Bagging 2

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai

Outline

- Review of Decision Trees
- Bagging
- Out of Bag Error (OOB)
- Variable Importance

Bagging

One way to adjust for the high variance of the output of an experiment is to perform the experiment multiple times and then average the results.

The same idea can be applied to high variance models:

1. **Bootstrap:** we generate multiple samples of training data, via bootstrapping. We train a deeper decision tree on each sample of data.

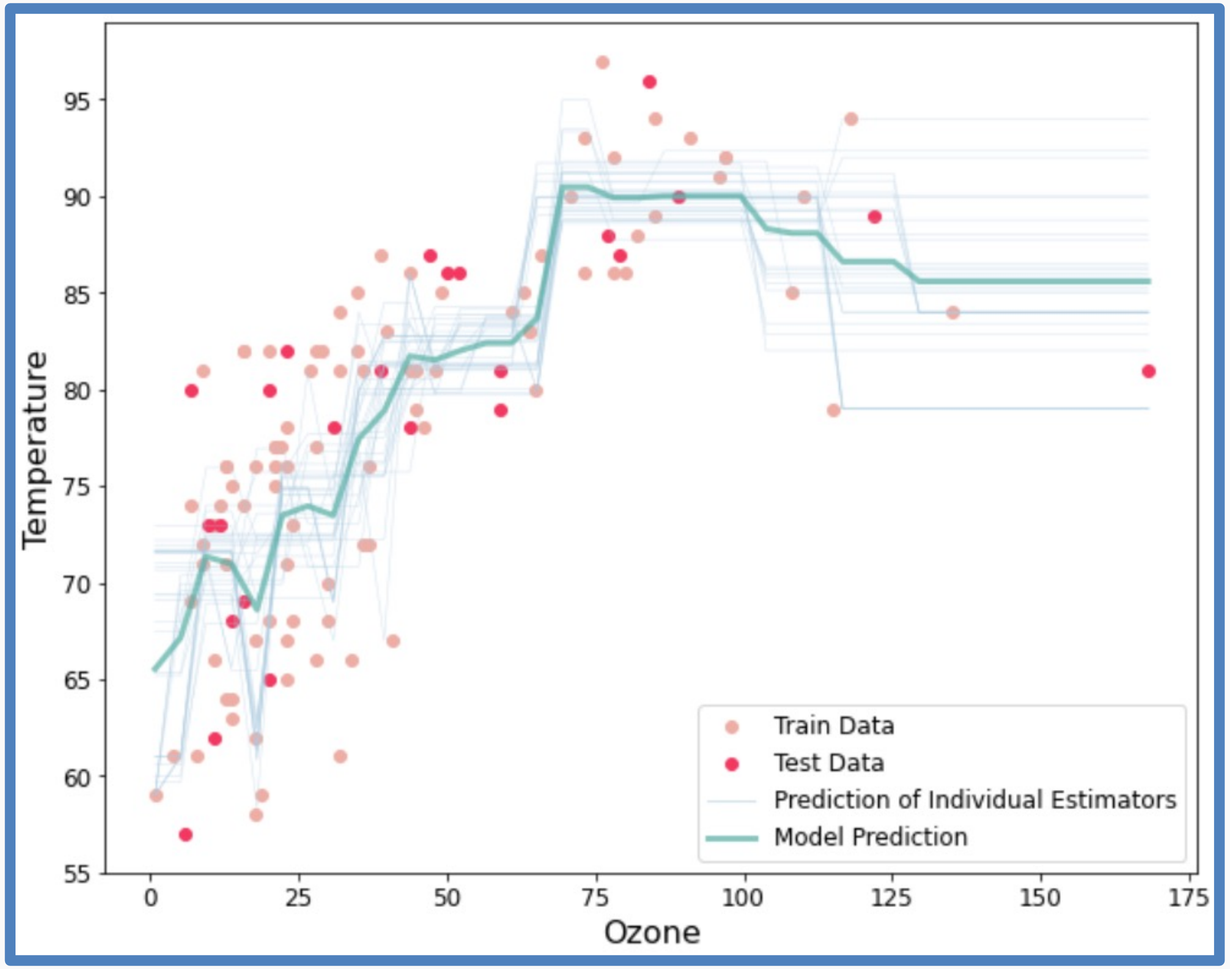
2. **Aggregate:** for a given input, we output the averaged outputs of all the models for that input.

This method is called ***Bagging*** (Breiman, 1996), short for, of course, Bootstrap Aggregating.

