Decision trees tend to overfit

i.e., New training data usually lead to very different trees.

Solution:

Introduce randomised algorithms.

1. Bagging ≡ Bootstrap aggregation

$S = \{ \vec{x}^{(i)}, y^{(i)} \} \quad i = 1, \ldots, N$ Our samples

(1) Bootstrapping: Produce B samples from S all of size N by random sampling with replacement: $S_b \; b = 1, \ldots, B$

(2) Models: From each of the B samples obtain a DT:

$$\{ \hat{f}^{DT}_b \}_{b=1}^{B}$$

In each S_b there will be repeated samples and absent samples from the original S

(3) Aggregate the model:

(i) $$\hat{f}(\vec{x}^{in}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b^{DT}(\vec{x}^{in})$$

Regression.

(ii) Classification:

$$\vec{x}^{in} \longmapsto \left[\vec{\pi}_1^{DT}, \dots, \vec{\pi}_B^{DT} \right] \qquad \vec{\pi}_b = \begin{pmatrix} \pi_{b,1} \\ \vdots \\ \pi_{b,Q} \end{pmatrix}$$

Q classes

$$\vec{\pi} = \frac{1}{B} \sum_{b=1}^{B} \vec{\pi}_b^{DT}$$ (aggregated probability

$$\hat{y} = \arg\max_q \vec{\pi}$$

These outcomes ~~[illegible]~~ are still correlated usually because of dominance of some descriptors.

If the $\{ \hat{f}_b^{DT}(\vec{x}^{in}) \}$ for a given $\vec{x}^{in}$ are all correlated, then averaging them does not reduce the variance beyond a limit.

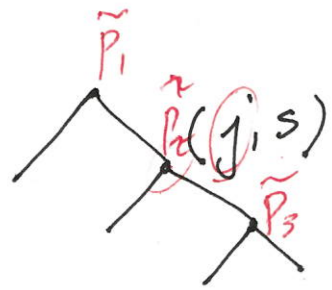
Hence we need to uncorrelate them more radically.

Random forests add an additional level of randomisation.

Steps:

(1) Generate a bootstrap sample S_b

(2)



In DT

$j \in \underbrace{[1, \dots, p]}_{\text{maximize over all predictors}}$

In RF:

$\underbrace{x_j \in \{x_i\}_{i \in RF}}$ ← random subset of descriptors

$\underbrace{|RF| = \tilde{p} < p}_{\text{Chosen at random among the } [1, \dots, p]}$

At each split choose x_j from a random subset of descriptors: with $\tilde{p}$ descriptors.

$\hat{f}^{RF}$

$[\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \dots]_s$

(3) Repeat S times

$s = 1, \dots, S$

The result is an ensemble of trees:

$$\left\{ \hat{f}^{RF}_{[\tilde{p}_k]} \right\}_{s=1}^{S'} = \underline{\text{random forest}}$$

k is defined on each of the splits.

The predictor is the aggregation of

$$\left\{ \hat{f}^{RF}_{[\tilde{p}_k]} \right\}_{s=1}^{S'}$$

- Regression: average of $\left\{ \hat{f}^{RF}_{[\tilde{p}_k]} \right\}_{s=1}^{S'}$
- Classification: $\left\{ \vec{\pi}^{RF}_{[\tilde{p}_k]} \right\}_{s=1}^{S'}$

Some comments:

(1) loss of interpretability:
with aggregation we lose the direct connection with the descriptors.

* Parameters (hyper-parameters):

- S = # of trees. Not very sensitive

- Minimum ~~l~~ leaf size: $\begin{cases} \text{Classification} \mapsto 1 \\ \text{Regression} \mapsto 5 \end{cases}$

- $\tilde{p} < p$ $\begin{cases} \tilde{p} \sim \sqrt{p} & \text{prediction (regression)} \\ \tilde{p} \sim \frac{p}{3} & \text{classification.} \end{cases}$

  # of descriptors allowed at each split

$\left\{ \begin{array}{l} \text{If } \tilde{p} = 1 \text{, trees are very uncorrelated} \\ \text{If } \tilde{p} \simeq p \text{; trees are correlated} \end{array} \right.$

* Bootstrap buys us some extra knowledge:

  For every sample S_b, there are some 'out-of-bag' samples:

out-of-bag error ≡ OOB error ~~can~~ be used for validation

- Example of <u>ensemble methods</u>

  RF is an example of an ensemble method.

