

Small data technique

Lecture 14

Changho Suh

January 26, 2024

Challenge of DTs & ensemble learning

Outline

1. Investigate a challenge that arises in DTs.

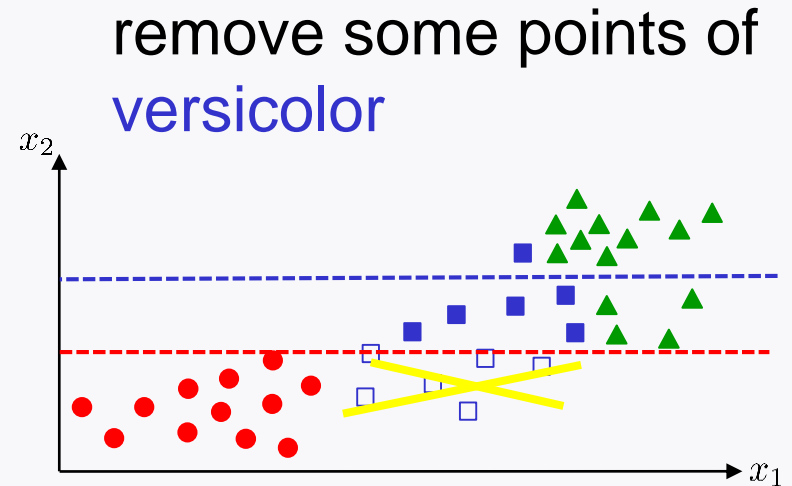
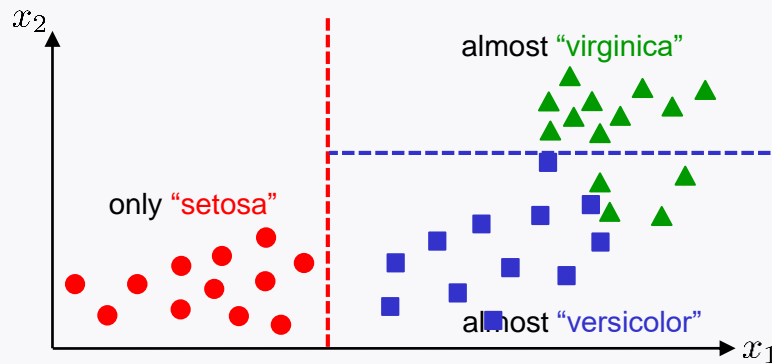
2. Explore a way to address the challenge:

Ensemble learning

Challenge

Sensitive to **small variations** of training data.

Example:



A solution to address variation sensitivity

Turns out:

Ensemble learning is a solution.

For the rest:

1. Study what **ensemble learning** is.
2. Study on powerful ensemble method:

Random forests (RFs)

Ensemble learning

Debate on a decision

How to decide when we have *diverse* opinions?

Often rely on **majority voting**.

Wisdom of the crowd: An aggregated decision is often better than even an expert's answer.

Can expect in the predictor context:

An aggregating prediction based on many predictors

→ A better prediction relative to the best predictor.

Ensemble learning

Ensemble: *A group* of predictors

Ensemble learning:

A technique that aggregates predictions of the ensemble.



Hard voting: Declare the one that gets **most votes**.

Soft voting: Declare the one with **highest probability** averaged over predictors

A way to obtain ensemble

Train each predictor on a **different subset** of the training set.

How to construct different subsets?

