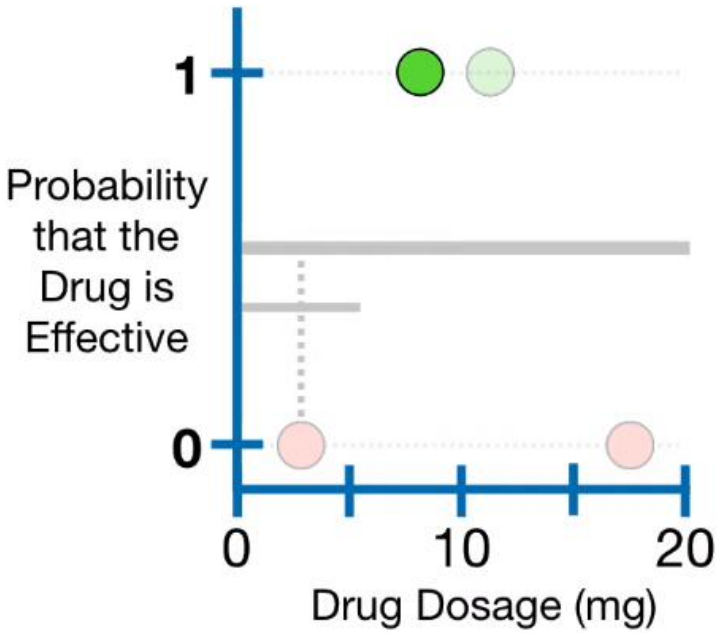
Now let's make a new prediction for this observation, with **Dosage = 8**.



Predicted Drug Effectiveness

0.5

Output = log(odds) = 0



We start with the original **log(odds)** prediction, **0**…

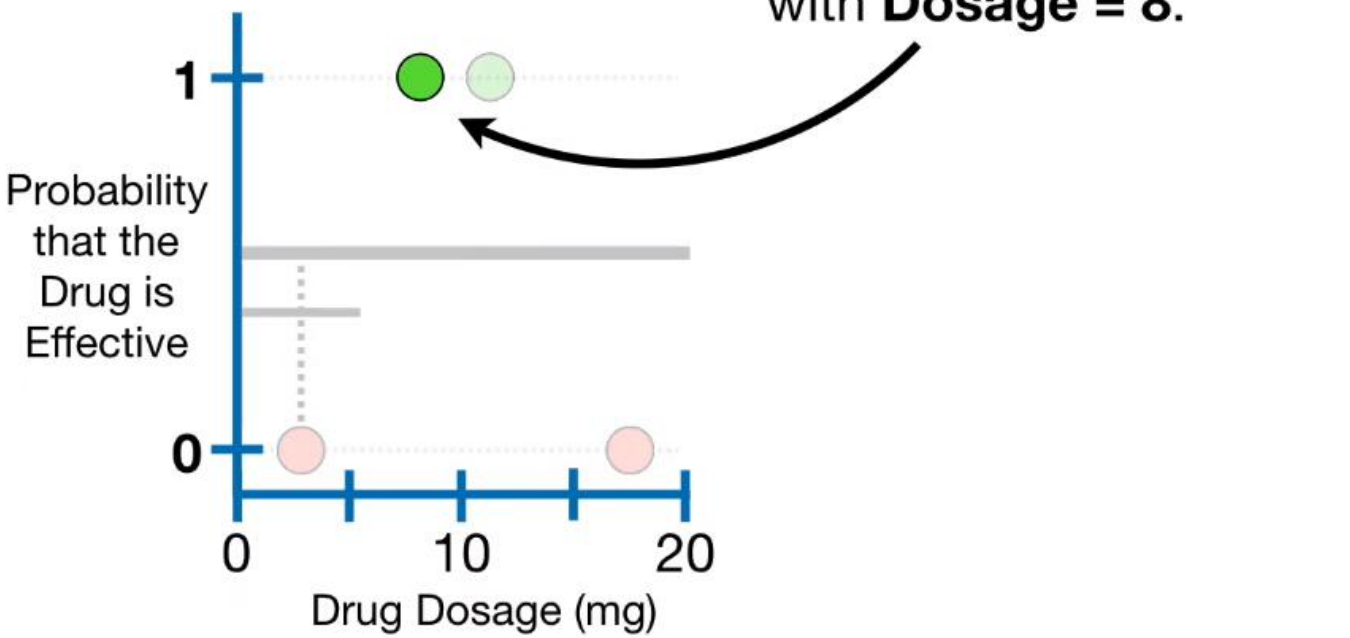
log(odds) Prediction = 0

Predicted Drug Effectiveness

0.5

Output = log(odds) = 0

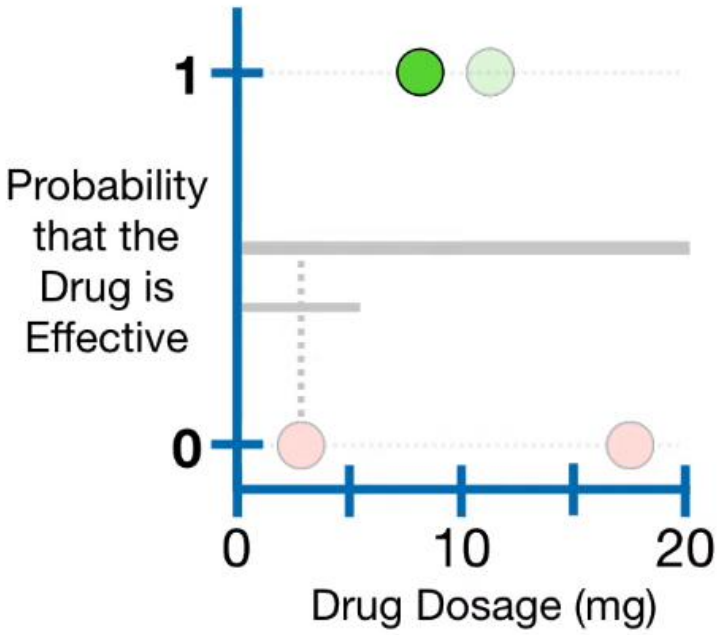
Now let's make a new prediction for this observation, with **Dosage = 8**.



Predicted Drug Effectiveness

0.5

Output = log(odds) = 0



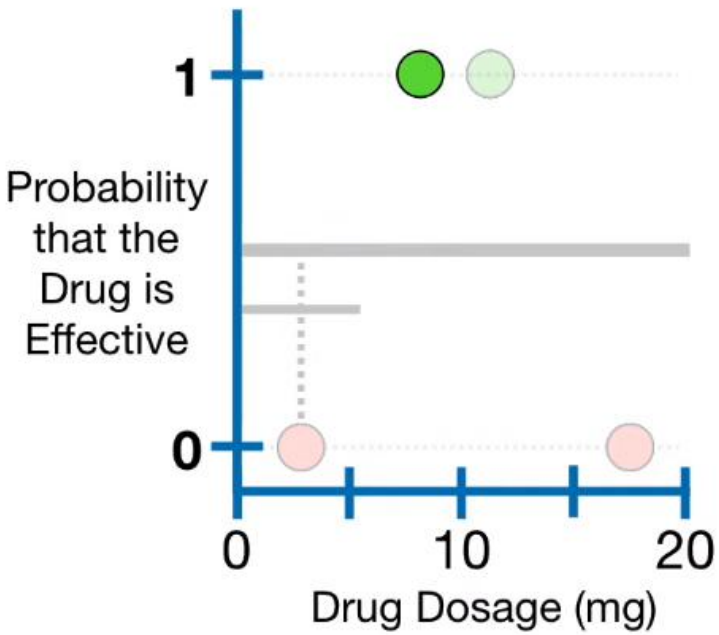
We start with the original **log(odds)** prediction, **0**…

log(odds) Prediction = 0

Predicted Drug Effectiveness

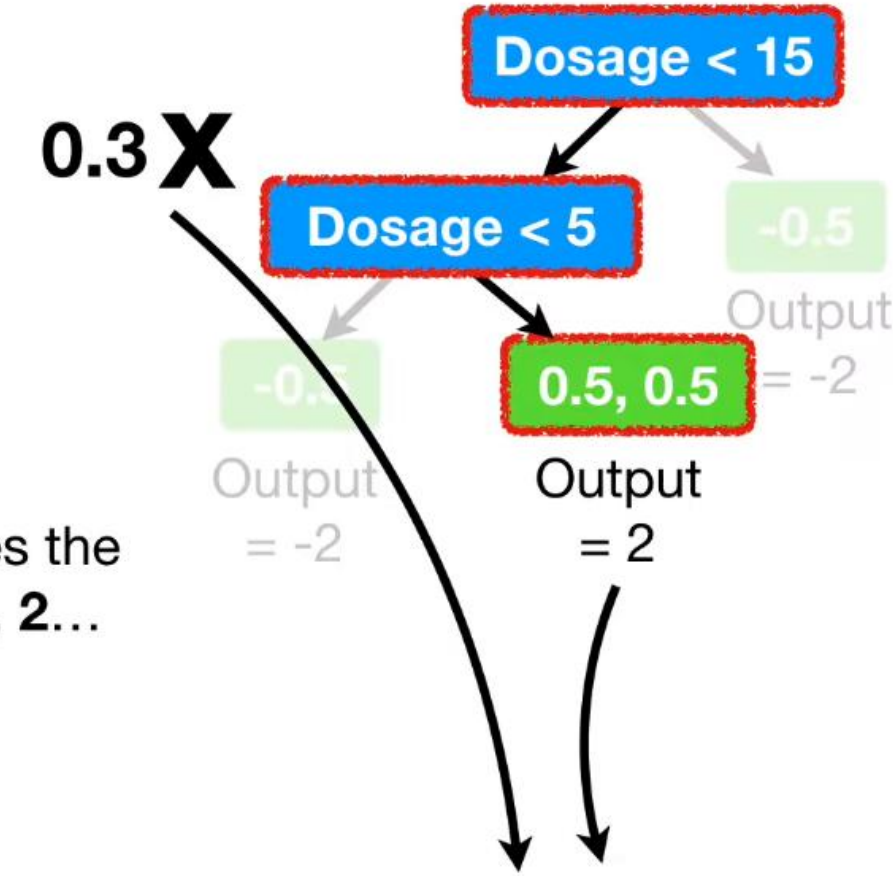
0.5 +

Output = log(odds) = 0



…plus **0.3** times the **Output Value**, **2**…

0.3 X



log(odds) Prediction = 0 + (0.3 x 2)