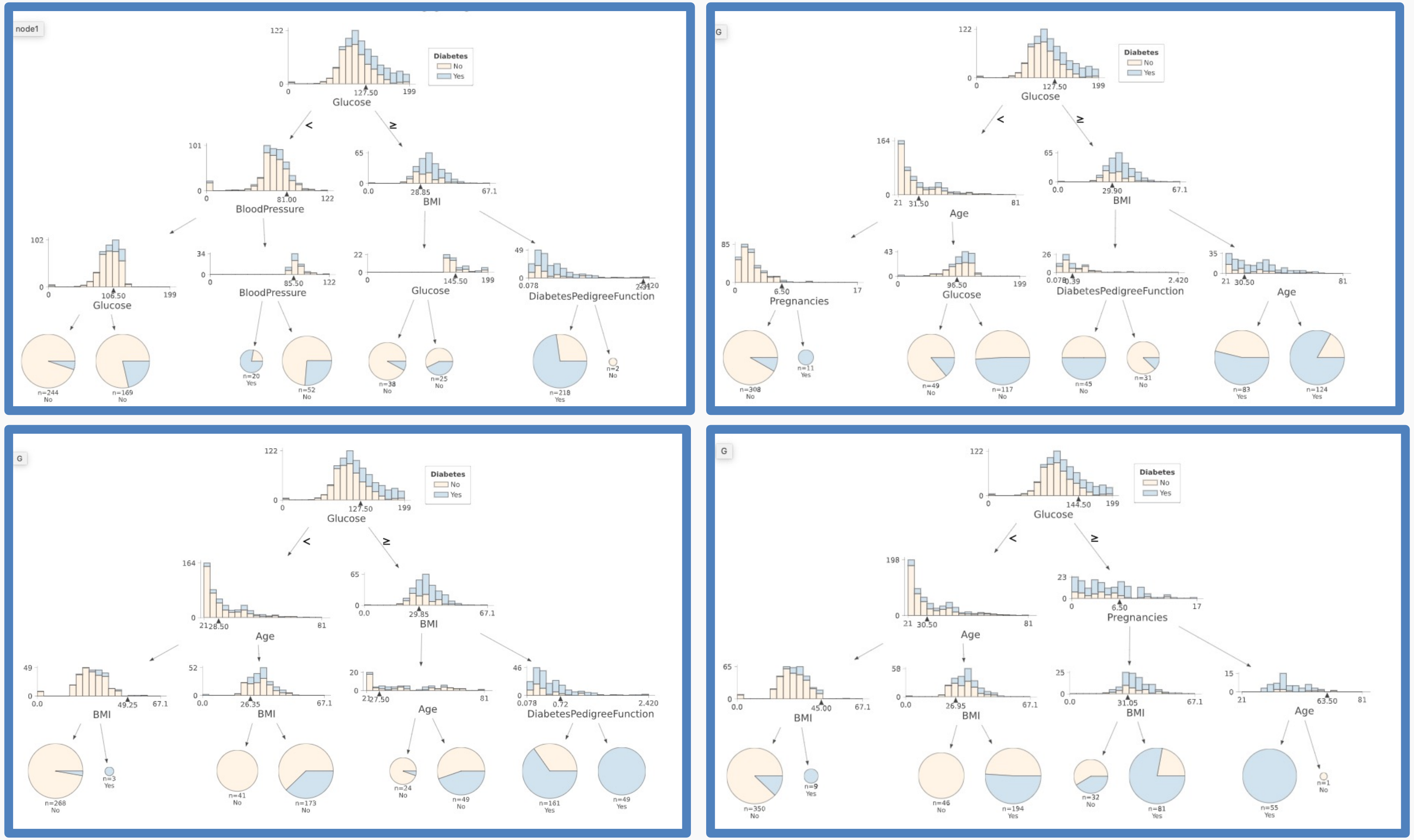
For classification, we return the class that is outputted by the plurality of the models. For regression we return the average of the outputs for each tree.

Bagging (regression)

The resulting tree is the average of all tree (estimators).



Bagging (classification)



For each bootstrap, we build a decision tree. The results is a combination (majority) of the predictions from all trees.

Bagging

Question: Do you see any problems?

- If trees are too shallow it can still underfit.
- Still some overfitting if the trees are too large.
- Interpretability:

  The **major drawback** of bagging (and other ***ensemble methods*** that we will study) is that the averaged model is no longer easily interpretable - i.e. one can no longer trace the 'logic' of an output through a series of decisions based on predictor values!