1. A way to choose *partial examples*:

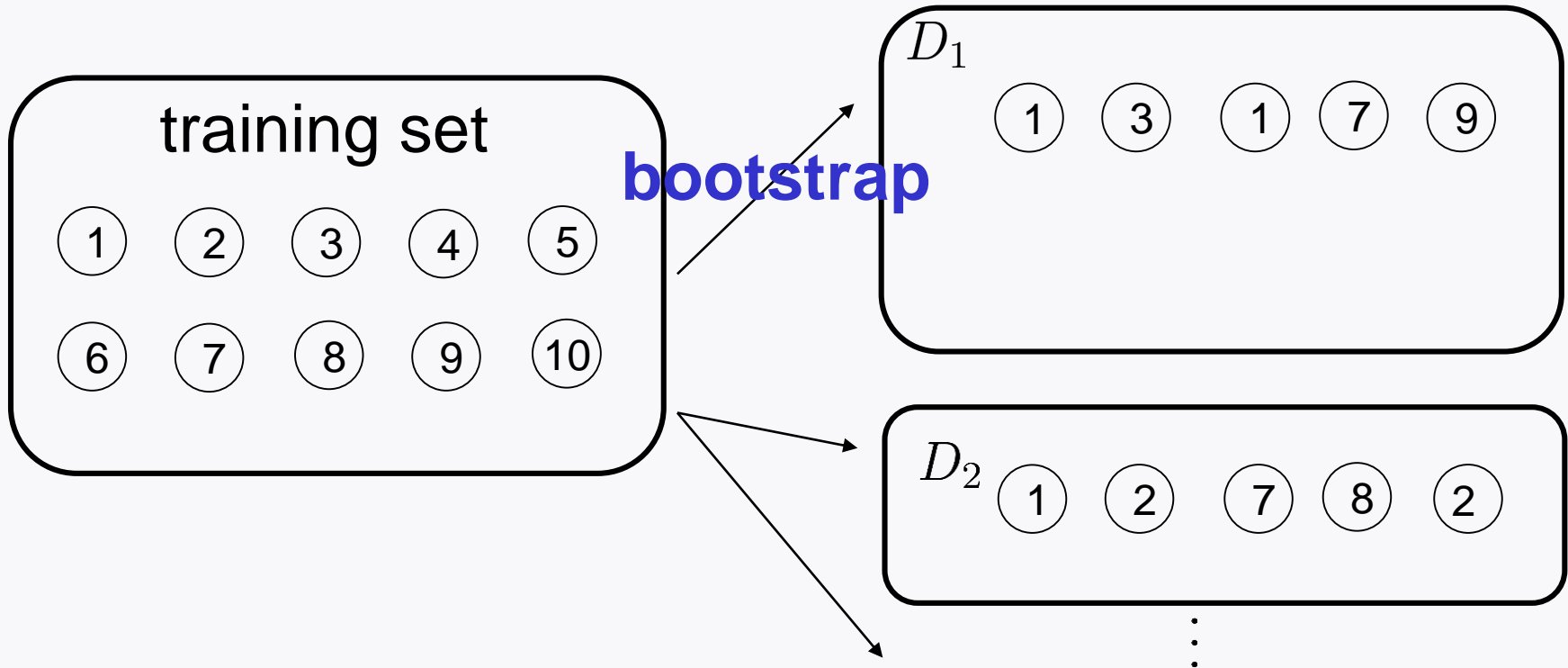
Bootstrap

2. A way to choose *partial features*:

Random subspace method

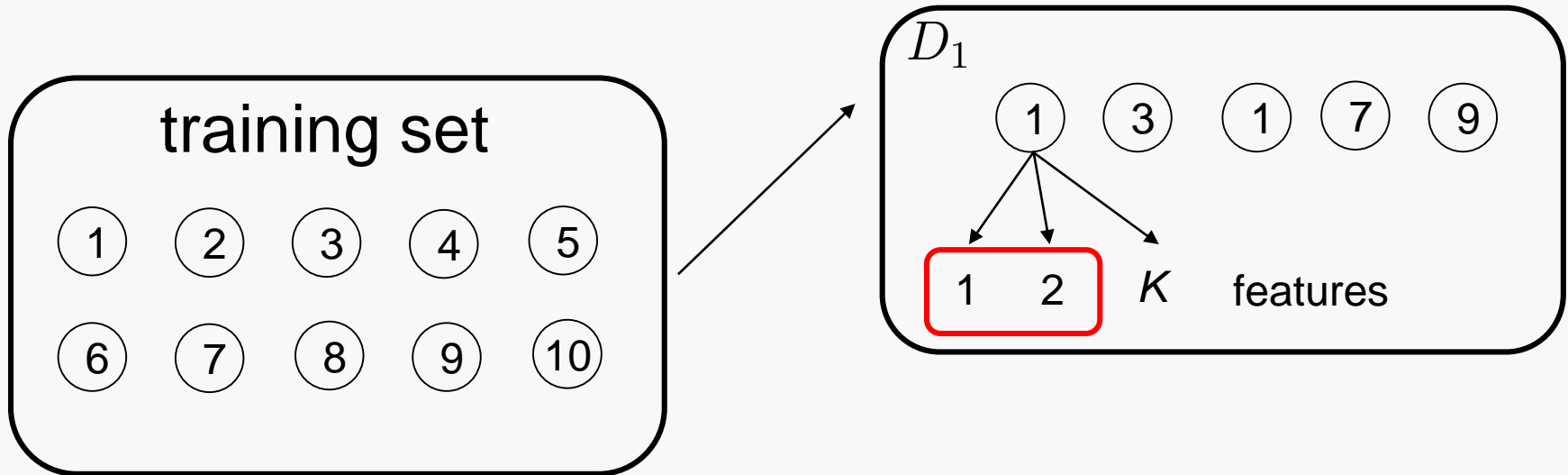
RF = Bootstrap + random subspace

Sampled uniformly at random *w/ replacement*

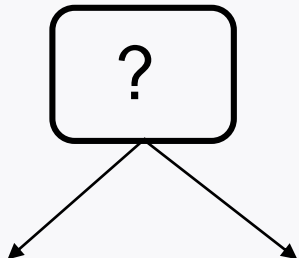


RF = Bootstrap + **random subspace**