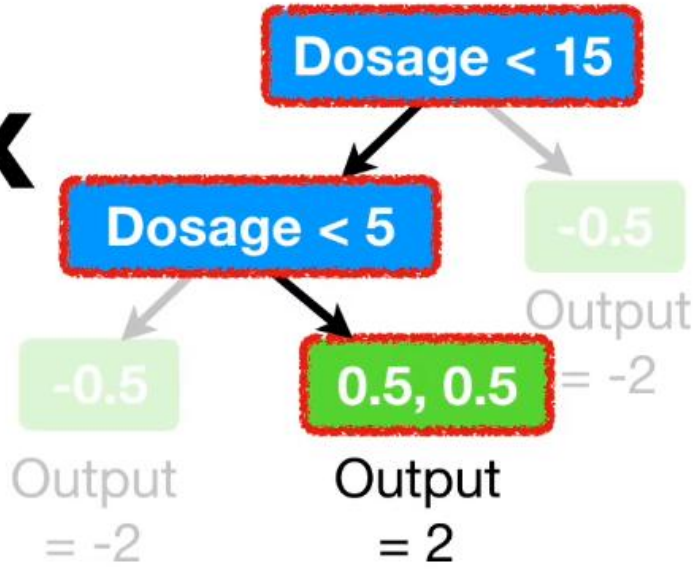
Predicted Drug Effectiveness

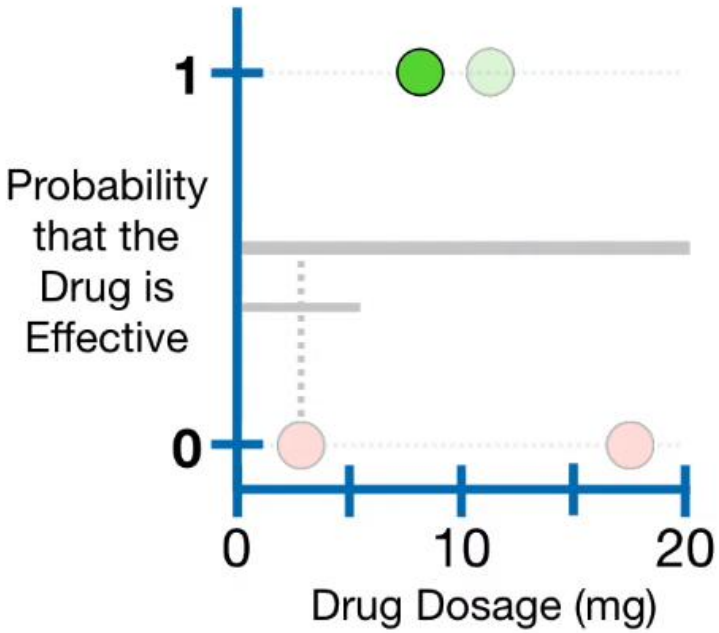
0.5

+

Output = log(odds) = 0

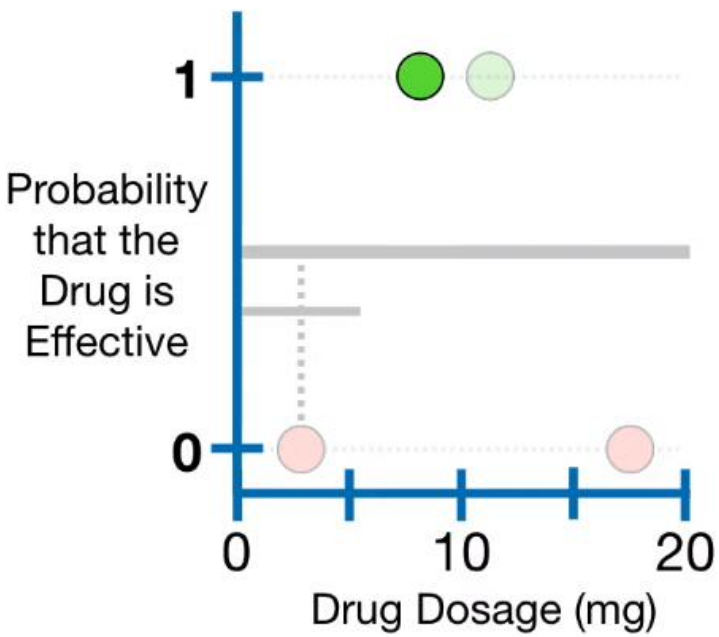
0.3 X





…and the predicted **log(odds) = 0.6**.

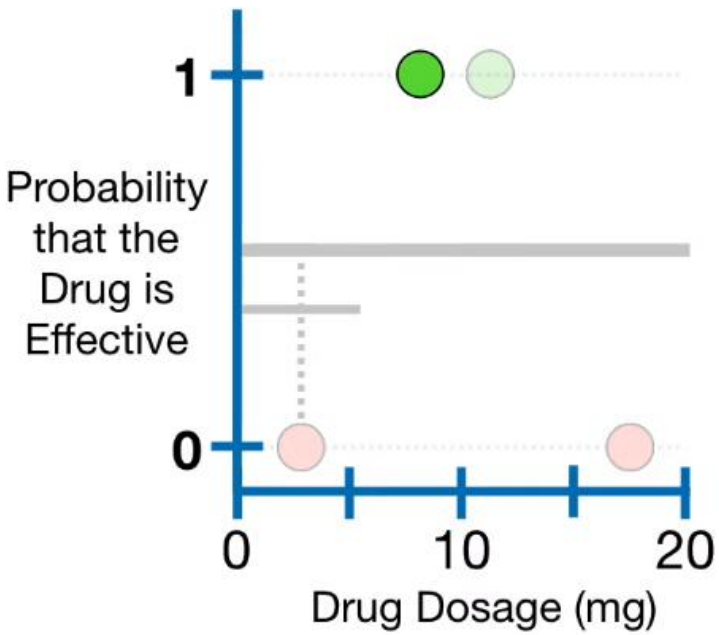
log(odds) Prediction = 0 + (0.3 × 2) = 0.6



Now we convert the **log(odds)** to a probability…

$$\text{Probability} = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}$$

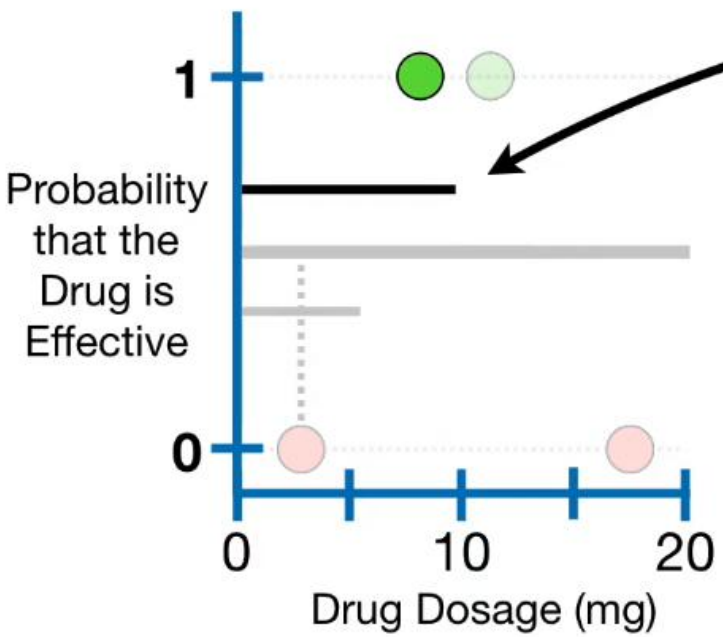
$$\log(\text{odds}) \text{ Prediction} = 0 + (0.3 \times 2) = 0.6$$



Now we convert the **log(odds)** to a probability…

$$\text{Probability} = \frac{e^{0.6}}{1 + e^{0.6}} = 0.65$$

$$\log(\text{odds}) \text{ Prediction} = 0 + (0.3 \times 2) = 0.6$$



…and we get **0.65**.

$$\text{Probability} = \frac{e^{0.6}}{1 + e^{0.6}} = 0.65$$

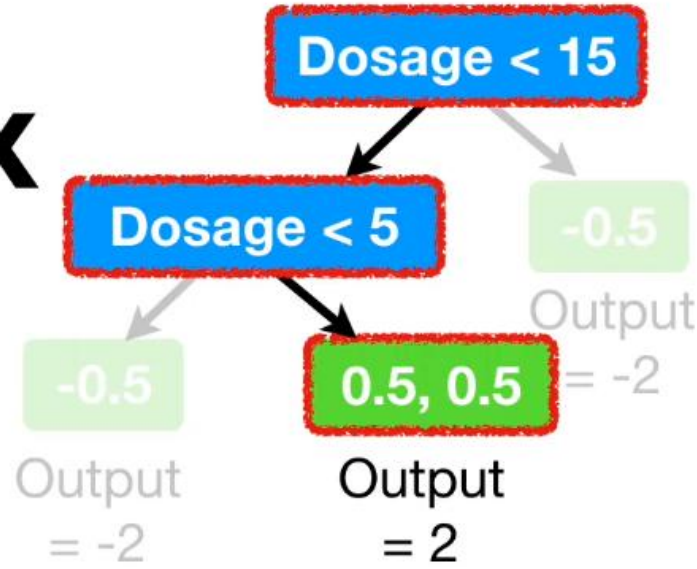
$$\log(\text{odds}) \text{ Prediction} = 0 + (0.3 \times 2) = 0.6$$

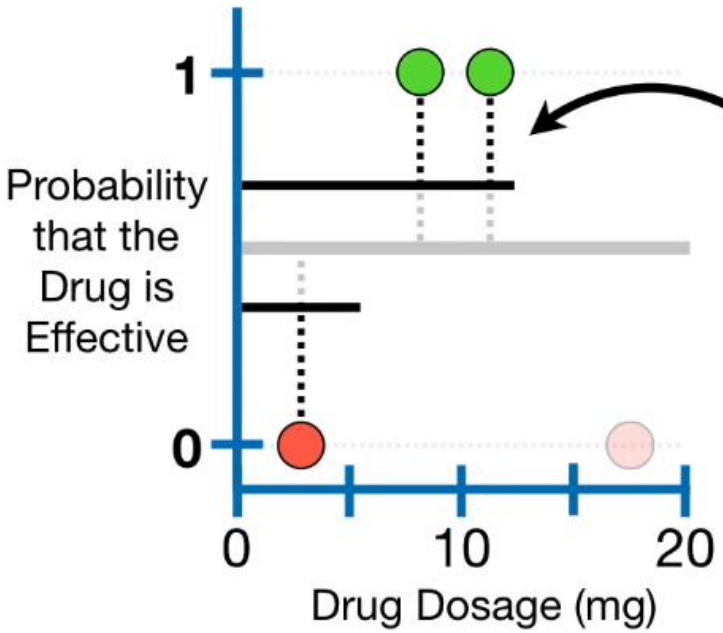
Predicted Drug Effectiveness

0.5 +

Output = log(odds) = 0

0.3 X





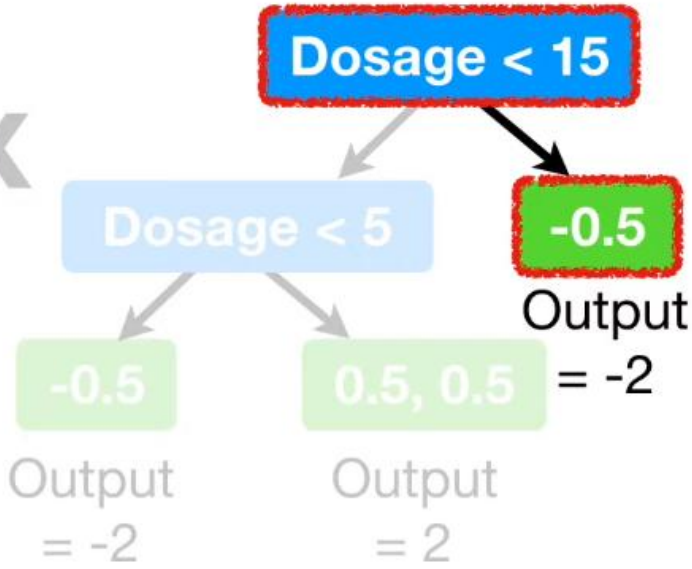
Likewise, the new predictions for the remaining observations have smaller **Residuals** than before.

