ADAPTIVE BASIS FUNCTION MODELS

KERNEL METHODS RELY TOO MUCH ON HAVING A GOOD KERNEL LOL. LET'S LEARN RELEVANT FEATURES DIRECTLY FROM DATA.

$f(x) = w_0 + \sum^{M} w_m \phi_m(x)$, ϕ IS BASIS FUNCTION LEARNED FROM DATA. $\phi_m(x) = \phi(x, v_m)$ ARE PARAMETRIC. MODELS NON LINEAR ANYMORE.

ONLY LOCAL ESTIMATES OF MLE AND MAP FOR θ WHOLE PARAMETERS $\{w, m\}$

CLASSIFICATION AND REGRESSION TREES

RECURSIVELY PARTITION INPUT SPACE, DEFINING LOCAL MODEL IN EACH REGION. EACH LEAF IS A REGION. AXIS PARALLEL SPLITS. PIECEWISE CONSTANT SURFACES.

- **GENERAL ALGORITHM**

OPTIMAL PARTITIONING IS NP-COMPLETE. → GREEDY ALGOS. CART, C4.5, ID3. SPLIT FUNCTION CHOOSES BEST FEATURE AND BEST VALUE FOR FEATURE TO SPLIT ON.

- STOPPING HEURISTICS → IS REDUCTION COST TOO SMALL? > MAX DEPTH? RESPONSE DISTRIBUTIONS SUFFICIENTLY HOMOGENEOUS? N° CASES IN RESPONSES TOO SMALL?
- **REGRESSION COST:** $\text{COST}(D) = \sum (y_i - \bar{y})^2$, $\bar{y} = \frac{1}{|D|} \sum y_i$ MEAN OF RESPONSE VARIABLE. ELSE FIT REGRESSION MODEL AT EACH LEAF WITH INPUTS ROOT → LEAF AND MEASURE RESIDUAL ERROR.
- **CLASSIFICATION COST:** FIT A MULTINOULLI BY ESTIMATING CLASS-CONDITIONAL PROBABILITIES $\hat{\pi}_c = \frac{1}{|D|} \sum \mathbb{I}(y_i = c)$

  ERROR MEASURES:
  - MISCLASSIFICATION RATE $\hat{y}_i = \text{ARGMAX}_c \hat{\pi}_c \rightarrow \frac{1}{|D|} \sum \mathbb{I}(y_i \neq \hat{y}) = 1 - \hat{\pi}_{\hat{y}}$
  - ENTROPY / DEVIANCE: $H(\hat{\pi}) = -\sum^{C} \hat{\pi}_c \log \hat{\pi}_c$ MIN ENTROPY ⟷ MAX INFORMATION GAIN $H(Y) - H(Y|X, \ldots)$

    $\hat{\pi}_c$ IS MLE
  - GINI INDEX: $\sum^{C} \hat{\pi}_c (1 - \hat{\pi}_c) = 1 - \sum \hat{\pi}_c^2$ EXPECTED ERROR RATE. BEHAVES SIMILARLY TO X-ENTROPY, LEAD TO MORE PURE NODES

- **PRUNING** EARLY STOPPING IS MYOPIC. GROW A FULL TREE THEN PRUNE. PRUNE BRANCHES GIVING LEAST INCREASE IN ERROR. PICK SMALLEST TREE WHOSE CV ERROR IS WITHIN 1 STDERR OF MINIMUM.

- **PROS:**
  - EASY TO INTERPRET
  - HANDLE MIXED TYPES OF INPUTS
  - AUTO VAR SELECTION
  - ROBUST TO OUTLIERS
  - SCALE WELL
  - CAN MISSING INPUTS

- **CONS**
  - NOT VERY ACCURATE
  - UNSTABLE BECAUSE BUILT HIERARCHICALLY → ERRORS PROPAGATE TOP TO BOTTOM

RANDOM FORESTS

LET'S REDUCE VARIANCE BY AVERAGING MANY ESTIMATES

BAGGING: TRAIN M TREES ON DIFFERENT DATA SUBSETS RANDOM w/ REPLACEMENT. $f(x) = \sum \frac{1}{M} f_m(x)$. BOOTSTRAP AGGREGATING → HIGH-CORR PREDICTORS

RANDOM FORESTS: TREES ARE TRAINED ON RANDOM SUBSETS OF VARS AND DATA CASES. AWESOME.

