SPORF Ppt Notes

Objective: discover predictable patterns in high dimensional data

Formal definition:

Let: $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$, $y_i \in \mathcal{Y} \subseteq \mathbb{R}^2$

Given: $D_n = (x_i, y_i)$ pairs for $i \in 1, \ldots, n$

↳ assume each pair is sampled independently & identically from some joint distribution, F_{XY}

↳ use D_n to obtain estimate of $F_{Y|X}$

Intuition

↳ Intuitive Desiderata of Supervised Learning Procedures

1. performant under ~~any~~ joint distribution
2. Is interpretable
3. is computationally efficient

↳ Linear 2-way classification in 1 dimension

↳ build a classifier on D_n

↳ try to get all possible splits, compute the "score" for each

↳ split on the best choice (highest score)

↳ predict the class of a new x

↳ $g(x) = 1$ if $x >$ threshold

↳ $g(x) = 0$ if $x <$ threshold

↳ Decision tree in 1 dimension

↳ build a tree on D_n

↳ try all possible splits

↳ compute the "score" for each

↳ split on the best choice (highest score)

↳ create 2 daughter nodes

↳ repeat on daughter nodes

↳ predict the class of a new x

↳ push down the tree

↳ select the plurality class that the node x lands in

↳ Random Forest (RF) in 1 dimension

↳ build forest on D_n

↳ subsample the data to select $m < n$ points
↳ build a tree on each
↳ predict the class of a new x
↳ push down each tree
↳ select the plurality vote of the trees

↳ Random Forest in 1D
↳ what score function should I use?
↳ purity
↳ how deep should each tree be?
↳ as deep as possible
↳ how many trees?
↳ about 1000
↳ how does it scale?
↳ linearly in n, # of trees, dimension of data
↳ not fly any problems in 1D

↳ Linear 2-way classif. in 2D
↳ build classifier on D_n
↳ try all possible angles
↳ compute the "score" for each
↳ split on the best choice (highest score)
↳ predict the class of a new x
↳ $g(x) = 1$ if $x >$ the line
↳ $g(x) = 0$ if $x <$ the line

↳ decision tree in 2D
↳ build a tree on D_n
↳ for each dimension:
↳ try all possible splits
↳ compute the "score" for each
↳ split on the best choice (highest score) to create 2 daughter nodes

↳ repeat on daughter nodes

↳ predict the class of a new x

↳ push down the tree

↳ select the plurality class for the node x lands in

↳ RF in 2D

↳ build a forest on D_n

↳ subsample the data to select $m < n$ points

↳ build a tree on each

↳ predict the class of a new x

↳ push down each tree

↳ select the plurality vote of the trees