Bagging

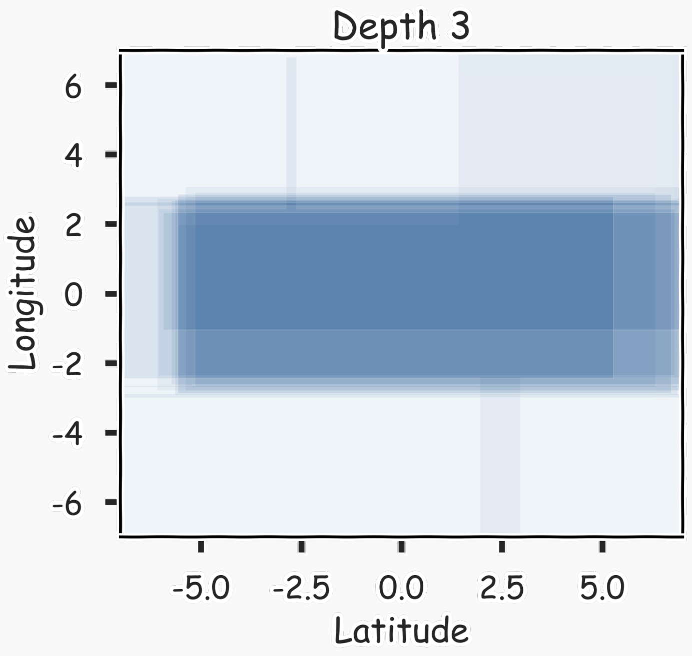
Question: Do you see any problems?

- If trees are too shallow it can still underfit.
- Still some overfitting if the trees are too large.
- Interpretability:

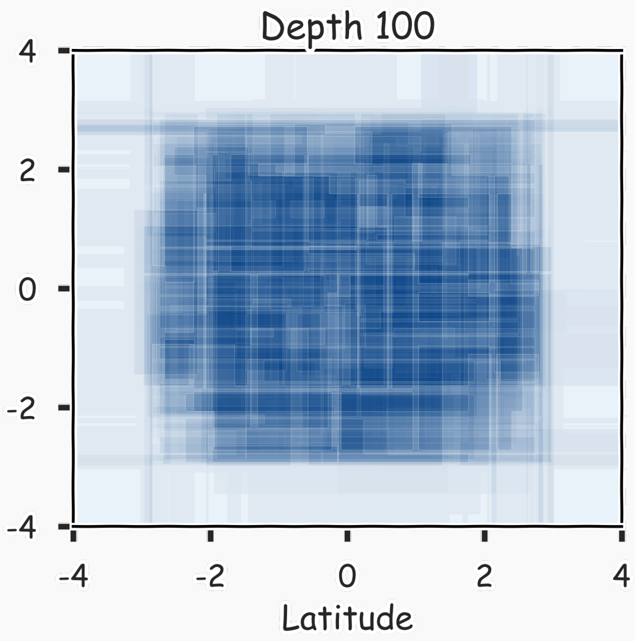
  The **major drawback** of bagging (and other ***ensemble methods*** that we will study) is that the averaged model is no longer easily interpretable - i.e. one can no longer trace the 'logic' of an output through a series of decisions based on predictor values!

Case of underfitting/overfitting

Underfitting



Overfitting



We decide on the complexity of the model using Cross Validation

Bagging

Question: Do you see any problems?

- If trees are too shallow it can still underfit.
- Still some overfitting if the trees are too large.
- Interpretability:

  The **major drawback** of bagging (and other ***ensemble methods*** that we will study) is that the averaged model is no longer easily interpretable - i.e., one can no longer trace the 'logic' of an output through a series of decisions based on predictor values!

Outline

- Review of Decision Trees
- Bagging
- **Out of Bag Error (OOB)**
- Variable Importance

Out of Bag Error (OOB)

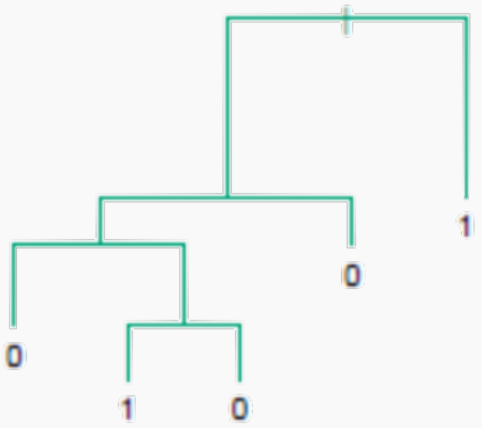
Original Data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Bootstrap Sample 1