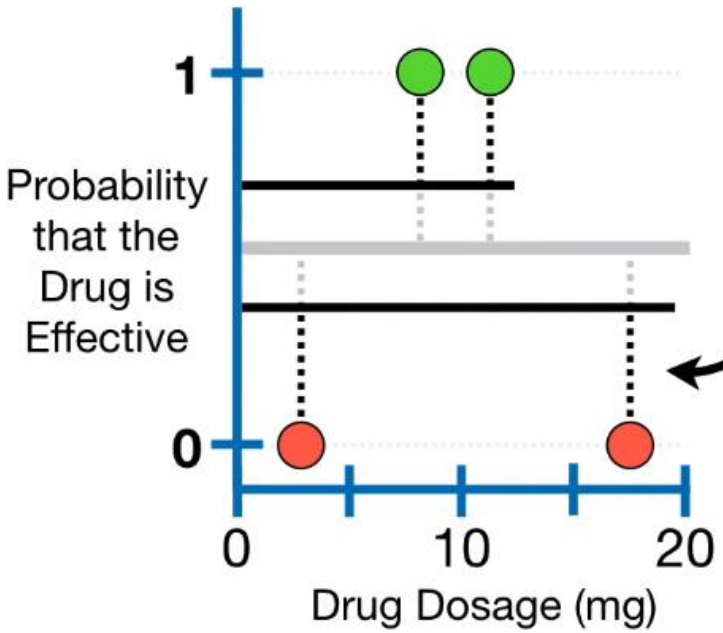
Predicted Drug Effectiveness

0.5 +

Output = log(odds) = 0

0.3 X





Likewise, the new predictions for the remaining observations have smaller **Residuals** than before.

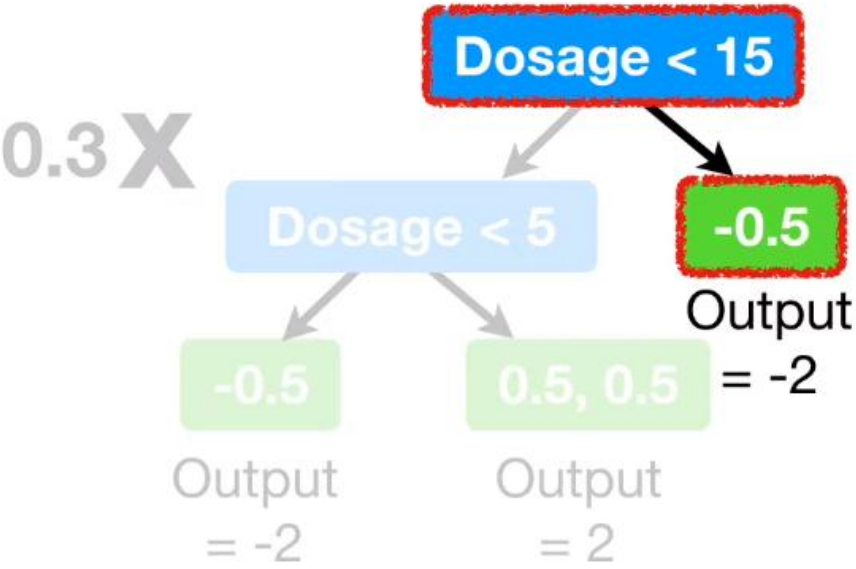
Predicted Drug Effectiveness

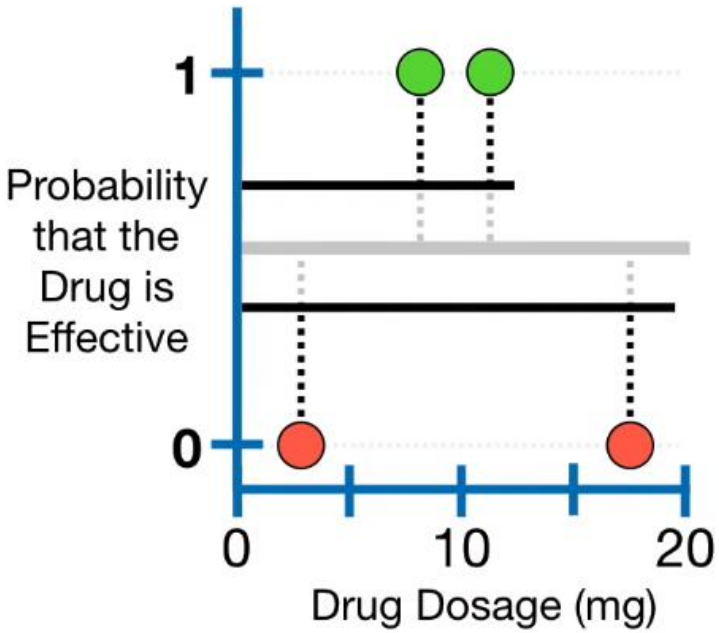
0.5

+

Output = log(odds) = 0

0.3 X





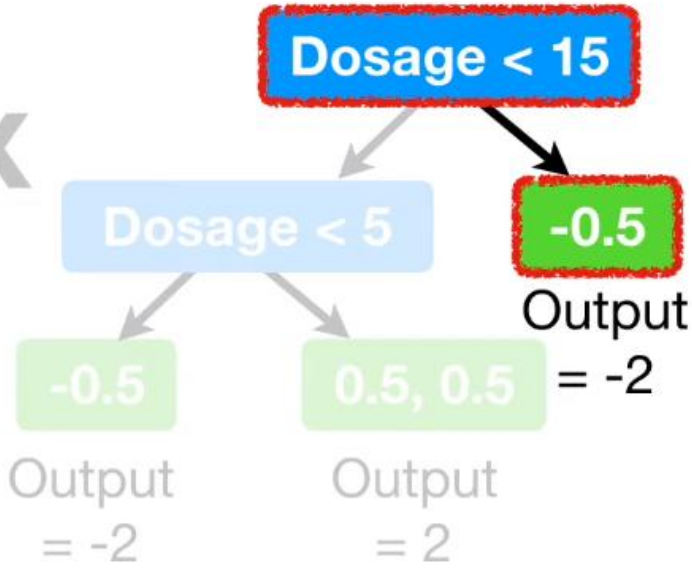
BAM!!!

