Random Forest

Random Forest is an ensemble learning method that combines many decision trees. The math behind it comes from bootstrap sampling, feature randomness, and majority voting. These ideas reduce variance and create a stronger, more stable model.

1. Bootstrap Sampling (Bagging)

Each tree is trained on a random sample of the dataset created by sampling with replacement.

If original dataset is:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

A bootstrap sample D_b is generated by:

$$D_b = \{(x_i, y_i), \dots, (x_{i_m}, y_{i_m})\}$$

where each i_k is sampled from $\{1, \dots, N\}$ with replacement.

Some points repeat, some disappear — this makes every tree see a different version of the world.

---

2. Feature Subsampling

When a tree tries to split a node, Random Forest restricts the possible features to a random subset.

If there are F total features, each split sees only:

$$k = \sqrt{F} \quad \text{(for classification)}$$

This creates diversity among trees because each tree uses different features at split.

---

3. Decision Tree Split Mathematics

Each trees uses the same math as a normal decision tree.

Entropy:

$$H(S) = -\sum_{c} p(c) \log_2 p(c)$$

Weighted Child Entropy:

$$H_{split} = \frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R)$$

Information Gain:

$$IG = H(S) - H_{split}$$

Every tree finds splits that maximize Information Gain.

Because every tree sees different samples and features, their splits differ naturally.

---

4. Ensemble Prediction (Majority Vote)

Once all trees make a prediction, the forest combines them.

For classification:

If T trees produce predictions:

$$h_1(x), h_2(x), \ldots, h_T(x)$$

The final prediction is the most common class:

$$H_{forest}(x) = \text{mode}(h_1(x), h_2(x), \ldots, h_T(x))$$

For regression:

Take the average:

$$H_{forest}(x) = \frac{1}{T} \sum_{i=1}^{T} h_i(x)$$

5. Why Random Forest Works (Mathematical Insight)

A single decision tree has low bias but high variance - it overfits.

Bagging reduces variance by averaging many noisy predictions.

Variance decreases roughly as:

$$\text{Var}_{\text{ensemble}} \approx \frac{\text{Var}_{\text{tree}}}{T}$$

(assuming trees are not too correlated)

Feature subsampling reduces correlation between trees, making this reduction stronger.

So even though each individual tree is a "weak learner", the forest becomes a strong model.

---

6. Out-of-Bag Error

Because each tree sees about 63% of the data (due to sampling with replacement), the remaining 37% can be used to estimate the model's performance without a separate test set.

This works because the probability a point is not chosen in a bootstrap sample is:

$$\left(1 - \frac{1}{N}\right)^N \approx e^{-1} \approx 0.37$$