Sampled uniformly at random *w/ replacement*

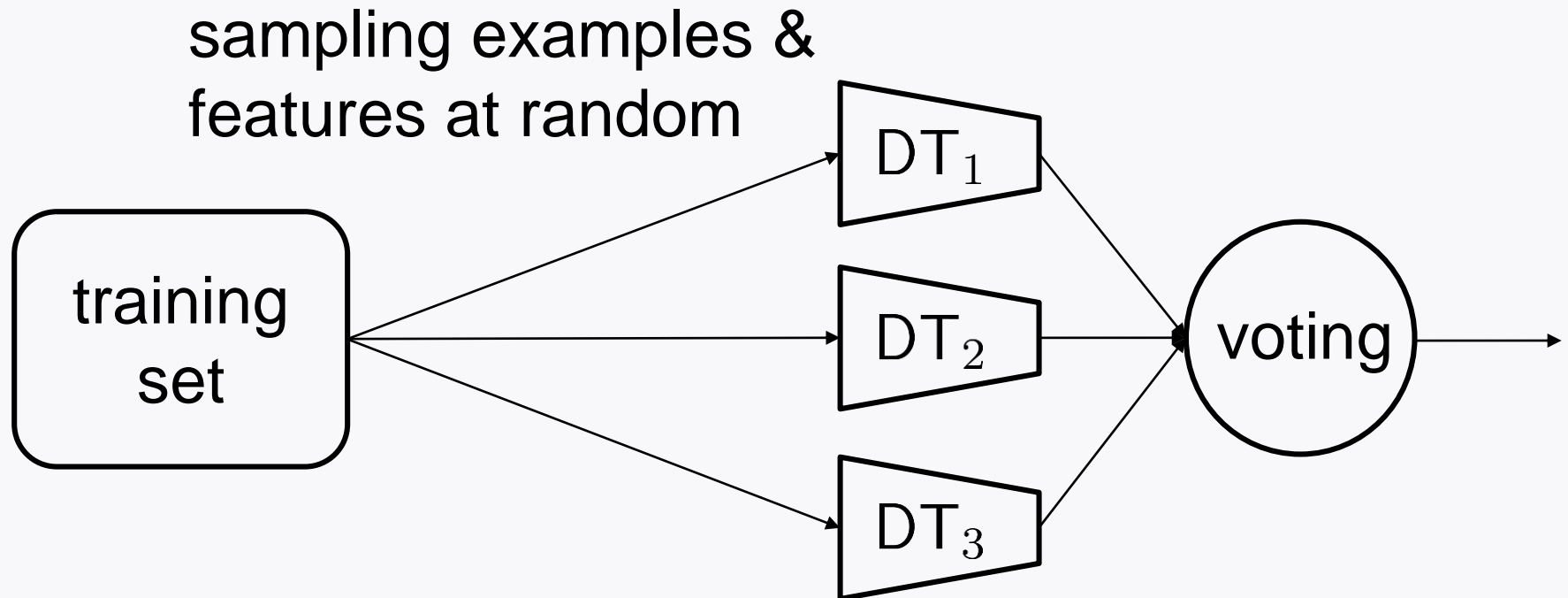


Decision Tree (DT) w/ D_1



Split a node considering a **random subset of features.**

RF in picture



Look ahead

Study **RF** in depth:

1. Investigate **hyperparameters**;
2. Study a measure for model *interpretation*:
Feature Importance