X	Y
X_4	y_4
X_{14}	y_{14}
X_{11}	y_{11}
X_2	y_2
X_{35}	y_{35}
⋮	⋮
X_k	y_k

Decision Tree 1



Used and unused data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Out of Bag Error (OOB)

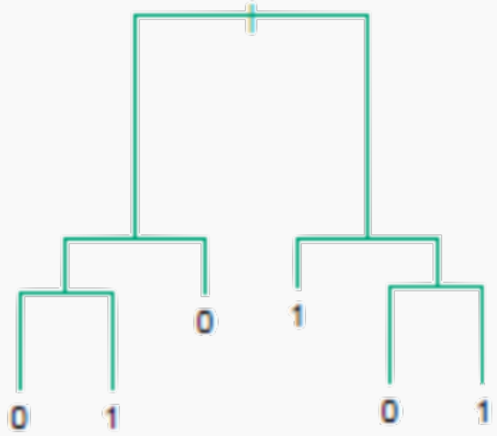
Original Data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Bootstrap Sample 2

X	Y
X_5	y_5
X_3	y_3
X_{12}	y_{12}
X_{43}	y_{43}
X_1	y_1
⋮	⋮
X_k	y_k

Decision Tree 2



Used and unused data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Out of Bag Error (OOB)

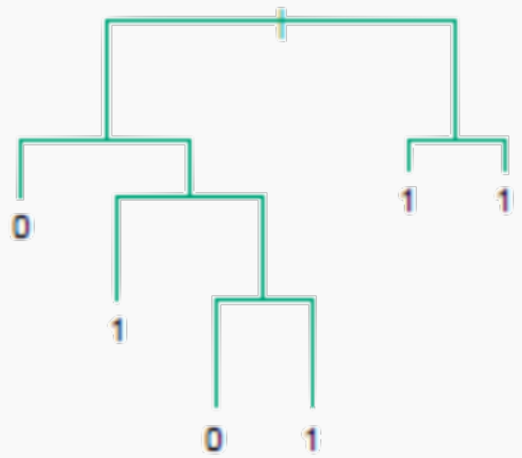
Original Data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Bootstrap Sample 3

X	Y
X_9	y_9
X_4	y_4
X_1	y_1
X_1	y_1
X_{65}	y_{65}
⋮	⋮
X_k	y_k

Decision Tree 3



Used and unused data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

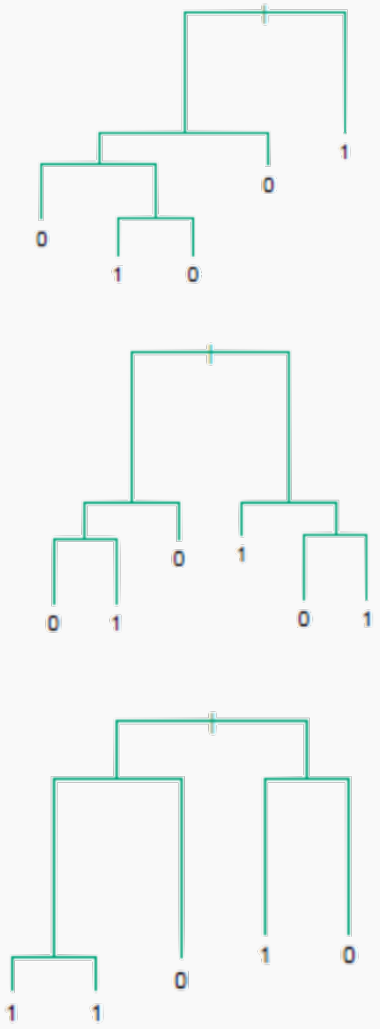
Point-wise out-of-bag error

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
$\vdots$	$\vdots$
X_i	y_i
$\vdots$	$\vdots$
X_n	y_n

Point-wise out-of-bag error

B Trees that did not see $\{X_i, y_i\}$

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
$\vdots$	$\vdots$
X_i	y_i
$\vdots$	$\vdots$
X_n	y_n



Classification

$$\hat{y}_{i,pw} = \text{majority}(\hat{y}_i) \qquad e_i = \mathbb{I}(\hat{y}_{i,pw} \neq y_i)$$

Regression

$$\hat{y}_{i,pw} = \sum_{j \in B} \hat{y}_{i,j} \qquad e_i = \left(y_i - \hat{y}_{i,pw}\right)^2$$

OOB Error

We average the point-wise out-of-bag error over the full training set.

