Lecture 21 announcements

- Groups have been created on D2L
  - Each project team is a group (already created)
  - All students working on the collaborative Kaggle topic are 1 group - you need to sign yourself up to be included
- Next Homework will be HW 12, to be posted next week.
- Discussion 12 will be pre-recorded today at 5:00 PM in OHE 100C, and played back next week at the usual time and place. Also will be posted on D2L as usual.

Lecture 21 outline

- Finish Random Forest
- Start Boosting

RANDOM FOREST ALGORITHM

[Based on Hastie, et al., 2nd Ed, Alg. 15.1].

1. FOR $b = 1$ TO B

   (a) DRAW A SAMPLE DATASET $\mathcal{D}^*$ AT RANDOM, WITH REPLACEMENT, OF SIZE N^*, FROM $\mathcal{D}_{Tr}$. TYPICALLY, $N^* = N_{Tr}$.

   ~~(b) (ii)~~

   (b) GROW A RANDOM-FOREST TREE T_b, USING $\mathcal{D}^*$:

   CYCLE THROUGH EACH REGION R_m;

   FOR EACH R_m:

   (i) SELECT d FEATURES AT RANDOM (~~FOR~~ FROM ALL D FEATURES)

   - COMMON VALUES:

     $d \approx \sqrt{D}$, OR $d = \lfloor \sqrt{D} \rfloor$

   (ii) USE CART APPROACH TO SPLIT R_m BY FINDING OPTIMAL j, t_k, $\underline{w}$

   (iii) SPLIT THE TREE NODE INTO TWO DAUGHTER NODES.

~~(iv)~~

(iv) ITERATE UNTIL A HALTING CONDITION IS REACHED

- SEE CART HALTING CONDITIONS.

2. OUTPUT THE SET OF TREES $\{T_b,\ b=1,2,\cdots,B\}$.

3. TO USE RESULTING SET FOR PREDICTION:

REGRESSION –

$$\hat{f}(\underline{x}) = \frac{1}{B}\sum_{b=1}^{B} \underbrace{\hat{f}_b(\underline{x})}_{\text{PREDICTION FROM TREE}}$$

$$= \frac{1}{B}\sum_{b=1}^{B} \overbrace{\sum_{m=1}^{M} w_m^{(b)}\, \mathbb{I}\left(\underline{x} \in R_m^{(b)}\right)}^{T_b(\underline{x})}$$

$$= \frac{1}{B}\sum_{b=1}^{B} \sum_{m=1}^{M} w_m^{(b)}\, \phi\left(\underline{x}, \underline{v}_m^{(b)}\right)$$

STILL PIECEWISE CONSTANT, BUT A LOT MORE "PIECES" OR REGIONS THAN FOR 1 CART TREE.

CLASSIFICATION —

CLASS PREDICTION:

IF $\hat{y}^{(b)}(\underline{x})$ IS CLASS ASSIGNMENT FROM TREE T_b, THEN

$$\hat{y}(\underline{x}) = \arg\max_{c} \sum_{b=1}^{B} \mathbb{I}\left[\hat{y}^{(b)}(\underline{x}) = c\right]$$

($\Leftrightarrow$ TAKE VOTE, $\hat{y}$ = WINNING CLASS)

CAN ALSO ESTIMATE CLASS POSTERIOR PROBABILITIES $p(\hat{y}=c \mid \underline{x}, \mathcal{D})$, BY:

LET $p_c^{(b)}(\underline{x})$ = FREQ. OF OCCURRENCE OF DATA POINTS $y_i = c$ IN $R_m^{(b)}$ THAT CONTAINS $\underline{x}$, FROM TREE T_b.

