Decision Tree → Random Forest

- Combining weak to strong learners.
  via random forest
- gained huge popularity
- scalibility
  - considered as an ensemble of decision Tree
  - Combines weak learners to a robust model

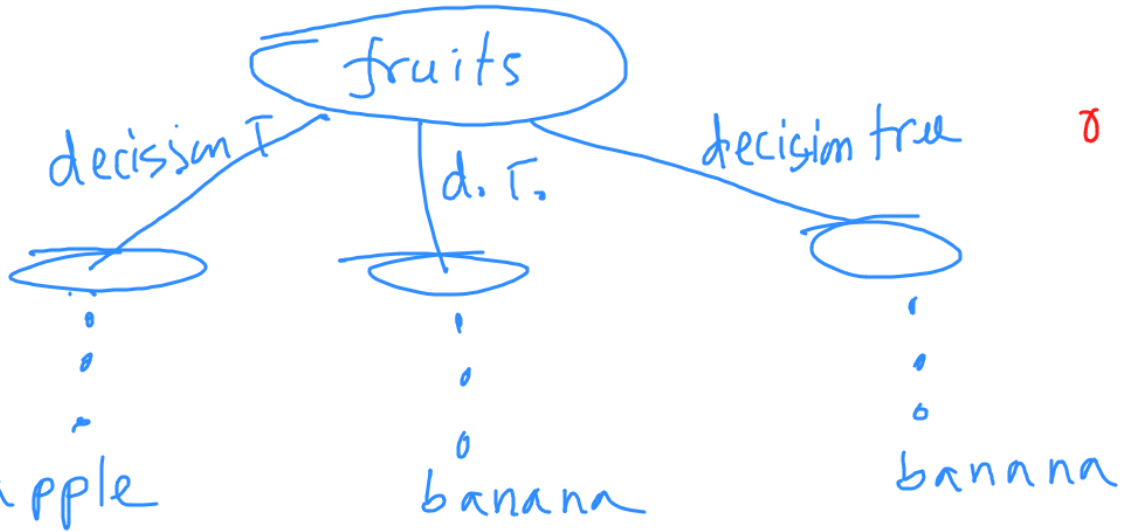
Alg.

1. Draw a random bootstrap sample of size n (randomly choose n samples from the training set with replacement.)
   → same samples can be used more than once
2. Grow a decision tree. At each node
   i) Randomly select d features without replacement.
   ii) Split the node using the features that provids the best spilt.

3. Repeat the steps 1 to 2 K times

4. Aggregate the prediction by each tree based on majority voting.

Ex.

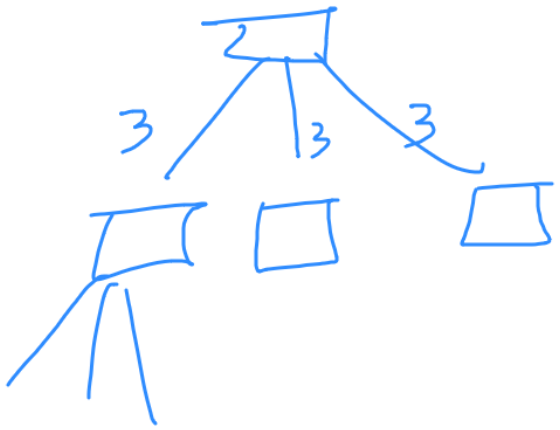


- fruit classification
- Majority voting
  - banana

- We don't have to worry so much about choosing good hyperparameter values.
  - need to fix only $\underline{k}$.
  - we can control the bias-variance trade-off via the sample size $\underline{n}$.
    - standard value of $d = \sqrt{m}$, m is the no. of features in the training set.
    - Larger the no. of trees, the better the performance at increased computation cost.

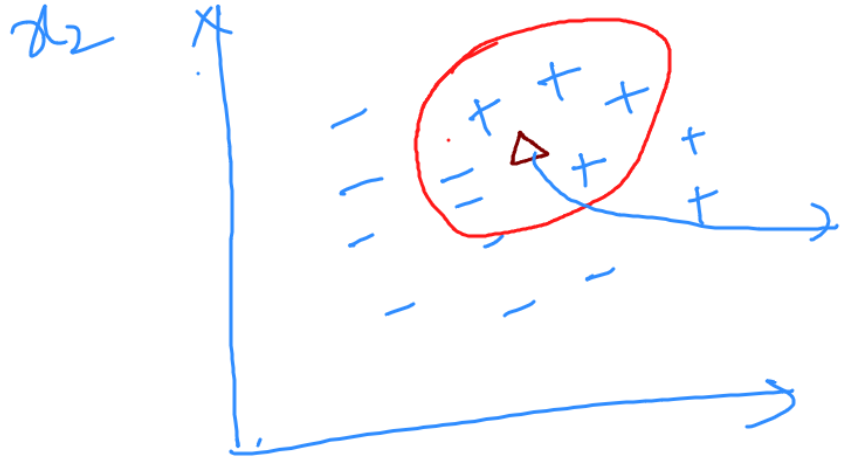
$m = 14$,

$d = \sqrt{m} = 3$



K-nearest neighbours

- fundamentally different from the algs we discussed so far.
- it doesn't learn a disriminative function
- memorizes the training data.



$K \Rightarrow 5$

new data → output class '+'

Algo:

(I) Choose the number of K and a distance metric. → 5

(II) Find the K nearest neighbors of the sample that we want to classify.

(III) Assign the class label by majority voting.

Comments

- The right choice of K is crucial to find a good balance b/t^{n} over- and underfitting.
- ~~Affected~~ by outliers.
- needs to apply scaling.

Distance Metric

$$d(x^{(i)}, x^{(j)}) = \sqrt{\sum_k \left(x_k^{(i)} - x_k^{(j)}\right)^2}$$

↳ Eucidian

Minkowski

$$d(x^{(i)}, x^{(j)}) = \sqrt[p]{\sum \left| x_k^{(i)} - x_k^{(j)} \right|^p}$$

if $p = 2$, Euclidian
$p = 1$, Manhattan

Parametric vs non-parametric models

estimate para ($\underline{w}$)

Ex. Perceptron, SVM, Logistic Reg

decision tree, Random forest

KNN