Classification

$$Error_{OOB} = \frac{1}{B}\sum_{i}^{B} e_i = \frac{1}{B}\sum_{i}^{B} \mathbb{I}(\hat{y}_{i,pw} \neq \mathrm{y}_{\mathrm{i}})$$

Regression

$$Error_{OOB} = \frac{1}{B}\sum_{i}^{B} e_i = \frac{1}{B}\sum_{i}^{B} \left(y_i - \hat{y}_{i,pw}\right)^2$$

Out-of-Bag Error

Bagging is an example of an ***ensemble method***, a method of building a single model by training and aggregating multiple models.

With ensemble methods, we get a new metric for assessing the predictive performance of the model, the ***out-of-bag error***.

Given a training set and an ensemble of models, each trained on a bootstrap sample, we compute the ***out-of-bag error*** of the averaged model by

1. For each point in the training set, we average the predicted output for this point over the models whose bootstrap training set excludes this point. We compute the error or squared error of this averaged prediction. Call this the point-wise out-of-bag error.

2. We average the point-wise out-of-bag error over the full training set.

Bagging

Question: Do you see any problems?

- If trees are too shallow it can still underfit.
- Still some overfitting if the trees are too large.
- Interpretability:

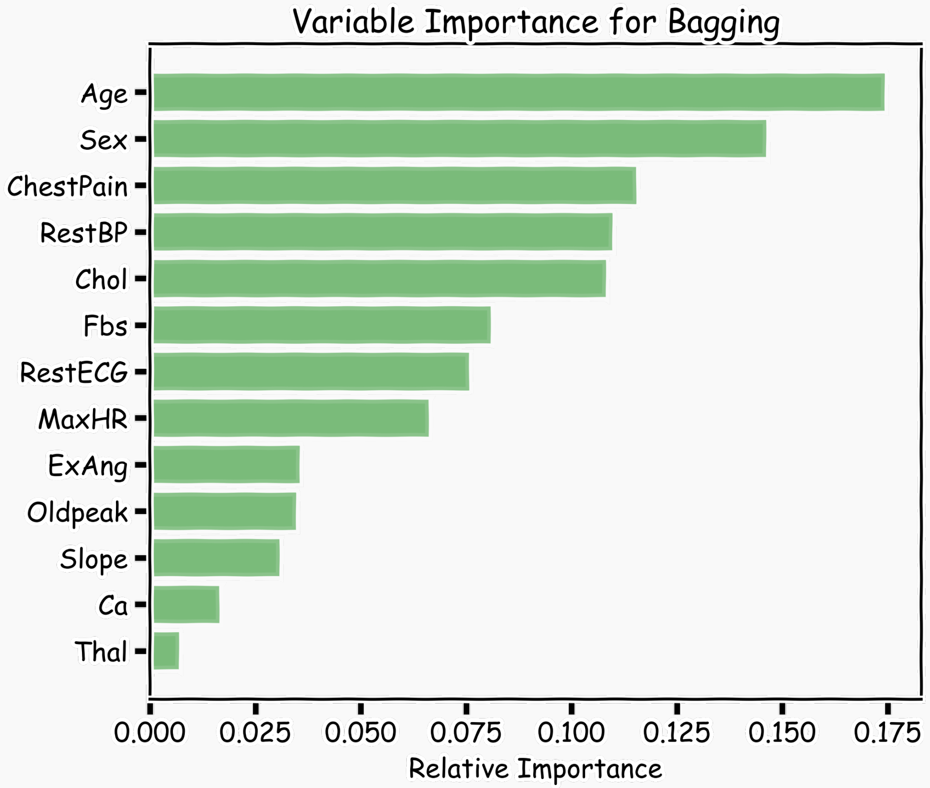
  The **major drawback** of bagging (and other ***ensemble methods*** that we will study) is that the averaged model is no longer easily interpretable - i.e. one can no longer trace the ‘logic’ of an output through a series of decisions based on predictor values!

Outline

- Review of Decision Trees
- Bagging
- Out of Bag Error (OOB)
- **Variable Importance**

Variable Importance for Bagging

Calculate the total amount that the MSE (for regression) or Gini index (for classification) is decreased due to splits over a given predictor, averaged over all B trees.