Others: $\left\{ \begin{array}{l} \text{(1) mixing different classifiers} \\ \text{(2) Generating ensemble of trees by } \underline{\text{boosting}} \end{array} \right.$

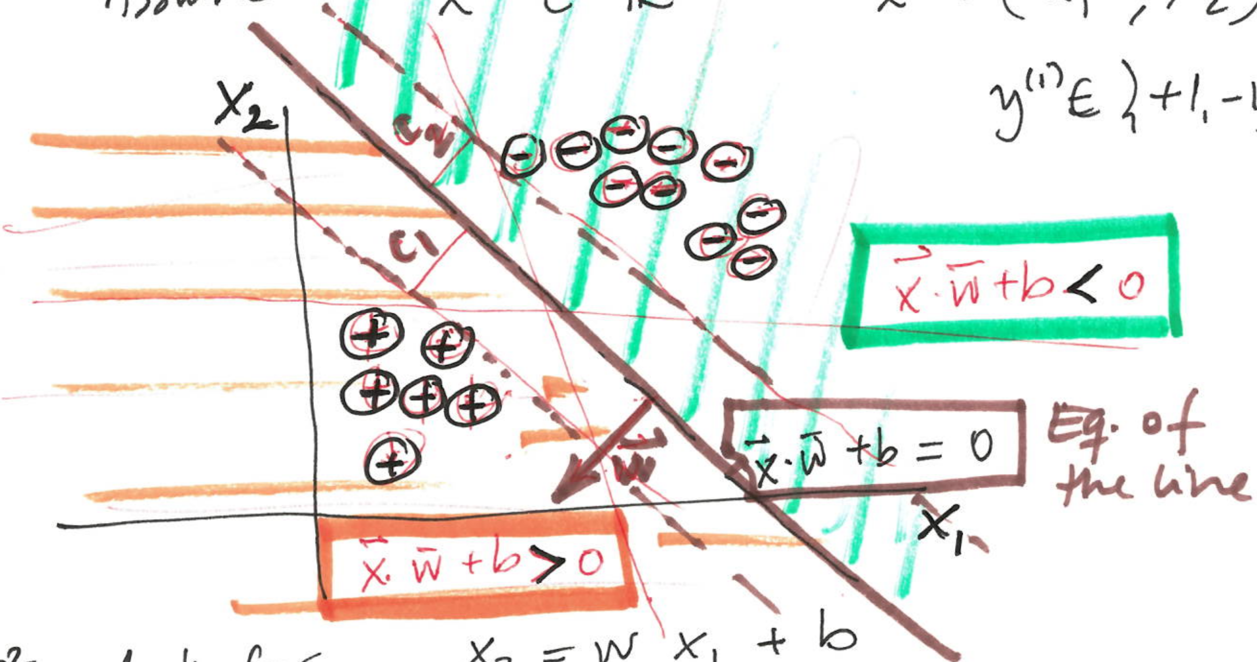
Support vector machines (SVM's) - Vapnik

Classification as geometric separation.

$\vec{x}^{(i)} = (x_1^{(i)}, \dots, x_p^{(i)})$ $\quad y^{(i)} \in \{-1, +1\}$

Assume $\vec{x}^{(i)} \in \mathbb{R}^2$ $\quad \vec{x}^{(i)} = (x_1^{(i)}, x_2^{(i)})$

$y^{(i)} \in \{+1, -1\}$



In $\mathbb{R}^2$: look for a line : $x_2 = w\, x_1 + b$

$$\vec{x} \cdot \vec{w} + b = 0$$

$$\vec{w} = (w, -1)$$

$$(w, -1) \cdot (x_1, x_2) + b = 0 \quad \checkmark$$

For $\vec{x} \in \mathbb{R}^p$

$$\left[\vec{x}^T \cdot \vec{\beta} = \right] \vec{x} \cdot \vec{w} + b = 0$$

Equation of a hyperplane

$$\vec{x} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} \quad \vec{\beta} = \begin{pmatrix} b \\ w_1 \\ \vdots \\ w_p \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \vec{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_p \end{pmatrix}$$

Find the hyperplane

$$\vec{x} \cdot \underline{\underline{\vec{w}}} + \underline{\underline{b}} = 0$$

$(\vec{w}, b)$ to be found.

$$c_1, c_2 \in \mathbb{R}^+ \quad \left\{ \begin{array}{ll} \vec{x}^{(i)} \cdot \vec{w} + b > c_1 & \text{if } y^{(i)} = +1 \\ \vec{x}^{(i)} \cdot \vec{w} + b < -c_2 & \text{if } y^{(i)} = -1 \end{array} \right\}$$