AT EACH POINT $\underline{x}$, TAKE AVE:

$$p(\hat{y}=c \mid \underline{x}, \mathcal{D}) \approx \frac{1}{B} \sum_{b=1}^{B} p_c^{(b)}(\underline{x}).$$

— [FIG. 15.1, HASTIE] —

BOOSTING [MURPHY 16.4.0-16.4.4]
[ALSO - HASTIE et al., Ch. 10]

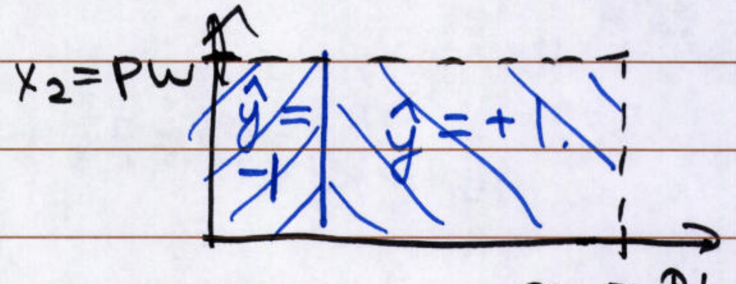
IS ALSO AN ADPTIVE BASIS FCN. MODEL:

$$\hat{f}(\underline{x}) = w_0 + \sum_{m=1}^{M} w_m \, \phi(\underline{x}; \underline{\gamma}_m) \tag{1}$$

EACH $\phi(\underline{x}, \underline{\gamma}_m)$:

- IS A SIMPLE CLASSIFIER THAT CAN CLASSIFY THE ~~ENTIRE~~ ENTIRE FEATURE SPACE.
- IS A "WEAK LEARNER" THAT IS ONLY REQUIRED TO DO BETTER THAN CHANCE.

e.g., WEAK LEARNER IS TYPICALLY A "DECISION STUMP" - A 1-STAGE CART RESULTING IN 1 NODE AND 2 LEAVES (OR VARIANTS WITH >2 LEAVES).



$\phi_m(\underline{x})$ ARE FOUND SEQUENTIALLY.

NOTATION

FOR 2-CLASS PROBLEM, LET CLASS LABELS BE ± 1.

$$(2)\quad \hat{f}(\underline{x}) = f_0 + \sum_{m=1}^{M} \beta_m \underbrace{\phi_m(\underline{x})}$$

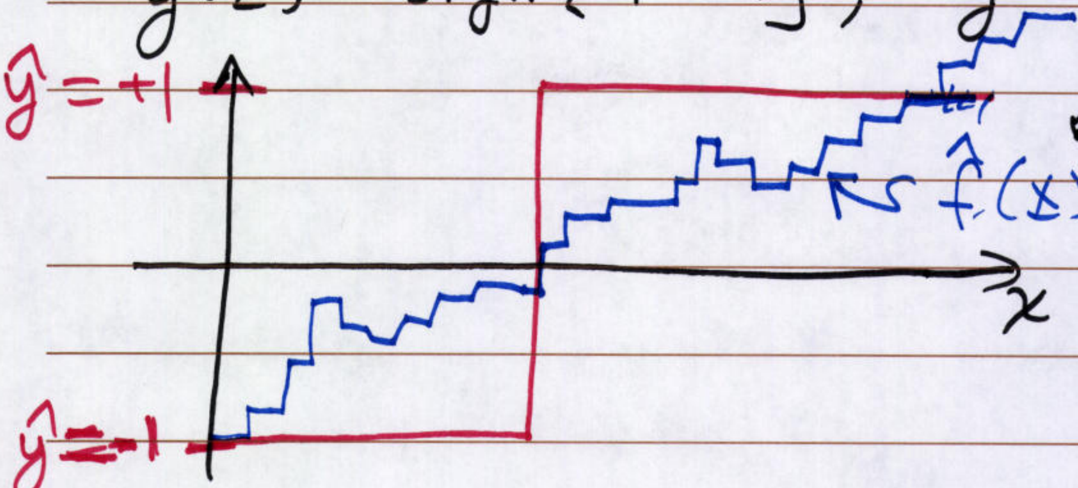
"WEIGHT" OR IMPORTANCE OF m^{th} CLASSIFIER ($\beta_m \in \mathbb{R}$)

SIMPLE BASE CLASSIFIER $\phi_m \in \{-1,+1\}$.

$$= f_0 + \sum_{m=1}^{M} \beta_m \, \phi(\underline{x}; \underline{\gamma}_m)$$

FINAL CLASSIFIER:

$$\hat{y}(\underline{x}) = \text{sign}\{\hat{f}(\underline{x})\}, \quad \hat{y} \in \{-1,+1\}$$



$\hat{f}(\underline{x})$ CARRIES MORE INFO THAN $\hat{y}(\underline{x})$.