100 trees, max_depth=10

Improving on Bagging

In practice, the ensembles of trees in Bagging tend to be highly correlated.

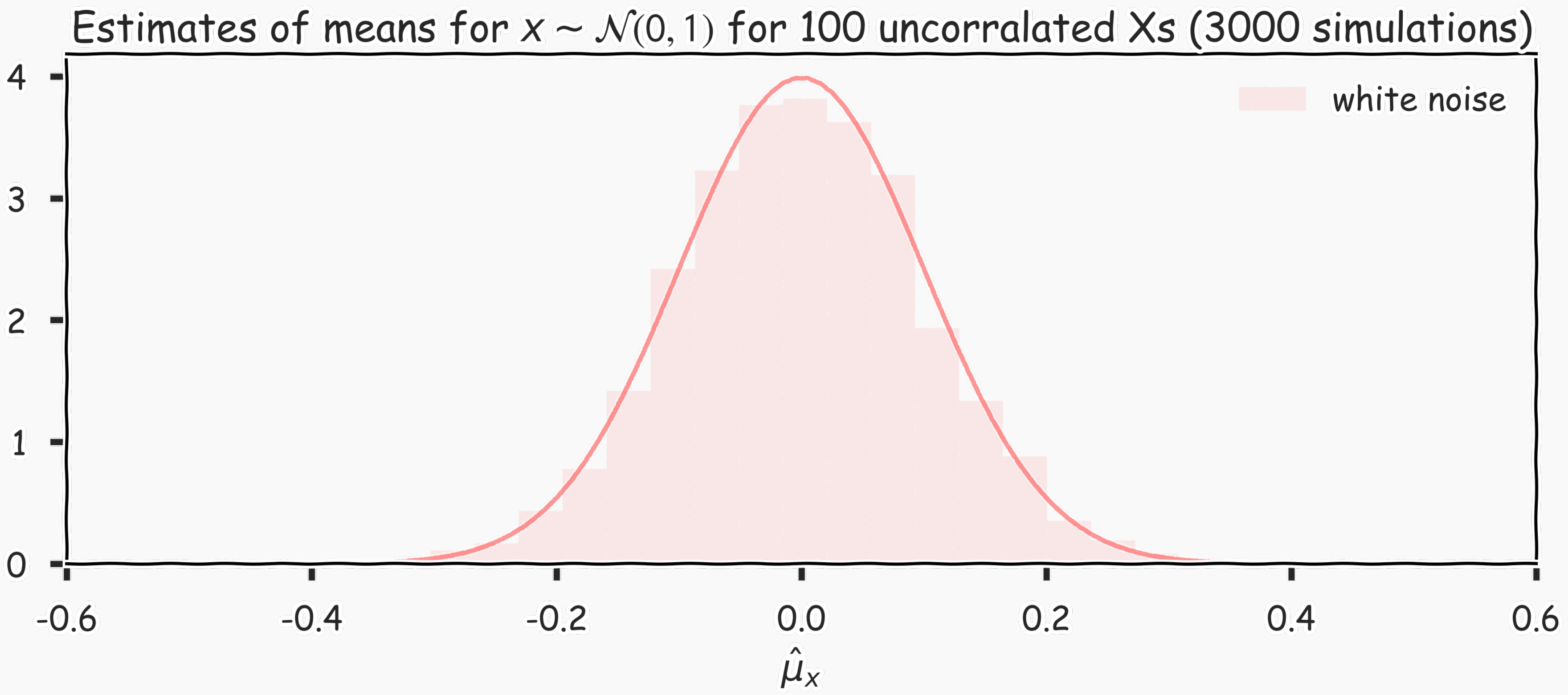
Suppose we have an extremely strong predictor, x_j , in the training set amongst moderate predictors. Then the greedy learning algorithm ensures that most of the models in the ensemble will choose to split on x_j in early iterations.

However, we assumed that each tree in the ensemble is independently and identically distributed, with the expected output of the averaged model the same as the expected output of any one of the trees.

Improving on Bagging

Recall, for B number of identically and independently distributed variable, X, with variance σ^2, the variance of the estimate of the mean is :

$$\text{var}(\hat{\mu}_x) = \frac{\sigma^2}{B}$$



Improving on Bagging

For B number of identically but not independently distributed variables with pairwise correlation ρ and variance σ^2, the variance of their mean is

$$\mathrm{var}(\hat{\mu}_x) \propto \sigma^2(1+\rho^2)/B$$