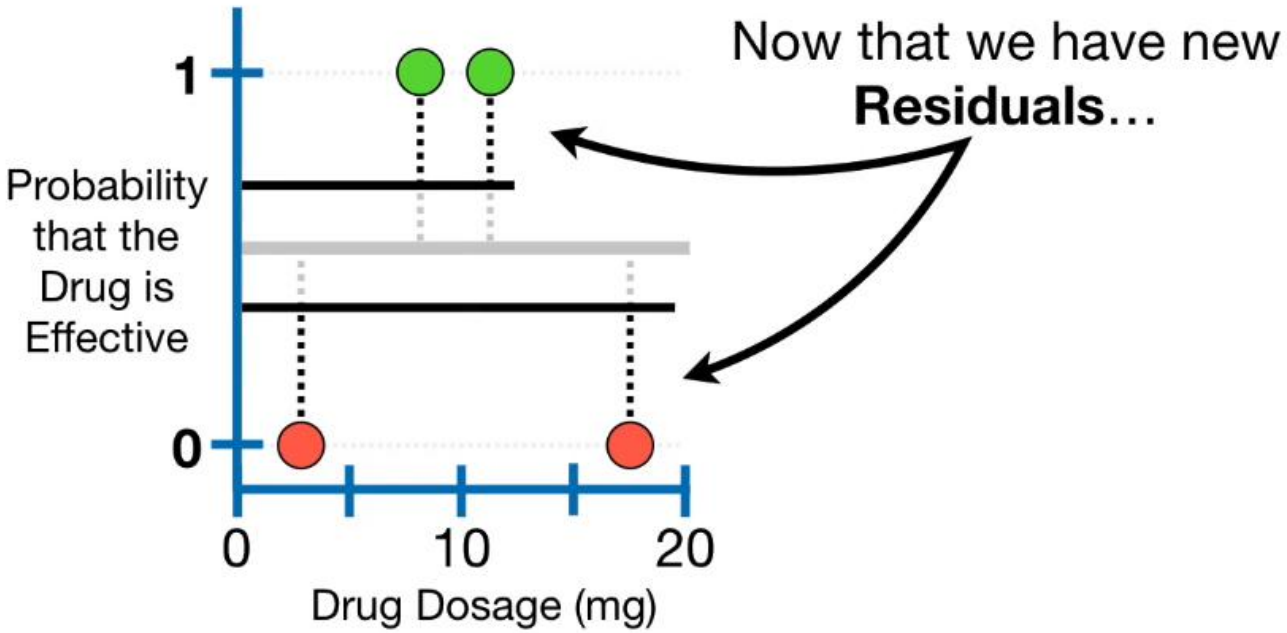
Predicted Drug Effectiveness

0.5 +

Output = log(odds) = 0

0.3 X





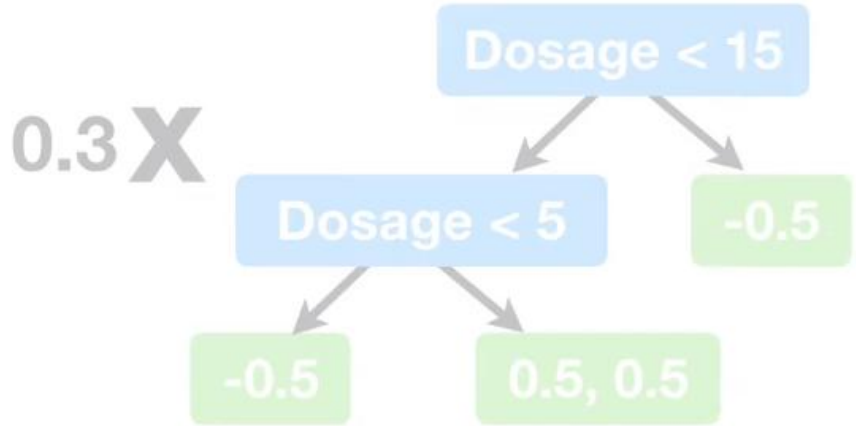
Predicted Drug Effectiveness

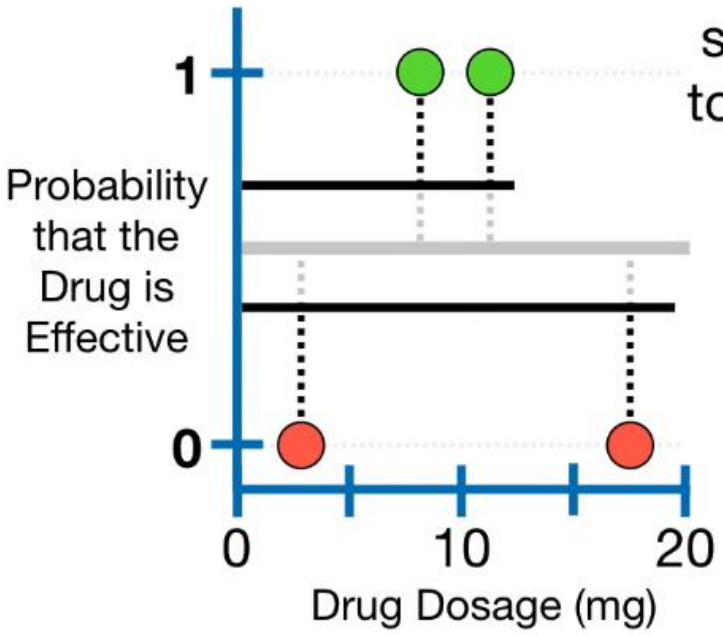
0.5

+

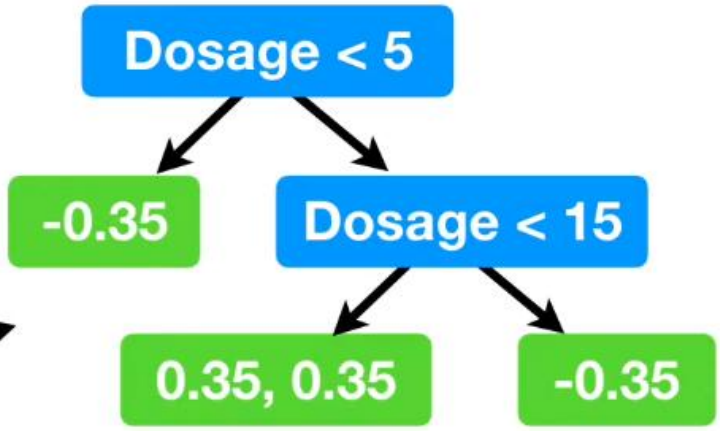
Output = log(odds) = 0

0.3 X



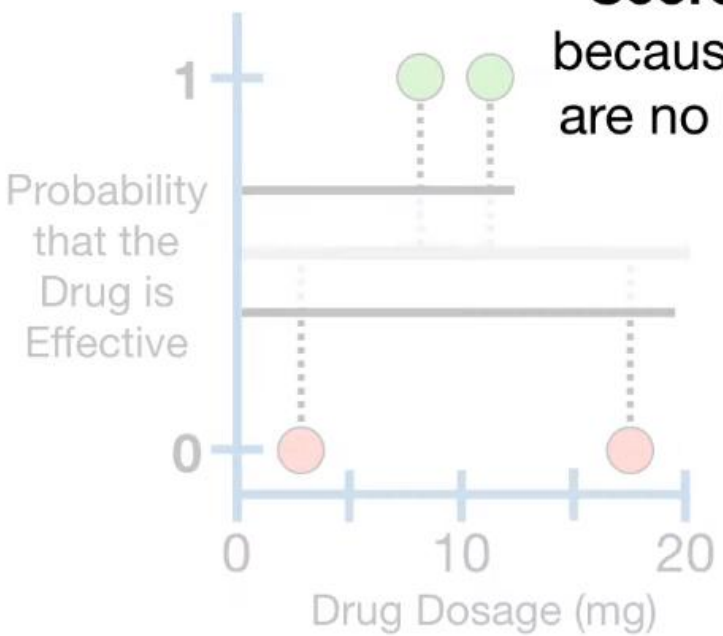


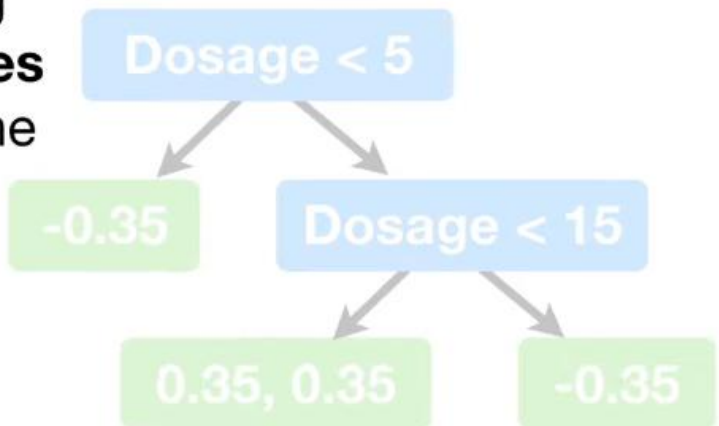
…we can build a second tree that is fit to the new **Residuals**.



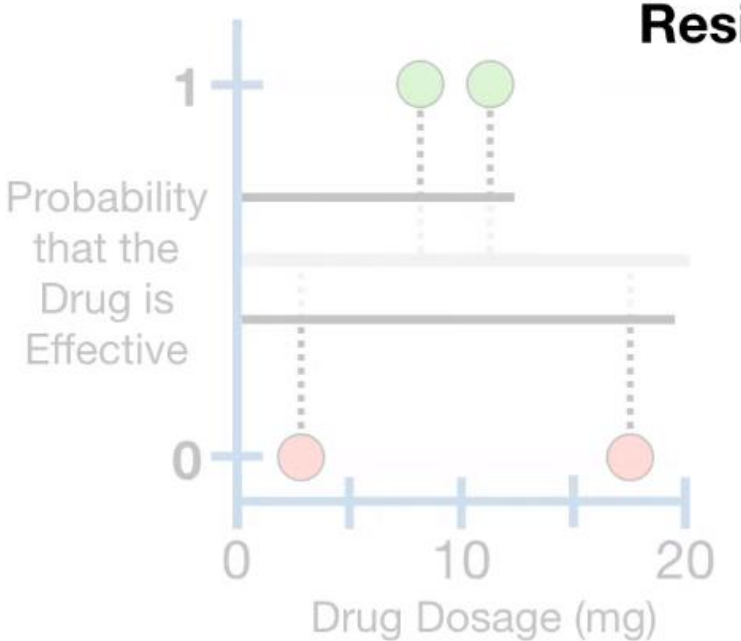
$$\text{Similarity Score} = \frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

NOTE: When we build the second tree, calculating the **Similarity Scores** is a little more interesting because the **Previous Probabilities** are no longer the same for all of the observations.

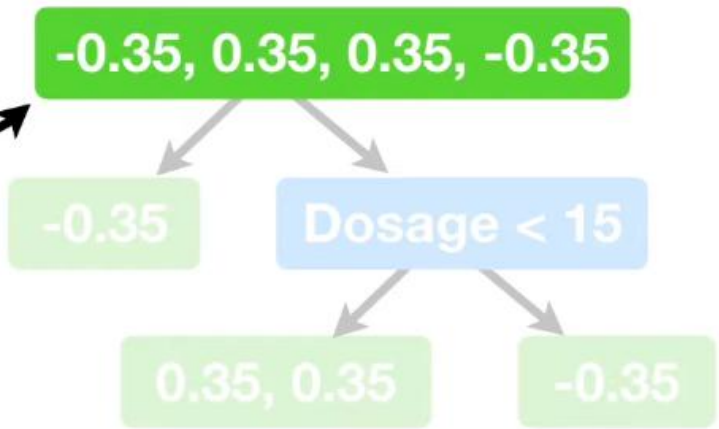




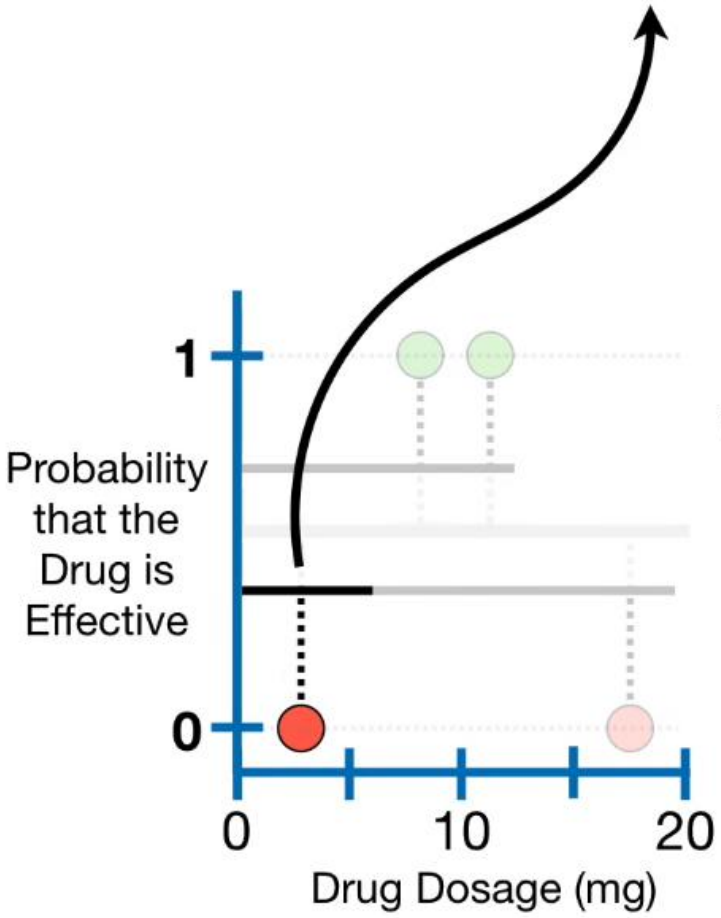
$$\text{Similarity Score} = \frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$



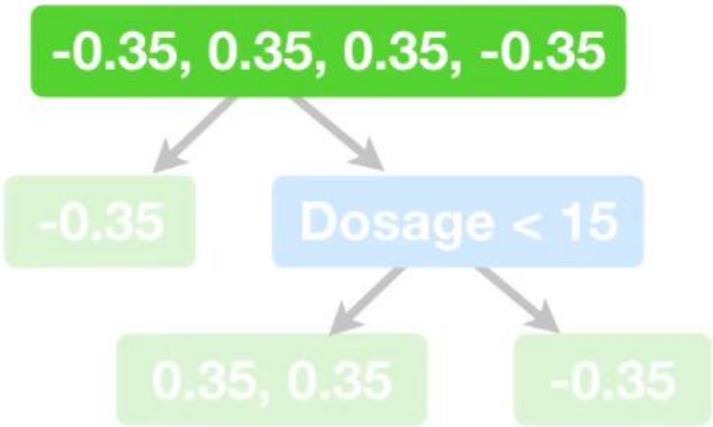
For example, since all of the **Residuals** start in the root of the tree…



$$\text{Similarity Score} = \frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

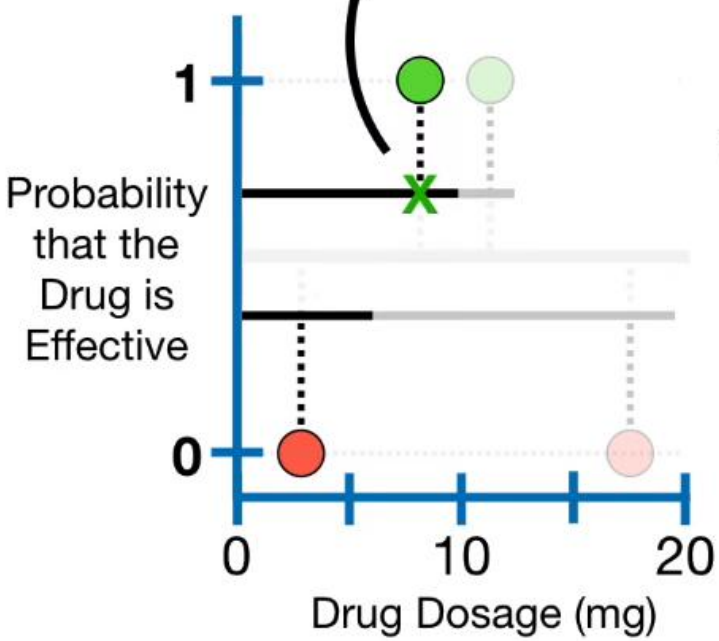


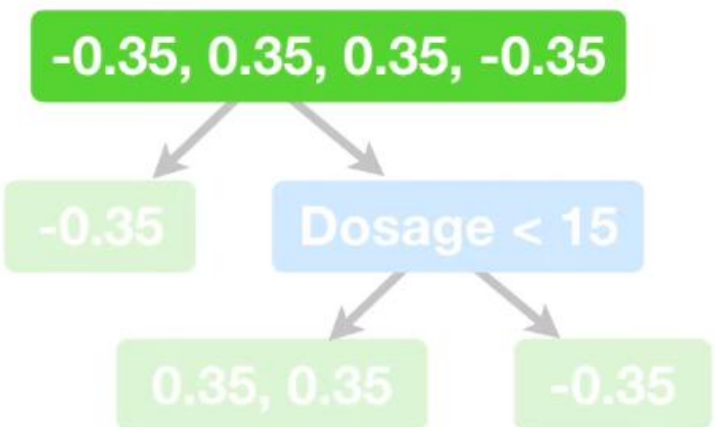
…we would plug in the previously predicted probabilities for each observation into the denominator, and this time they are not all the same.