- BAYESIANLY WE CAN PERFORM APPROX INFERENCE OVER SPACE OF TREES (STRUCTURE + PARAMS) OR ENSEMBLE OF TREES BART

- **HIERARCHICAL MIXTURE OF EXPERTS**

CAN BE ALTERNATIVE TO TREES. DIFFERENT EXPERT ON EACH PARTITION. ANY NESTED LINEAR DECISION BOUNDARIES. NOT JUST AXIS PARALLEL. GLOBAL PREDICTION IS AVG OF ALL EXPERTS. CAN USE EM BECAUSE IT'S SMOOTH CONTINUOUS PROBLEM.

GENERALIZED ADDITIVE MODELS

$f(x) = \alpha + f_1(x) \dots + f_D(x_D)$. f IS SCATTERPLOT SMOOTHER. MAPPED TO $p(y|x)$ WITH LINK FUNCTIONS AS IN GLM.
BASIS FUNCTIONS ARE SMOOTHING/REGRESSION SPLINES. **BACKFITTING** ALGORITHM. IF X FULL RANK OBJECTIVE IS CONVEX
$O(NDT)$ • **MARS:** ALLOWS FOR INTERACTION EFFECTS, ANOVA MODEL. $f(x) = \beta_0 + \sum^D f_j(x_j) + \sum f_{jk}(x_j, x_k) + \sum f_{jkl}(x_j, x_k, x_l)$
→ **SIMILAR TO CARTS** BUT WITH PIECEWISE LINEAR BASIS FUNCTION VS STEPS FUNCTIONS.

BACKFITTING: ITERATIVELY UPDATE f_j USING RESIDUALS $(-f_j)$ AS TARGET VECTOR
NORMALIZE AT 0 MEAN WITH EACH STEP "WEIGHTED IRLS"

MARS • GAM + INTERACTION + BASIS FCN + CARTS + SPLINES

BOOSTING

GREEDY APPROACH FOR FINDING ADAPTIVE BASIS FUNCTION MODELS. ϕ_m GENERATED BY WEAK/BASE LEARNER. WL IS APPLIED SEQUENTIALLY TO WEIGHTED DATA. EARLIER MISCLASSIFIED DATA → MORE WEIGHT. CAN DO ON ANY REGR/CLASS ALGO BUT USUALLY ON CARTS. IT'S BESTEST. BOOSTING WORKS AS LONG AS WL DOES BETTER THAN CHANCE. VERY RESILIENT TO OVERFITTING. CAN BE SEEN AS GRADIENT DESCENT IN FUNCTION SPACE. CAN BE EXTENDED TO MANY LOSS FCNS.

BOOSTING PROBLEM: $\min_f \sum^N L(y_i, f(x_i))$. L IS LOSS, f IS ABM. TRUE CONDITIONALS CAN'T BE KNOWN. MINIMIZES THEIR EXPECTATION → **POPULATION MINIMIZER**

BOOSTING APPROXES LOGODDS RATIO SEQUENTIALLY

• $f_0(x) = \underset{\gamma}{\text{ARGMIN}} \sum^N L(y, f(x, \gamma))$

$(\beta_m, \gamma_m) = \text{ARGMIN} \sum^N L(y_i, f_{m-1}(x_i) + \beta\phi(x, \gamma))$

$f_m(x) = f_{m-1}(x) + \beta_m \phi(x, \gamma_m)$

• NO GOING BACK TO OPTIMIZE EARLIER PARAMS **FORWARD STAGEWISE ADDITIVE MODELING**
• M ITERATIONS, MONITOR PERF ON ~~SETTING~~ TESTSET AND STOP EARLY, AIC OR BIC

• **IN PRACTICE:** $f_m(x) = f_{m-1}(x) + \nu\beta_m\phi(x, \gamma_m)$, ν STEPSIZE $0 < \nu \leq 1$, **SHRINKAGE**

L2 BOOSTING: SQUARE ERROR LOSS, $L(\dots) = (\underbrace{(y_i - f_{m-1}(x_i))}_{\text{RESIDUAL}} - \phi(x, \gamma))^2$, $\beta = 1$ WLOG. WEAK LEARNER PREDICTS R AND FINDS NEW BASIS FCN

