EE660

Discussion 11

Today:

1. TF-IDF comments from discussion 10
2. CART
3. Random Forest

1.) TF-IDF

- Bag-of-words: x_i is a vector whose entry j is the number of times word j appears in doc. i.
- TF-IDF

$$tf(x_{ij}) = \log(1 + x_{ij})$$

$$idf(j) = \log \frac{N}{1 + \sum_{i=1}^{N} \mathbb{I}(x_{ij} > 0)}$$

$$tf\text{-}idf(\vec{i}) = \left[tf(x_{ij}) \cdot idf(j) \right]_{j=1}^{M}, \quad M \text{ total \# of words}$$

Kernel: cosine similarity

$$K(x_1, x_2) = \frac{x_1^T x_2}{\|x_1\|_2 \|x_2\|_2} \rightsquigarrow \text{bag-of-words}$$

$$K(x_1, x_2) = \frac{\phi(x_1) \cdot \phi(x_2)}{\|\phi(x_1)\|_2 \, \|\phi(x_2)\|_2}$$

$\phi(x_i) = tf\text{-}idf(i)$

~ // ~

CART

. Growing the tree

$$(j^*, t^*) = \underset{j \in \{1, \ldots D\}}{\arg\min} \; \min_{t \in T_j} \left[\begin{array}{l} cost(\text{left node}) + \\ cost(\text{right node}) \end{array} \right]$$

Best split on best feature.

. Since after the split region are indep. of each other (greedy algo.), the order in which regions are analyzed is (usually) not important

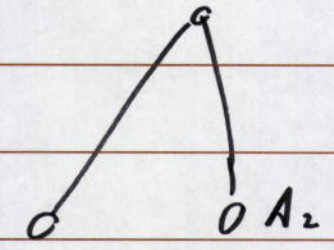
. Some stopping criteria from sklearn (attention to default values)

- max_depth, default = none
- min_sample_split, default = 2
- min_samples_leaf, default = 1
- max_leaf_nodes, default = none
- min_impurity_decrease, default = 0

. Cost choices: accuracy, entropy, gini index

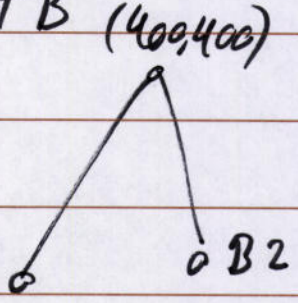
Example: class 0, class 1

Split A (400, 400)



(300, 100) (100, 300)

Split B (400, 400)



(200, 400) (200, 0)

	A1	A2	B1	B2
Π_0	3/4	1/4	1/3	1
Π_1	1/4	3/4	2/3	0
$\hat{y}$	0	1	1	0

Misclassification rate: $e_m = 1 - \bar{\Pi}_{\hat{y}}$

Gini index: $e_g = 2\bar{\Pi}_{\hat{y}}(1 - \bar{\Pi}_{\hat{y}})$

	A1	A2	B1	B2
e_m	1/4	1/4	1/3	0
e_g	3/8	3/8	4/9	0

$$cost(split) = \frac{e_1 \cdot N_1 + e_2 N_2}{N}$$

~~(cost (split A)~~

Misclassification:

$$cost(split\ A) = \frac{\frac{1}{4} \cdot 400 + \frac{1}{4} 400}{800} = \frac{1}{4}$$

$$cost(split\ B) = \frac{\frac{1}{3} \cdot 600 + \phi}{800} = \frac{1}{4}$$

$=$

Gini index:

$$cost(split\ A) = \frac{\frac{3}{8} \cdot 400 + \frac{3}{8} \cdot 400}{800} = \frac{3}{8}$$