$$\text{Similarity Score} = \frac{(\sum \text{Residual}_i)^2}{(0.35 \times (1-0.35)) + (0.65 \times (1-0.65))}$$

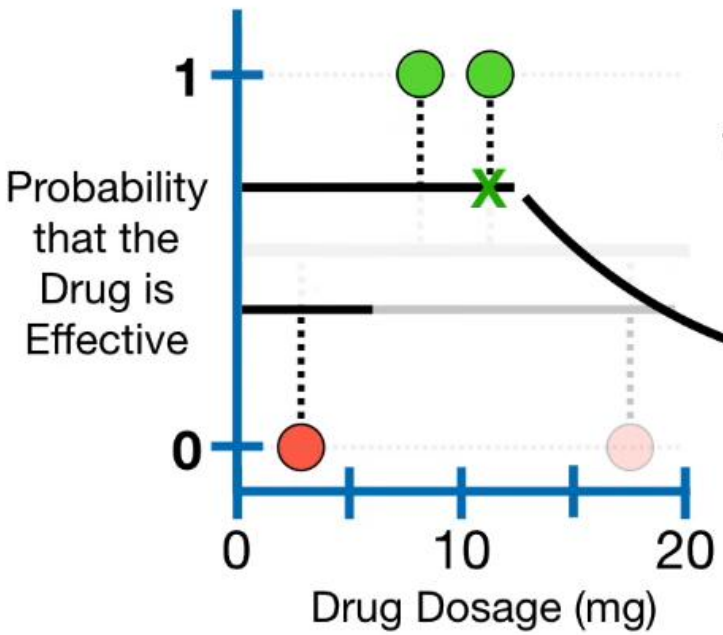
…we would plug in the previously predicted probabilities for each observation into the denominator, and this time they are not all the same.

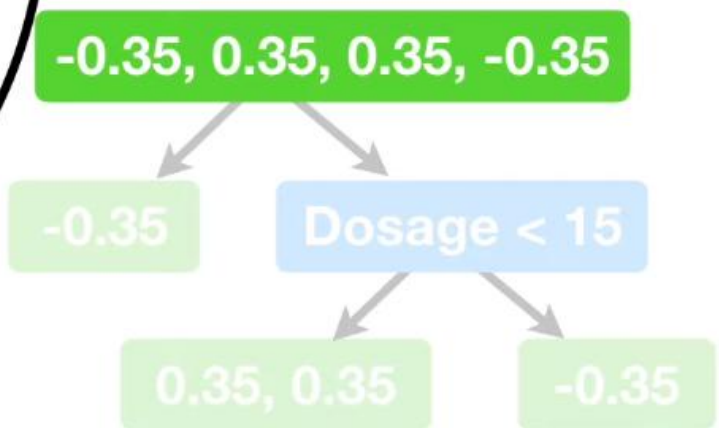




$$\text{Similarity Score} = \frac{(\sum \text{Residual}_i)^2}{(0.35 \times (1-0.35)) + (0.65 \times (1-0.65)) + (0.65 \times (1-0.65))}$$

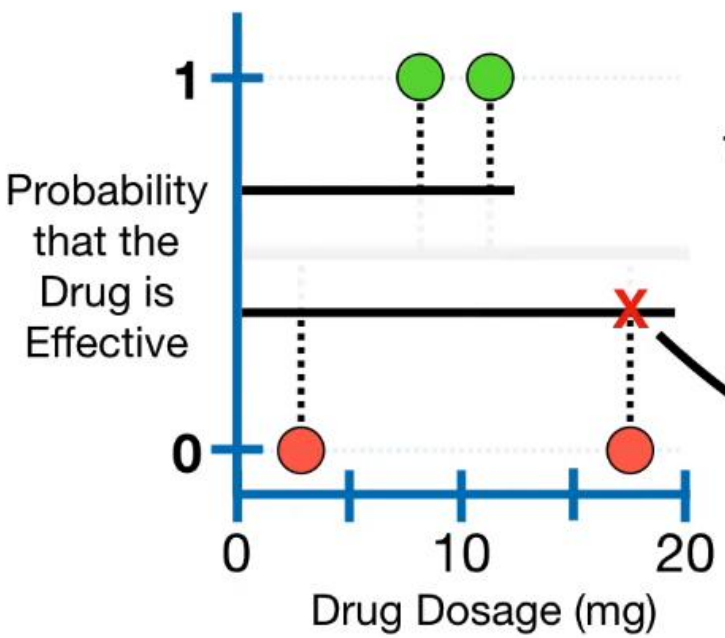
…we would plug in the previously predicted probabilities for each observation into the denominator, and this time they are not all the same.

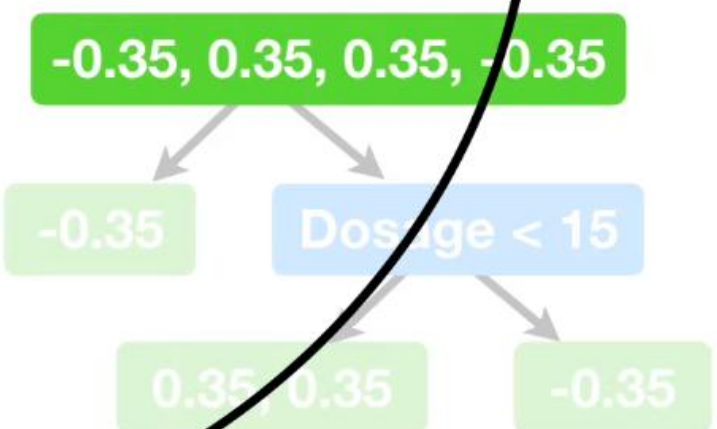




$$\text{Similarity Score} = \frac{(\sum \text{Residual}_i)^2}{(0.35 \times (1-0.35)) + (0.65 \times (1-0.65)) + (0.65 \times (1-0.65)) + (0.35 \times (1-0.35)) + \lambda}$$

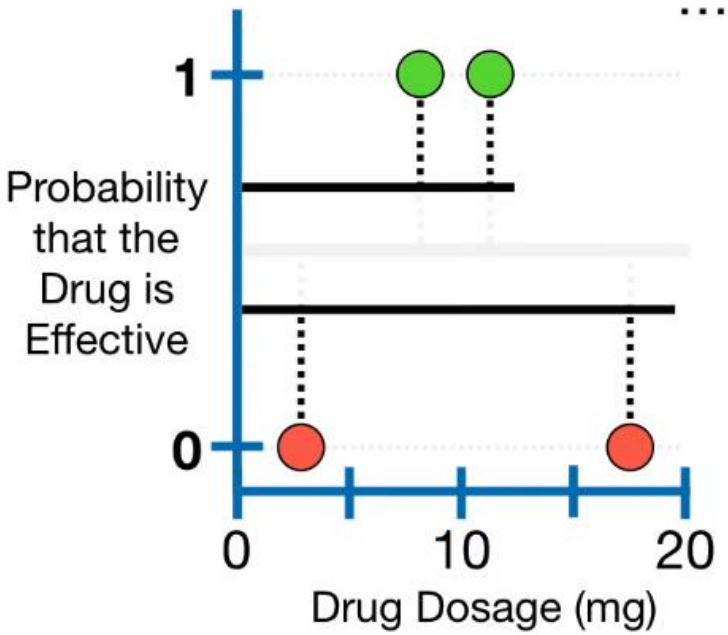
…we would plug in the previously predicted probabilities for each observation into the denominator, and this time they are not all the same.

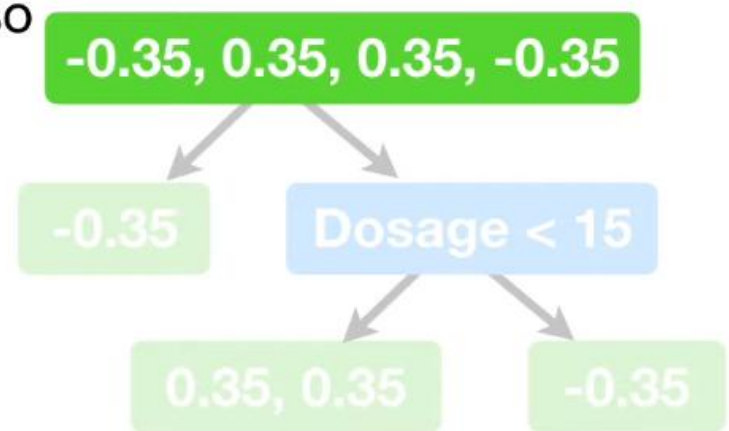




$$\text{Output Value} = \frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

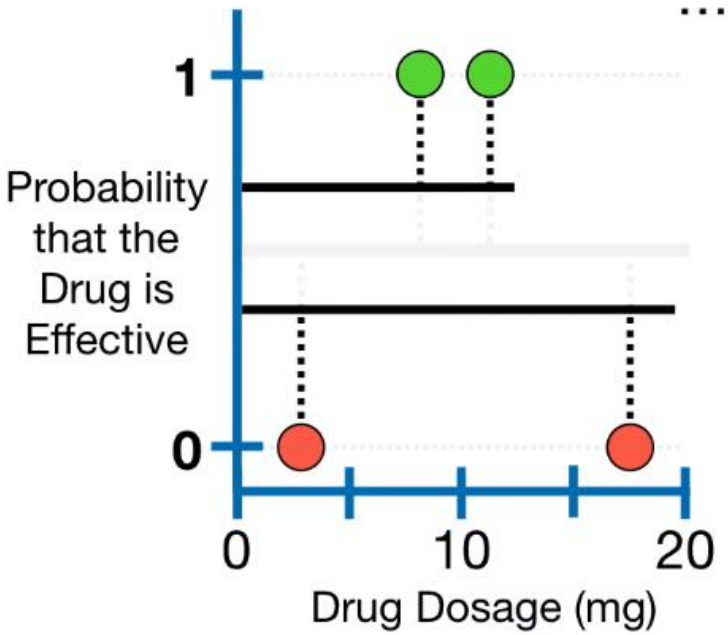
…the denominator would also contain a mixture of previously predicted probabilities.