ADABOOST BINARY CLASSIF, EXPONENTIAL LOSS. $L_m(\phi) = \sum^N w_{i,m} \exp(-\beta \tilde{y} \phi(x_i))$, $w_{im} = \exp(-\tilde{y}_i f_{m-1}(x_i))$ WEIGHT 4 DATAPOINT. $\phi_m = \text{ARGMIN}\ w_{i,m}(y \neq \phi($
WEAK LEARNER TO WEIGHTED DATASET. SOLVE FOR $\beta_m = \frac{1}{2} \log \frac{1 - ERR}{ERR}$ → $f_m(x) = f_{m-1}(x) + \beta_m \phi_m(x)$
COMPLEX DECISION BOUNDARIES SUPER RESILIENT TO OVERFITTING.

LOGIT BOOST! EXP LOSS IS TOO HEAVY ON MISCLASSIFICATIONS. SENSITIVE TO OUTLIERS. USE **LOGLOSS** → MISTAKES HAVE LINEAR PENALTY. CAN EXTRACT PROBS.
$L(\phi_m) = \sum^N \log[1 + \exp(-2\tilde{y}_i(f_{m-1}(x) + \phi(x_i)))]$ NEWTON UPDATES

GRADIENT BOOSTING: 'GENERIC' MODEL 4 BOOSTING $\hat{f} = \text{ARGMIN}\ L(f)$. USUAL STAGEWISE. $g_{im} = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]$. $f = f_{m-1} - \rho_m g_m$. ρ_m STEPSIZE SO TO $\rho = \text{ARGMIN}\ L(f)$
→ **FIT A WEAK LEARNER** $\gamma_m = \text{ARGMIN} \sum^N (-g_{i,m} - \phi(x, \gamma))^2$. UPDATE $f_m(x) = f_{m-1}(x) + \nu\phi(x, \gamma_m)$. FLEXIBLE

SPARSE BOOSTING: PICK WL THAT BEST PREDICTS RESIDUAL VECTOR. FORWARD STAGEWISE LINEAR REGRESSION. $\beta_m = \beta_{m-1} + \nu(0, 0, 0 \dots \beta_{k(m)}, 0\, 0\, 0)$
$\nu \to 0$ IS LAR. INCREASING NO STEPS = REDUCING REGULARIZATION λ. + VARIABLE DELETION → LARS

MART: MULTIVARIATE ADAPTIVE REGRESSION TREES. GRADIENT BOOSTING + SHALLOW DECISION TREES/STUMPS. SHALLOW BC VARIANCE IS LOW AND BIAS IS FIXED WITH MOAR ROUNDS

• BOOSTING CAN BE SEEN AS ℓ_1 REGULARIZATION. ELIMINATES 'IRRELEVANT' FEATS. ADABOOST + ℓ_1 = **L1-ADABOOST**: GREEDILY ADDS BEST FEATS, THEN PRUNES WITH ℓ_1
→ ALSO BOOSTING MAXIMIZES MARGINS ON TRAINING DATA

• BAYESIANLY BOOSTING IS AKIN TO MIXTURE OF EXPERTS $p(y|x, \theta) = \sum^M \pi_m p(y|x, \gamma_m)$. EACH EXPERT IS WEAK LEARNER.

FEED FORWARD NEURAL NETWORKS - MLPs

SERIES OF LOGISTIC MODELS STACKED ON TOP OF EACHOTHER. FINAL LAYER IS LOGISTIC (CLASSIF) OR LINEAR (REGR).

- $P(y|x,\theta) = N(y|w^T z(x), \sigma^2)$
- $z(x) = g(Vx) = [g(v_1^T x) \dots g(v_H^T x)]$
- g = NONLINEARITY, ACTIVATION FCN
- $z(x)$ = HIDDEN LAYER
- H = NO. OF HIDDENS
- V = WEIGHT MATRIX INPUT → HIDDEN
- W = WEIGHT VECTOR HIDDEN → OUTPUT
- IF g LINEAR → COLLAPSE TO $y = W^T(Vx)$
- UNIVERSAL APPROXIMATORS: ANY FCN GIVEN ENOUGH H
- NONLINEARITIES = SIGM (OUTPUT, BIN) SOFTMAX (OUTPUT MULTICLASS) TANH (HIDDENS)

