


9.4 Caveats

If a given predictor is categorical with q classes, there are $2^{q-1} - 1$ possible splits into two groups which becomes computationally prohibitive for large q . Moreover, trees tend to favor these splits due to their high dimensionality, so should be avoided. Finally, importance measures such as node purity are biased to favor these inputs. But permutation-based measures are not.

Note trees are discontinuous predictors 

9.5 Random Forests

single tree



(x_i, y_i)

$x_i = (x_{i1}, \dots, x_{ip})^T$

p features [think large]



Find best $\{x_1, x_2, \dots, x_p\}$
to split over + location of split

[suppose variable j]

$x_j < s$

Find best $\{x_1, x_2, \dots, x_p\}$

to split over

/ 1

etc

$x_j \geq s$

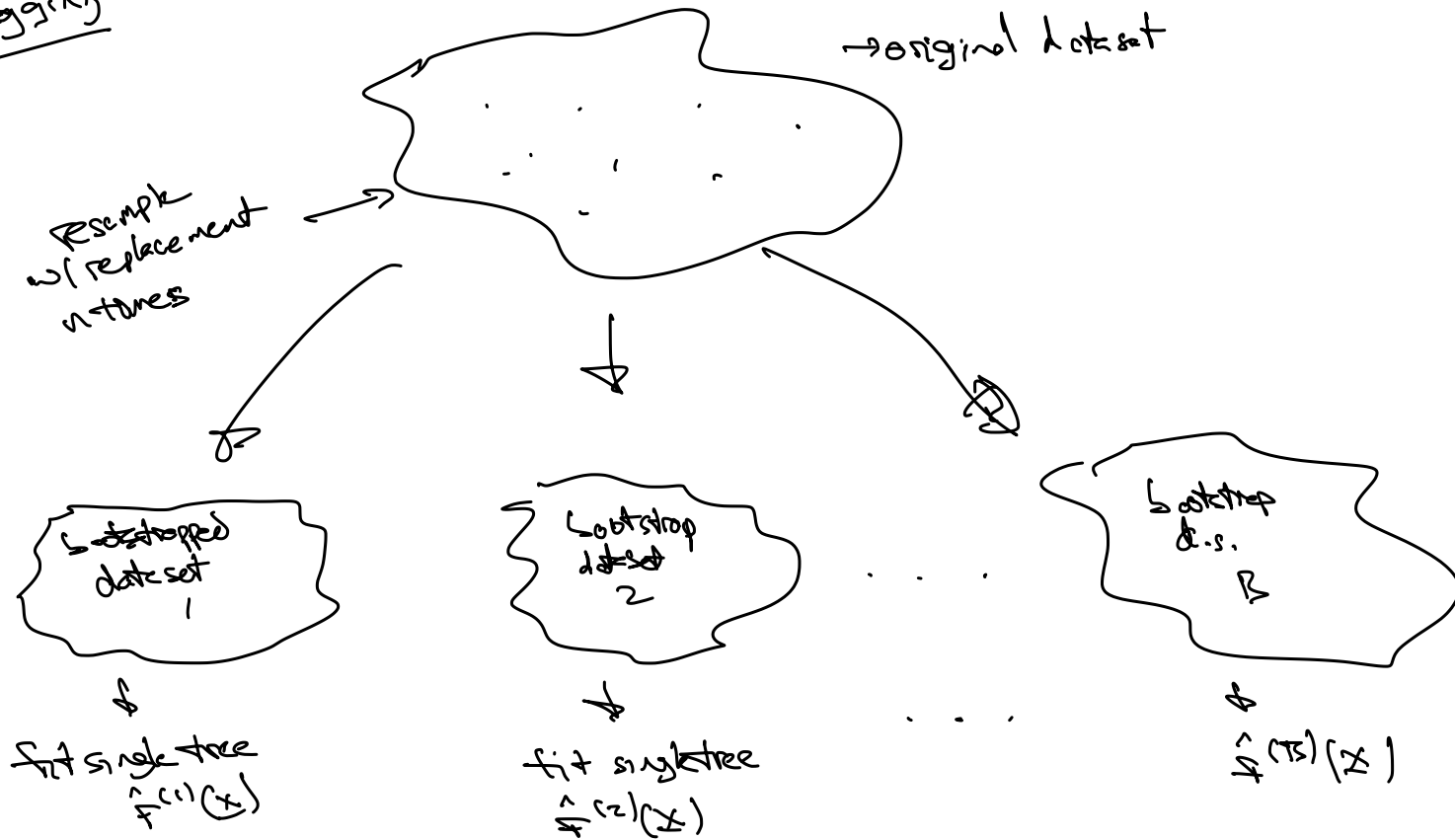
Find best $\{x_1, x_2, \dots, x_p\}$

to split over

/ 1

etc

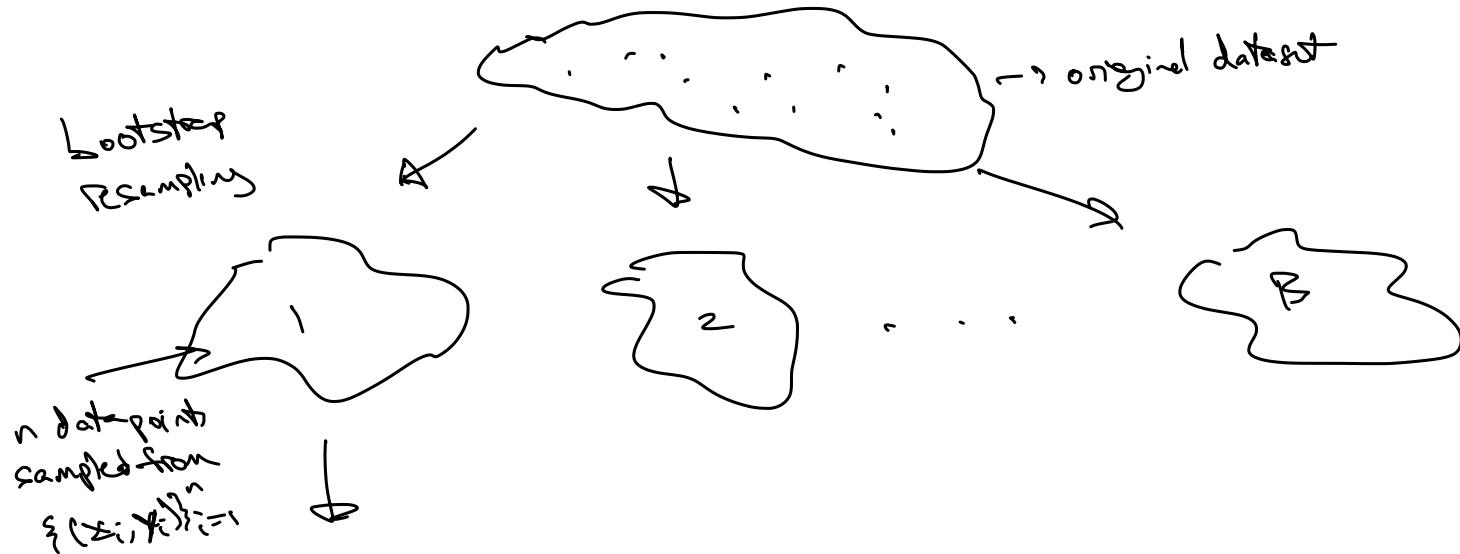
bagging



$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{(b)}(x)$$

random forest

A r.f. follows same approach as bagging,
but at every possible feature split for every
tree, rather than testing all features, we
randomly select a subset of $m < p$ to consider.
This has the effect of decorrelating trees.



Fit tree as follows:

Step 1 Randomly sample $m < p$ features

Ex x_1, \dots, x_{100} $m=4$

check all $x_j \in \{x_1, x_{12}, x_{18}, x_{72}\}$

for best split + location of split

$$x_{12} \leq s$$

resample 4 from
 $\{x_1, \dots, x_{100}\}$

Find $x_j \in \{x_6, x_8, x_{70}, x_{92}\}$

for best split

/ \

resample 4 from p features

$x_j \in \{x_{10}, x_{27}, x_{49}, x_{50}\}$

to split

/ \

results in

$\hat{f}^{(1)}(x), \dots, \hat{f}^{(B)}(x)$ for each B bootstrapped dataset,
the final model is:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{(b)}(x)$$

~~m~~ # of splits

regression $m \approx p/3$

classification $m \approx \sqrt{p}$