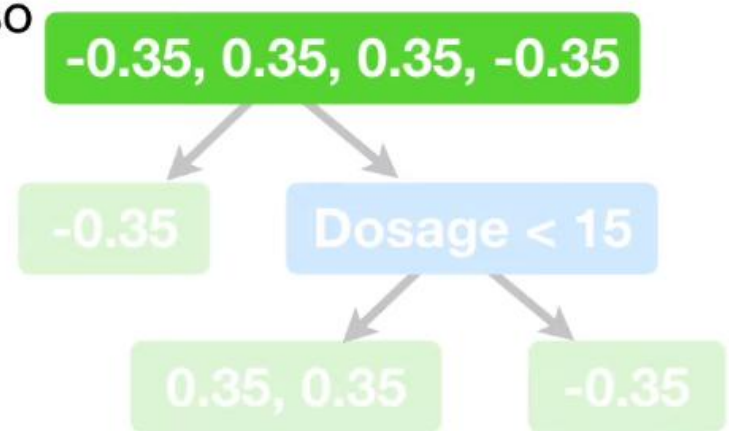




$$\text{Output Value} = \frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

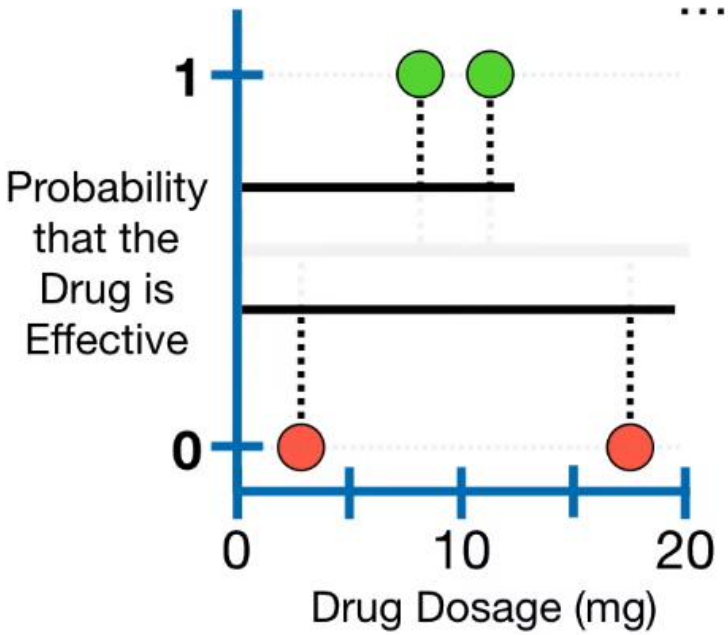
…the denominator would also contain a mixture of previously predicted probabilities.

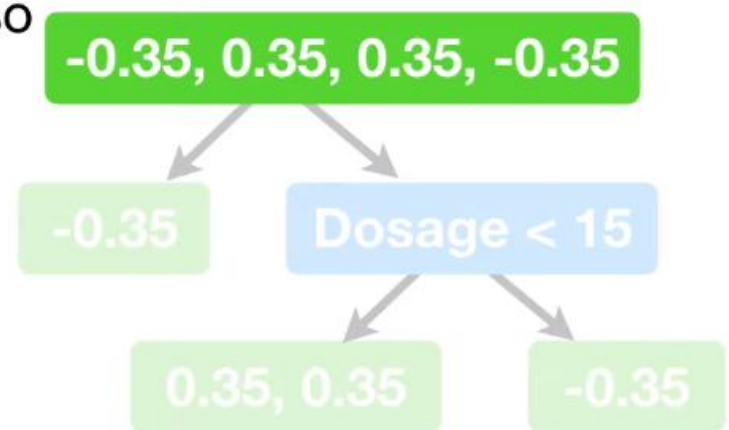




$$\text{Output Value} = \frac{\left(\sum \text{Residual}_i\right)}{(0.35 \times (1-0.35)) + (0.65 \times (1-0.65)) + (0.65 \times (1-0.65)) + (0.35 \times (1-0.35)) + \lambda}$$

…the denominator would also contain a mixture of previously predicted probabilities.