- BINARY CLASSIF: $P(y|x,\theta) = Ber(y|sigm(w^T z(x)))$ - LIKE GLM
- MULTI CLASSIF: VIA MUTUAL INHIBITION / SUM TO ONE / SOFTMAX $P(y|x\theta) = Cat(y|S(Wz(x)))$
- REGRESSION: $P(y|x,\theta) = N(y|W\phi(x,V), \sigma^2 I)$

- $|N|$ TRAINING $\approx 10\times |W|$ NO WEIGHTS

CONVNETS

PURPOSE OF HIDDENS → LEARN NONLINEAR COMBINATIONS OF INPUTS. FEATURE EXTRACTION. FEATS USED BY FINAL GLM. GOOD WHERE ORIGINAL INPUTS AREN'T VERY INFORMATIVE. IN CONVNETS HIDDENS HAVE **RECEPTIVE FIELDS** AND PARAMETERS ARE **TIED / SHARED** PER LAYER. EXHIBIT **TRANSLATIONAL INVARIANCE**. CONV LAYERS CREATE **FEATURE MAPS**. INCREASE RESILIENCE BY AUGMENTING DATASET WITH DISTORTED VERSION OF ORIGINAL DATA.
LENET5 INTRODUCES SUBSAMPLING BY POOLING / AVERAGING / DROPOUT BETWEEN CONV. LAYERS → SHIFT INVARIANCE. FINAL LAYER IS RBF NET. → VS. SIGM OR SOFTMAX

OTHERS

- SKIP ARCS → DIRECT ARCS BETWEEN IN AND OUT
- RECURRENT NETS → FEEDBACK CONNECTIONS, NONLINEAR DYNAMICAL SYSTEM
- **HOPFIELD NETWORK / ASSOCIATIVE MEMORY** IF WE ALLOW SYMMETRIC CONNECTIONS BETWEEN HIDDENS $\approx$ BOLTZMANN MACHINE (PROBABILISTIC)

BACKPROPAGATION

NLL OF MLP IS NONCONVEX. WE HAVE TO USE GRADIENT-BASED OPTIMIZERS LIKE SGD BECAUSE ONLINE.

MLP FORWARD MODEL: $\underbrace{x_n \xrightarrow{V} a_n \xrightarrow{g} z_n}_{\text{HIDDEN LAYER}} \underbrace{\xrightarrow{w} b_n \xrightarrow{h} \hat{y}_n}_{\text{OUTPUT LAYER}}$; $\theta = (V, W)$

ALGO: $\Delta O \to \Delta H \to W_{O \to H}, W_{H \to I}$

- **COMPUTE LOG LIKELIHOODS** REGRESSION = SQ. ERROR, CLASSIF = X-ENTROPY | $J(\theta) = -\sum_n \sum_k (\hat{y}_{nk}(\theta) - y_{nk})^2$ || $J(\theta) = -\sum_n \sum_k y_{nk} \log \hat{y}_{nk}(\theta)$
- **OUTPUT LAYER GRADIENT:** $\nabla_{w_k} J_n = \frac{\partial J_n}{\partial b_{nk}} z_n = \delta^w_{nk} z_n$, $\delta^w_{nk} = (\hat{y}_{nk} - y_{nk})$
- **INPUT LAYER GRADIENT:** $\nabla_{v_j} J_n = \frac{\partial J_n}{\partial a_{nj}} = \delta^v_{nj} x_n$; $\delta^v_{nj} = \frac{\partial J_n}{\partial a_{nj}} = \sum \frac{\partial J_n}{\partial b_{nk}} \cdot \frac{\partial b_{nk}}{\partial a_{nj}} = \sum \delta^w_{nk} w_{kj} g'(a_{nj})$ **CAN BE DONE VIA LAYER 2 ERRORS BACK TO W** → LOCALLY
- **FINAL:** $\nabla_\theta J(\theta) = \sum_n [\delta^v_n x_n, \delta^w_n z_n]$

IDENTIFIABILITY: COL CAZZO. NONLINEARITIES ARE SYMMETRIC / ODD. 2^H EQUIV SETTINGS FOR SIGN FLIPS + SWITCHING HIDDENS = TOTAL OF $H! 2^H$ EQUIVALENT PERMUTATIONS. PLUS LOCAL MINIMA DUE TO NLL NONCONVEXITY