$|\hat{f}(x)|$ PROVIDES A MEASURE OF CONFIDENCE.

INTERMEDIATE EXPRESSIONS FOR $\hat{f}$:

(3) $$\hat{f}_m(x) = f_0 + \sum_{m'=1}^{m} \beta_{m'} \phi_{m'}(\underline{x})$$

DATA POINTS: $(\underline{x}_i, \tilde{y}_i)$, $\tilde{y}_i \in \{-1, +1\}$.

TO FIND EACH $\phi_m(\underline{x})$,

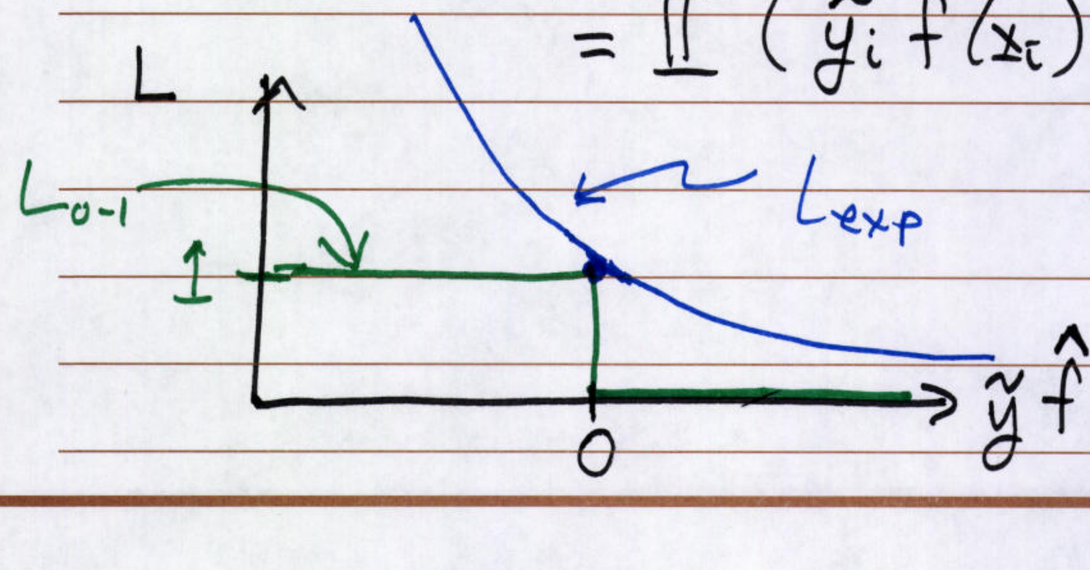
$$f_{obj}\left[\{\beta_m, m=1,\cdots,M\}, \{\underline{\gamma}_m, m=1,\cdots,M\}, \phi(\underline{x})\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} L(\tilde{y}_i, \hat{f}(x_i))$$

LOSS FUNCTION.

0-1 LOSS: $L_{0-1} = \mathbb{I}(\tilde{y}_i \neq \text{sign}\{\hat{f}(\underline{x}_i)\})$

$$= \mathbb{I}(\tilde{y}_i \hat{f}(\underline{x}_i) < 0)$$



~~Ex~~ f_{obj} using L_{0-1}: - NOT CONVEX
- NOT AMENABLE TO GRADIENT TECHNIQUES.

⇒ (4) EXPONENTIAL loss: $L_{exp} = \exp\{-\tilde{y}_i \cdot \hat{f}(\underline{x}_i)\}$

· ($\exists$ OTHER L FCNS.)

- CONVEX
- DIFFERENTIABLE
- PROVIDES CONFIDENCE MEASURE.