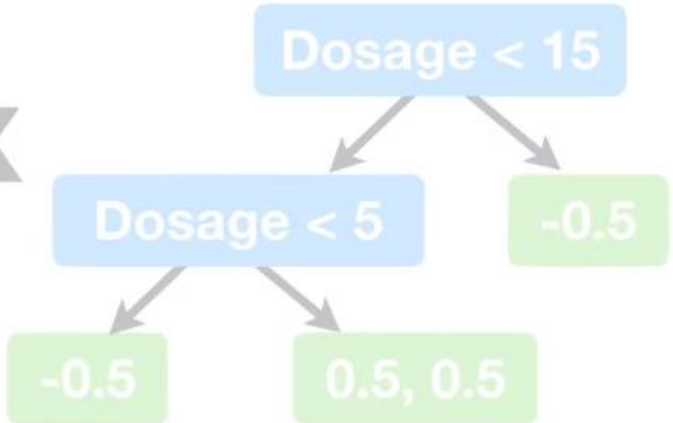
Predicted Drug Effectiveness

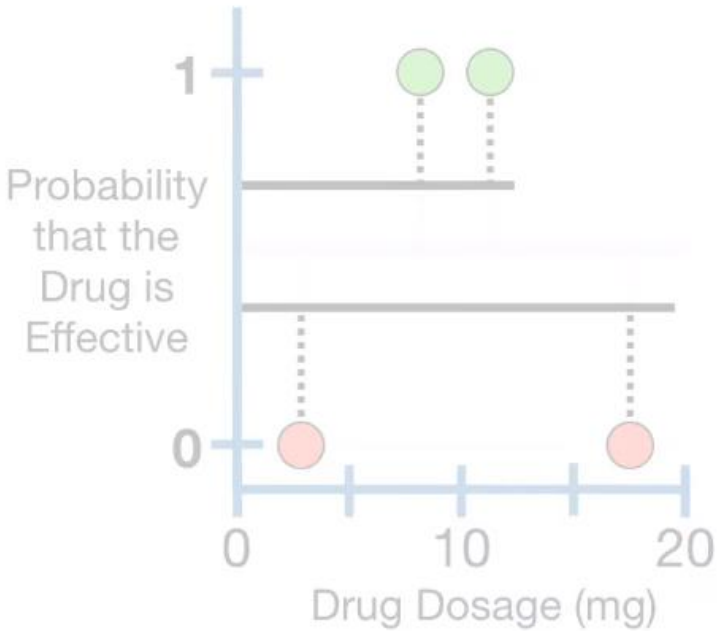
0.5

+

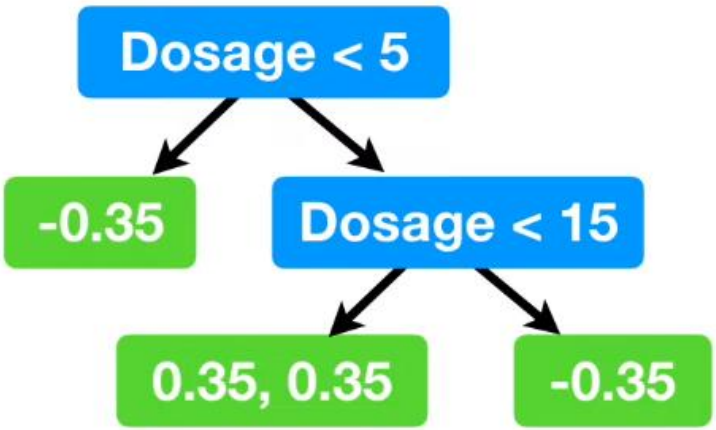
Output = log(odds) = 0

0.3 X





Now that we have a new tree…



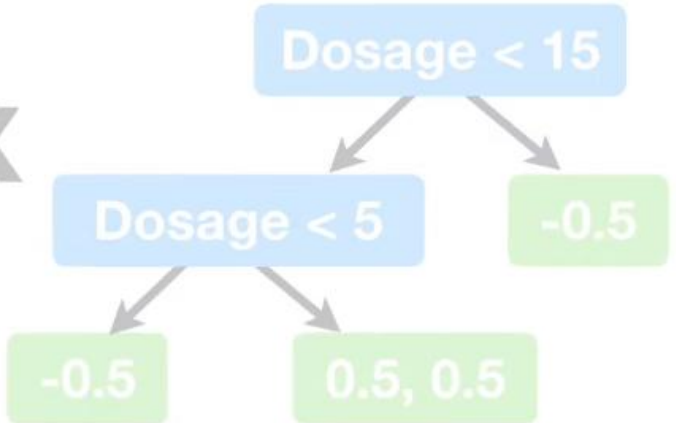
Predicted Drug Effectiveness

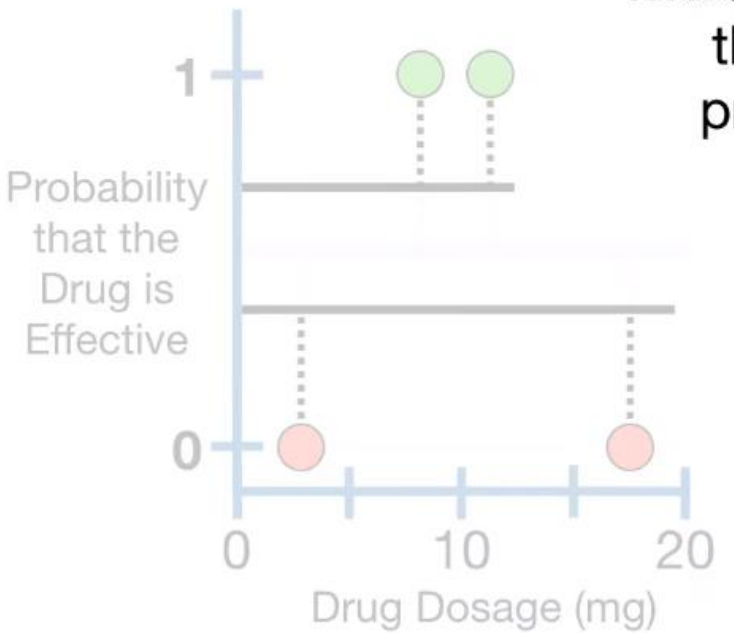
0.5

+

Output = log(odds) = 0

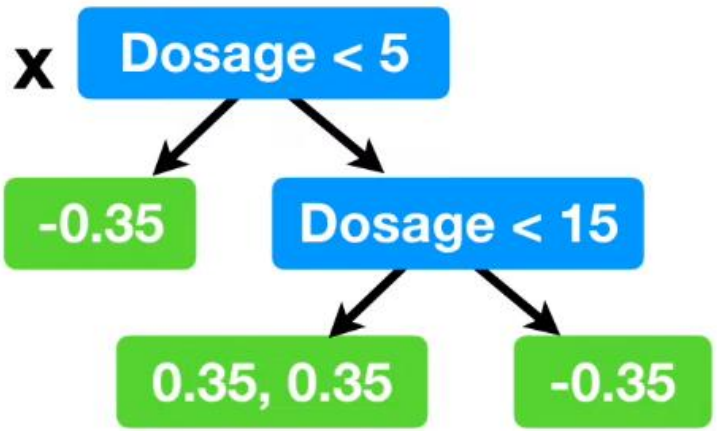
0.3 X





…we add it to all of the previous predictions…

+ 0.3 x



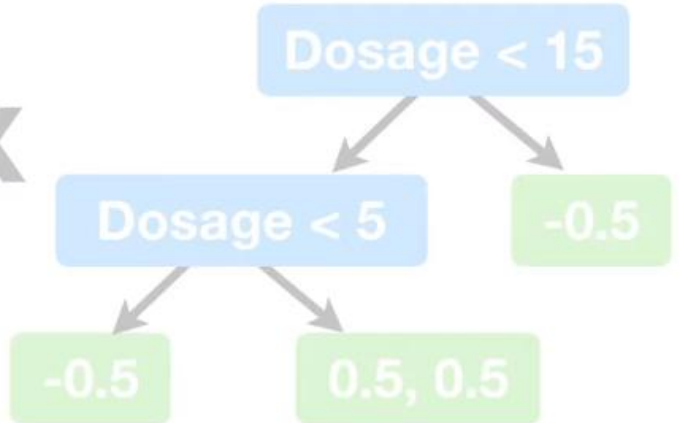
Predicted Drug Effectiveness

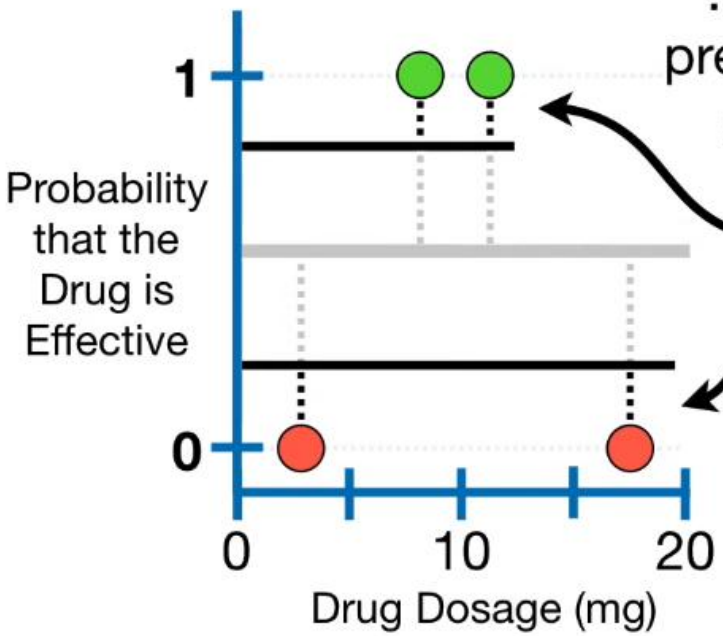
0.5

+

Output = log(odds) = 0

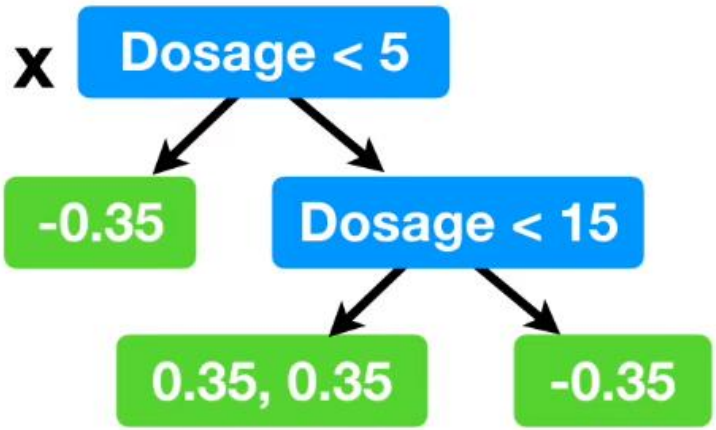
0.3 X





…and make new predictions that give us even smaller **Residuals**.

+ 0.3 x



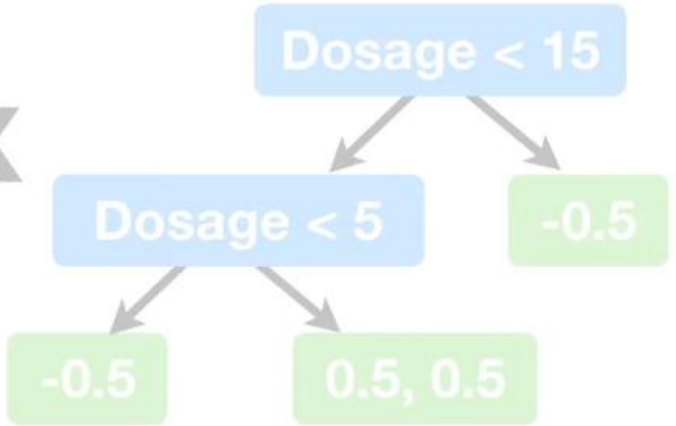
Predicted Drug Effectiveness

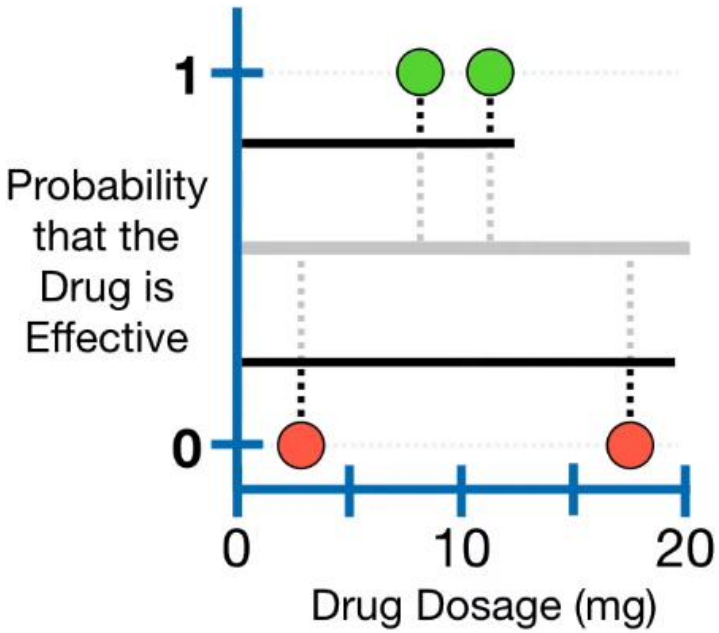
0.5

+

Output = log(odds) = 0

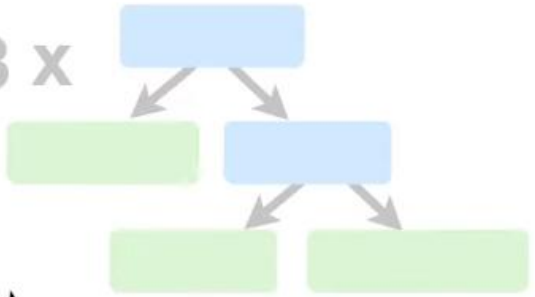
0.3 X



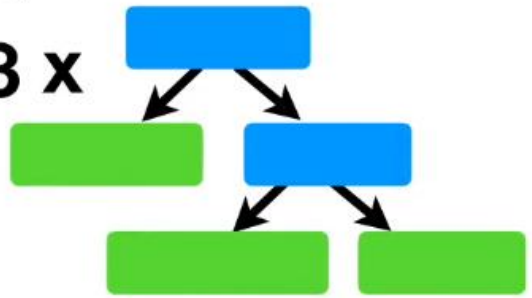


Then we build another tree based on the new **Residuals**…

+ 0.3 x



+ 0.3 x

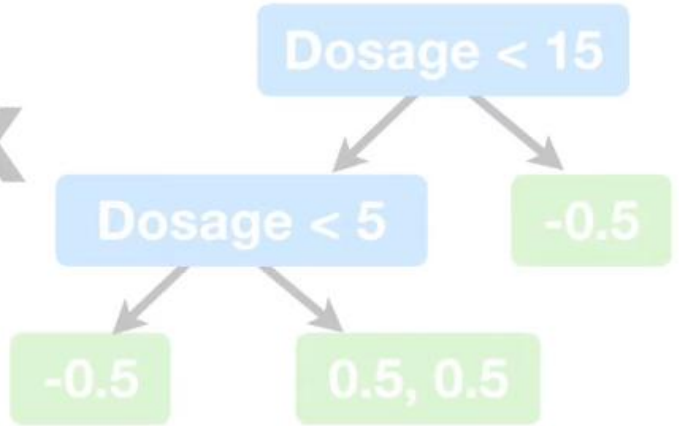


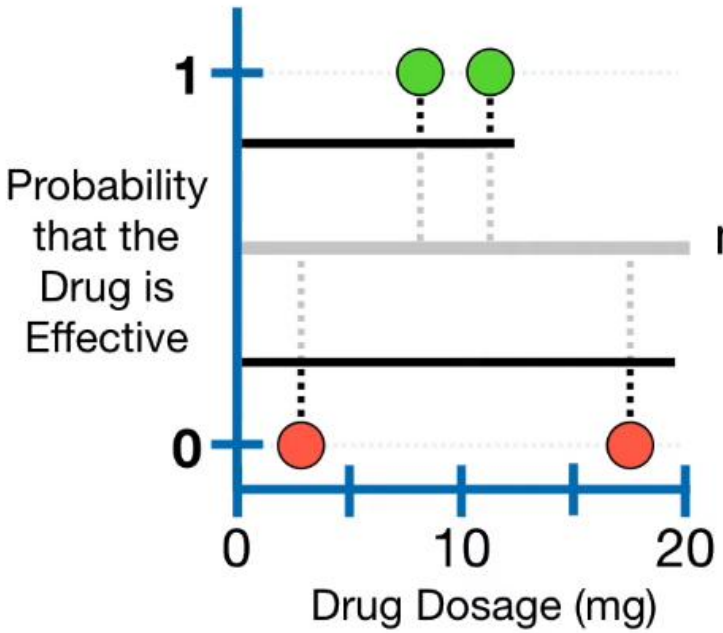
Predicted Drug Effectiveness

0.5 +

Output = log(odds) = 0

0.3 X

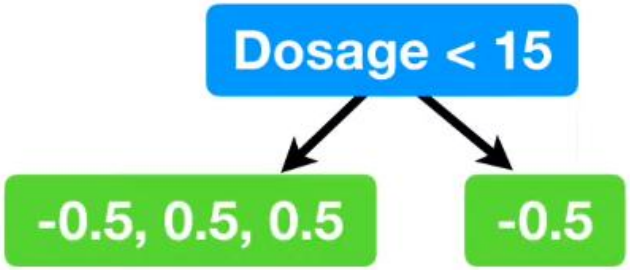




…and we keep building trees until the **Residuals** are super small, or we have reached the maximum number of trees.