REGULARIZATION: EARLY STOPPING (STOP WHEN TEST ERROR STARTS TO RISE) • IMPOSE PRIOR ON PARAMS, THEN MAP, LIKE $N(0, \alpha^{-1} I) \to \ell_2$ REG. **WEIGHT DECAY**

→ NLL BECOMES $J(\theta) = -\sum_n^N \log p(y_n | x_n, \theta) + \frac{\alpha}{2}\left[\sum_{ij} V_{ij}^2 + \sum_{jk} w_{jk}^2\right]$, $\nabla_\theta J(\theta) = \left[\sum_n \delta^v_n x_n + \alpha v_i, \sum_n \delta^w_n z_n + \alpha w\right]$

- STRONG REGULARIZATION → WHO CARES IF MANY H.
- SET H LARGE, THEN REGULARIZE
- SET α WITH CV OR EB
- STANDARDIZE INPUT FOR NICE PRIOR BEHAVIOR
- SAME REG. PARAMS FOR 1 AND 2 LAYER → NO INVARIANCE; SET DIFFERENT ONES FOR WEIGHTS AND BIASES PER LAYER

$P(\theta) = N(W|0, \frac{1}{\alpha_w} I) N(V|0, \frac{1}{\alpha_v} I) N(b|0, \frac{1}{\alpha_b} I) N(c|0, \frac{1}{\alpha_c} I)$

- **ENCOURAGE SPARSITY:** WITH ℓ_1 OR ARD OR AD-HOC
- ENCOURAGE PARAMS TO HAVE SIMILAR STATISTICAL PROPERTIES. $P(\theta)$ AS MIXTURE OF DIAGONAL GAUSSIANS → SAME CLUSTER, SAME μ, σ^2 **SOFT WEIGHT SHARING**

SEMI-SUPERVISED EMBEDDING

ENCOURAGE HIDDENS TO ASSIGN SIMILAR OBJECTS TO SIMILAR REPRESENTATIONS. $L(f_i, f_j, S) = \begin{cases} \|f_i - f_j\|^2 & S=1 \\ \max(0, m - \|f_i - f_j\|^2) & S=0 \end{cases}$

- $\sum \mathrm{NLL}(f(x_i) y_i) + \lambda \sum L(f(x_i), f(x_j), S_{ij})$
- PICK RANDOM LABELED EXAMPLE, AND GRADIENT NLL; PICK RANDOM SIMILAR UNLABELED EXAMPLES AND GRADIENT L_1; PICK RANDOM UNLABELED WITH HIGH PROB DISSIMILAR AND OPTIMIZE GRADIENT L_0

BAYESIAN INFERENCE

INTEGRATING VS OPTIMIZING PARAMS IS STRONGER REGULARIZATION. BAYESIAN MODEL SELECTION FOR HYPERPARAMS IS DESIRABLE AND FASTER. UNCERTAINTY IN PARAM DESIRABLE FOR CERTAIN PROBLEMS. ONLINE INFERENCE FOR ONLINE LEARNING. LAPLACE APPROXIMATION, HYBRID MONTECARLO, VARIATIONAL BAYES.

- REGRESSION POSTERIOR: $P(w|\alpha, \beta, D) \approx N(w|w_{MP}, A^{-1})$; A IS HESSIAN OF ERROR $= \nabla\nabla(w_{MP}) = \beta H + I\alpha$
- CLASSIF POSTERIOR: SAME FOR REGR BUT $\beta = 1$ AND E IS XENTROPY
- REGRESSION PRED. POSTERIOR: $P(y|x, \alpha, \beta, D) \approx N(y|f(x, w_{MP}), \sigma^2(x))$; $\sigma^2(x) = \beta^{-1} + g^T A^{-1} g$
- CLASSIF PRED POSTERIOR: $P(y=1|x, D) = \int \mathrm{sigm}(a) P(a|x, D)\, da \approx \mathrm{sigm}(\kappa(\sigma_a^2) b^T w_{MP})$, $\kappa(\sigma^2) \approx (1 + \pi\sigma^2/8)^{-1/2}$
- ARD: ONCE LAPLACE APPROX, CAN OPTIMIZE MARGINAL LIKELIHOOD WRT HYPERPARAMS α. ONE X WEIGHT VECTOR. EFFECT SIMILAR TO GROUP LASSO. CAN PRUNE OUT IRRELEVANT INPUT OR HIDDEN FEATURES. $P(\theta) = \prod^{D} N(v|0, \frac{1}{\alpha_{vi}} I) \prod^{H} N(w|0, \frac{1}{\alpha_{wj}} I)$