$$cost(split\ B) = \frac{\frac{4}{9} \cdot 600 + \phi}{800} = \frac{1}{3}$$

$cost(B) < cost(A)$

Random Forest

1. Bagging: grow different trees based on subsamples of your dataset with replacement

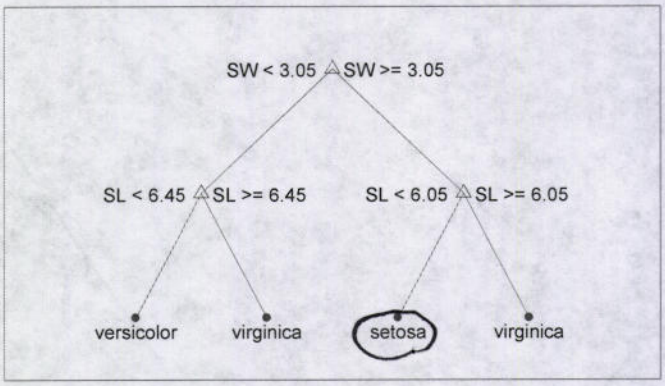
2. Random Forest: not only random subsets of data but also random subset of features at each node.

Some parameters:

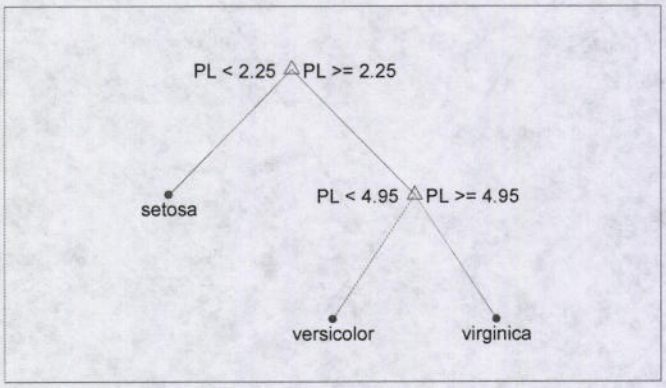
- all from CART

- n_estimators, default=10 : # of trees
- max_features, default= auto : # of considered features at each split

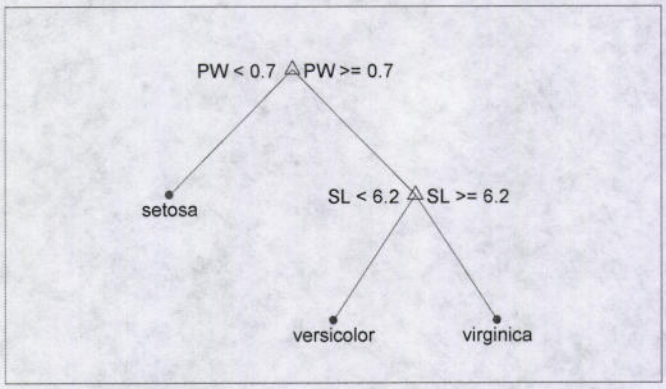
$$X = \begin{bmatrix} 5 & SW \\ 3 & SL \\ 2 & PW \\ 1 & PL \end{bmatrix}$$



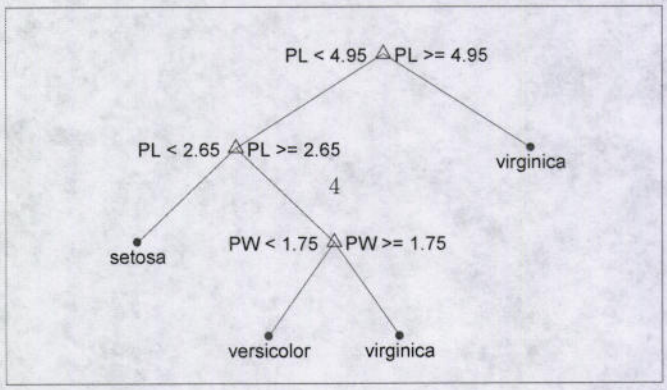
(a) Tree 1



(b) Tree 2



(c) Tree 3



(d) Tree 4

Figure 1: Random Forest simple example