Predicted Drug Effectiveness
0.5
+
Output = log(odds) = 0
0.3 X
Dosage < 15
Dosage < 5
-0.5
-0.5
0.5, 0.5
BAM!!!!!!!!!!!!
Proba
that
Drug is
Effective
0
0
10
20
Drug Dosage (mg)
+ 0.3 x

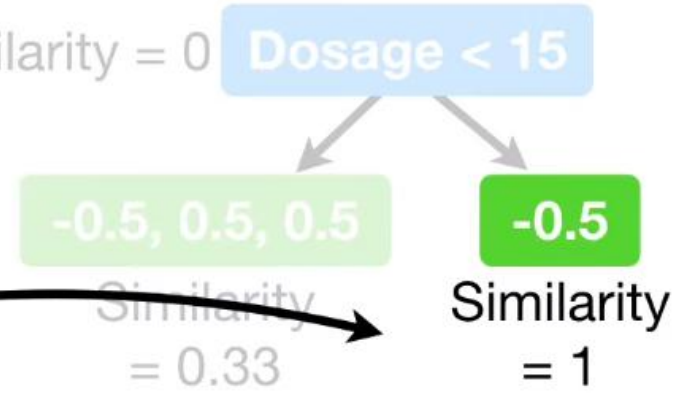
In summary, when building **XGBoost Trees** for **Classification**…



…we calculate **Similarity Scores**…

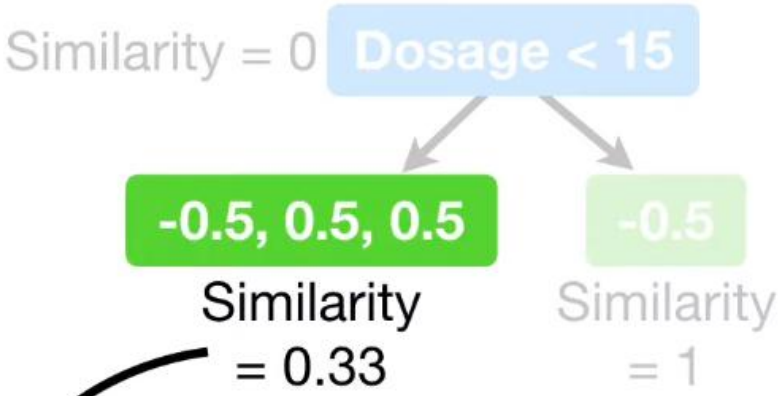
$$\text{Similarity Score} = \frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

…we calculate **Similarity Scores**…



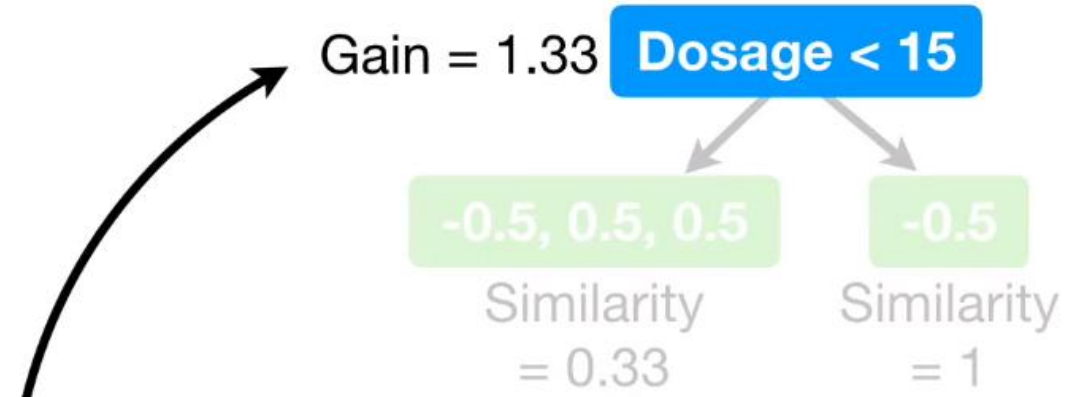
$$\text{Similarity Score} = \frac{(\sum \text{Residual}_i)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

… and **Gain** to determine how to split the data…



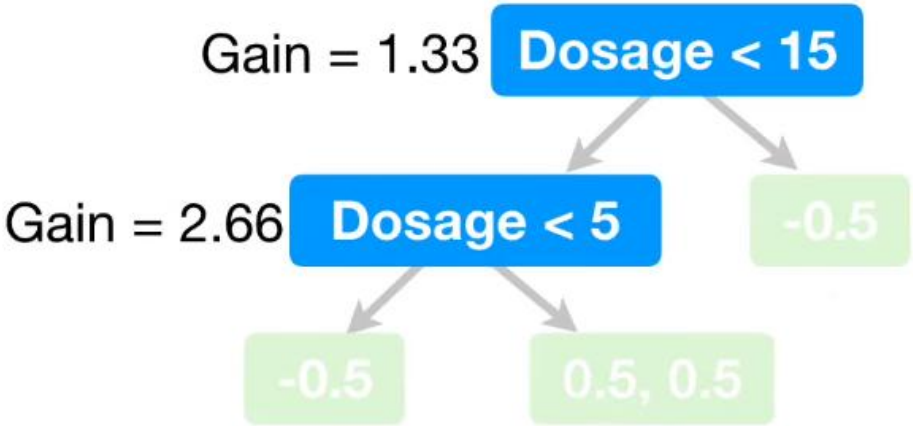
$$\text{Gain} = \text{Left}_{\text{Similarity}} + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

… and **Gain** to determine how to split the data…



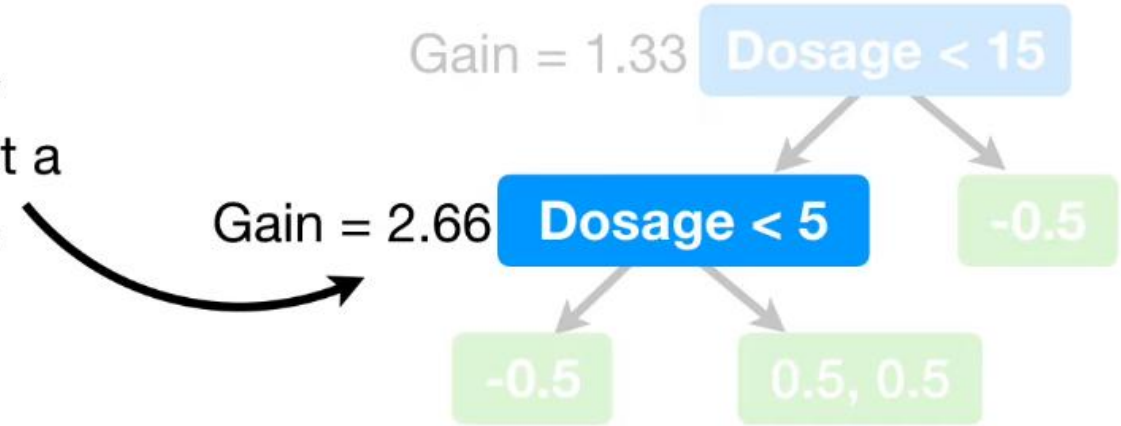
$$\text{Gain} = \text{Left}_{\text{Similarity}} + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

…and we prune the tree by calculating the difference between **Gain** values and a user defined **Tree Complexity Parameter**, ***γ*** **(gamma)**.



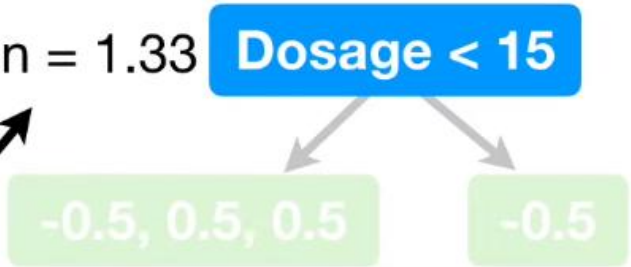
Gain - ***γ*** = { If positive, then do not prune.
If negative, then prune.

For example, if we subtract γ **(gamma)** from this **Gain** and get a negative value, we will prune, otherwise we're done.



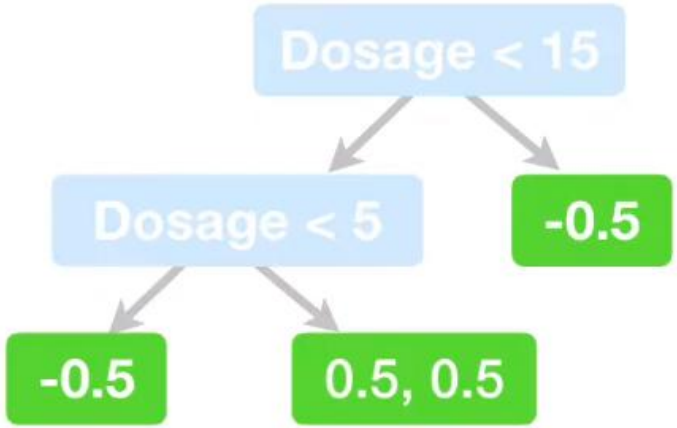
$$\text{Gain} - \gamma = \begin{cases} \text{If positive, then do not prune.} \\ \\ \text{If negative, then prune.} \end{cases}$$

If we prune, then we will subtract γ **(gamma)** from the next **Gain** value (etc. etc. etc.).



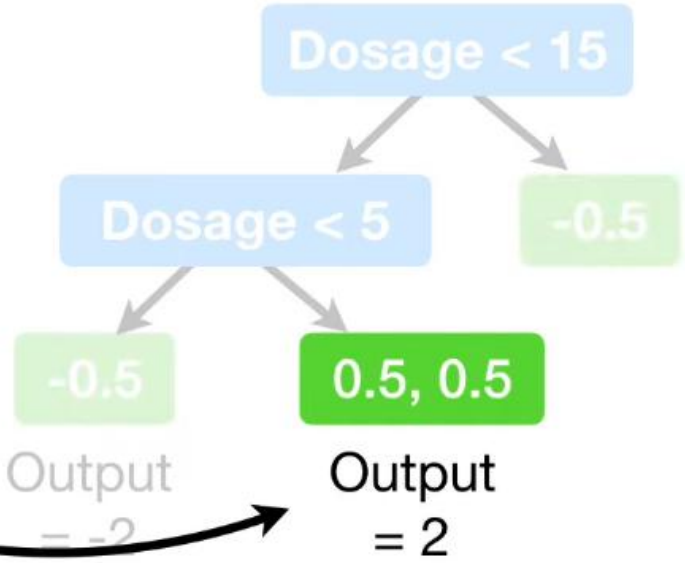
$$\text{Gain} - \gamma = \begin{cases} \text{If positive, then do not prune.} \\ \\ \text{If negative, then prune.} \end{cases}$$

Then we calculate the **Output Values** for the leaves…



$$\text{Output Value} = \frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

Then we calculate the **Output Values** for the leaves…



$$\text{Output Value} = \frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

…and lastly, **λ (lambda)** is a **Regularization Parameter** and when **λ > 0**, it results in more pruning, by shrinking the **Similarity Scores**, and smaller **Output Values** for the leaves.

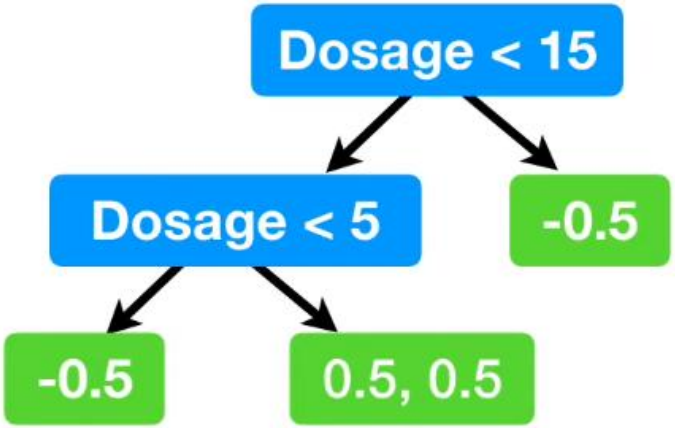
…and lastly, **λ (lambda)** is a **Regularization Parameter** and when **λ > 0**, it results in more pruning, by shrinking the **Similarity Scores**, and smaller **Output Values** for the leaves.

$$\text{Similarity Score} = \frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

…and lastly, **λ (lambda)** is a **Regularization Parameter** and when **λ > 0**, it results in more pruning, by shrinking the **Similarity Scores**, and smaller **Output Values** for the leaves.

$$\text{Similarity Score} = \frac{\left(\sum \text{Residual}_i\right)^2}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$

$$\text{Output Value} = \frac{\left(\sum \text{Residual}_i\right)}{\sum \left[\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)\right] + \lambda}$$



Oh, and I almost forgot, when using **XGBoost** for **Classification** we have to be aware that the minimum number of **Residuals** in a leaf is related to a metric called **Cover**, which is the denominator of the **Similarity Score**, minus **λ (lambda)**.

$$\text{Similarity Score} = \frac{(\sum \text{Residual}_i)^2}{\text{Cover} + \lambda}$$

$$\text{Cover} = \sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)]$$