RADIAL BASIS NETWORKS

MLP W/ RBF AS HIDDEN NODES. OUTPUT AS USUAL **TRAINING:** 1 - POSITION RBF NODES (RANDOM, K-MEANS, OTHER CLUSTERING) UNSUPERVISED
2 - FWD PASS
3 - UPDATE WEIGHTS (PERCEPTRON OR PSEUDOINVERSE) $G = (G^T G)^{-1} G^T$

- CAN DO REGRESSION, FCN APPROX, TIME SERIES PREDICTION, ...

SELF-ORGANIZING MAPS

IS AN ANN-LIKE THING USED FOR DIMENSIONALITY REDUCTION AND VISUALIZATION. UNSUPERVISED LEARNING. VIA VECTOR QUANTIZATION. INPUT SPACE $\longleftrightarrow$ MAP SPACE (2D/3D IND INDICES /

- **FEATURE MAPPING:** NEARBY NEURONS MAP | ACTIVATE TO SIMILAR FEATURES. INPUTS FULLY CONNECTED TO EACH MAP NEURON.
- **ALGORITHM:** SELECT BEST-MATCHING-UNIT | NEURON WRT PRESENTED INPUT.
  - UPDATE BMU WEIGHTS - UPDATE DAMPED WEIGHTS OF MAP-NEIGHBOURING NEURONS
  - REDUCE LEARNING RATE / NEIGHBOURHOOD RADIUS - UNTIL CONVERGENCE
  - WEIGHT INIT IS RANDOM OR VIA PCA OF DATA
- **SELF ORGANIZATION:** SPACE GETS GLOBALLY ORGANIZED BY ONLY LOCAL INTERACTIONS HAPPENING
- VARIANTS ON MAP TOPOLOGY | CONNECTEDNESS. LETTING THE MAP GROW OVER CYCLES
- **NEURAL GAS:** SIMILAR, BUT FEATURE VECTORS MOVE IN SPACE.

ENSEMBLE LEARNING

LEARNING A WEIGHTED COMBINATION OF MODELS $f(y|x,\pi) = \sum w_m \cdot f_m(y|x)$. COMMITTEE METHOD. NEURAL NETS, BOOSTING, CAN BE SEEN AS ENSEMBLES

- STACKING: $\hat{w} = \underset{w}{\text{ARGMIN}} \sum^{N} L(y_i, \sum^{M} w_m \cdot \hat{f}_m^{-i}(x))$

  $\hat{f}_m^{-i}$ → PREDICTOR BY REMOVING (x_i, y_i).

- ERROR CORRECTING OUTPUT CODES: ECOC, MULTICLASS CLASSIFICATION. USE MORE BIT THAN $B = \log_2 C$ FOR CLASS LABEL. MAXIMIZE HAMMING DISTANCE → RESILIENCE TO BIT-FLIPPING ERROR $\hat{c}(x) = \min_c \sum^{B} |C_{cb} - \hat{p}_b(x)|$

- BAYES MODEL AVERAGING: USE WEIGHTED AVERAGE OF PREDICTIONS MADE BY EACH MODEL. $p(y|x,D) = \sum^{M} p(y|x,m,D)\,p(m|D)$
  USUALLY INTRACTABLE, PICK FEW SAMPLE FROM POSTERIOR
  - NOT EQUIVALENT TO ENSEMBLE LEARNING → HERE WE ENLARGE THE MODEL SPACE, CONVEX COMBINATION OF BASE MODELS: $p(y|x,\pi) = \sum \pi_m p(y|x,m)$

INTERPRETATION; PERF ANALYSIS

BLACKBOX MODELS ARE NOT INTUITIVE TO INTERPRET.

- PARTIAL DEPENDENCE PLOT: $\bar{f}(x_j)$ vs x_j, WITH OTHER PREDICTORS AVERAGED OUT. $\bar{f}(x_j) = \frac{1}{N}\sum^{N} \hat{f}(x_j, x_{i,-j})$. $\hat{f}$ IS PREDICTED RESPONSE.
- RELATIVE IMPORTANCE OF PREDICTOR VARS: 'MODEL FREE' VARIABLE SELECTION. IN TREES COUNT HOW OFTEN VARS ARE USED.