Lecture 20 announcements

- HW 10 is due Thursday
- HW 6 and HW 7 grading are completed. Scores will show on D2L shortly.

Lecture 20 outline

- CART (part 2)
- Variance of an average (e.g., an average of trees)
- Random Forest (part 1)

CART (part 2)

FOR CLASSIFICATION, USE A DIFFERENT COST FCN, E.G.:

$$\text{cost}\{(\underline{x}_i, y_i) \in R_{m'}\}$$

$$= \frac{1}{N_{R_{m'}}} \sum_{\underline{x}_i \in R_{m'}} \mathbb{I}\left[y_i \neq \hat{y}(R_{m'})\right]$$

CLASS ASSIGNMENT IN $R_{m'}$

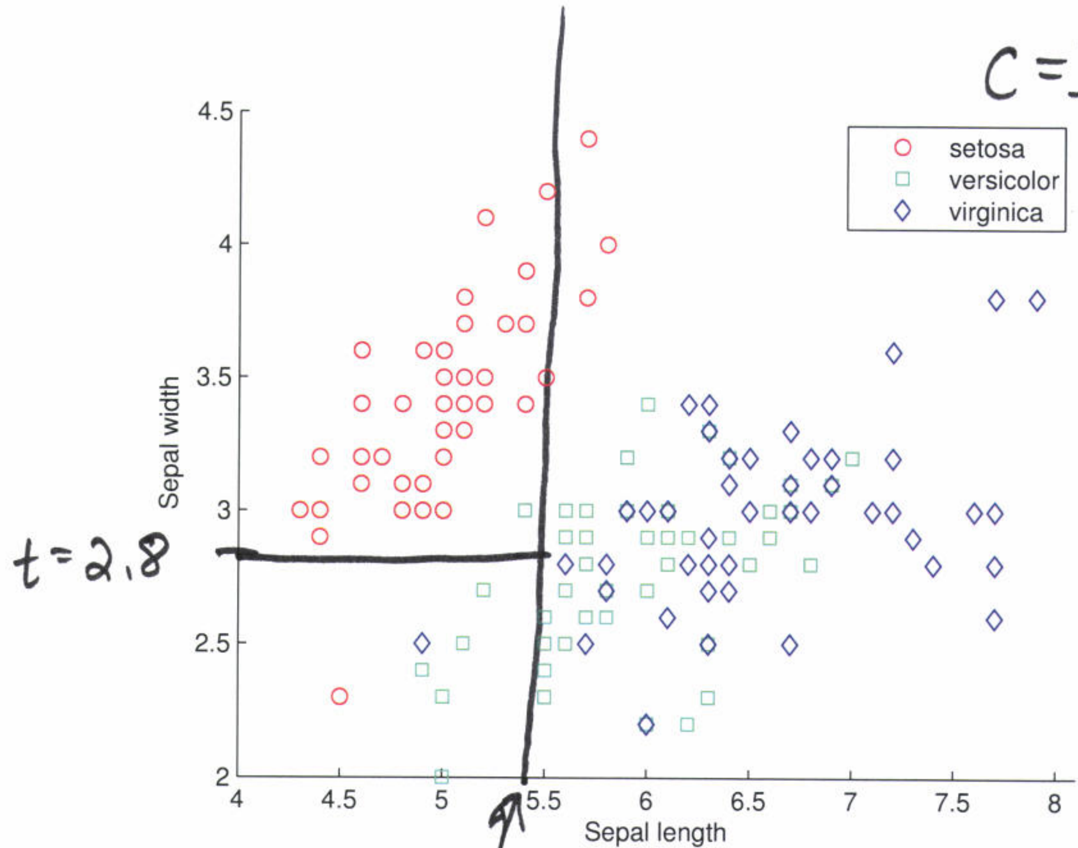
(CLASS W/ MOST DATA PTS. IN $R_{m'}$)

= CLASSIFICATION ERROR RATE IN $R_{m'}$.

OR OTHER METRIC (IN TEXT).

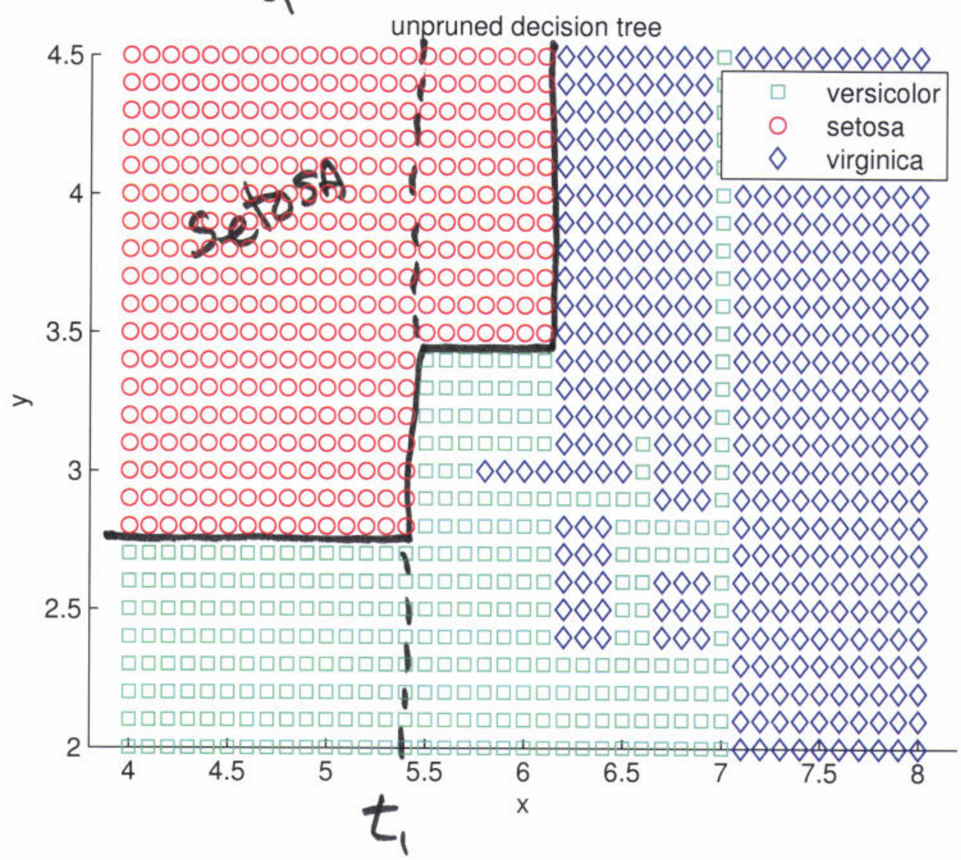
CART WITH THESE COST FCNS. WORKS FOR $C > 2$ CLASSES ALSO.

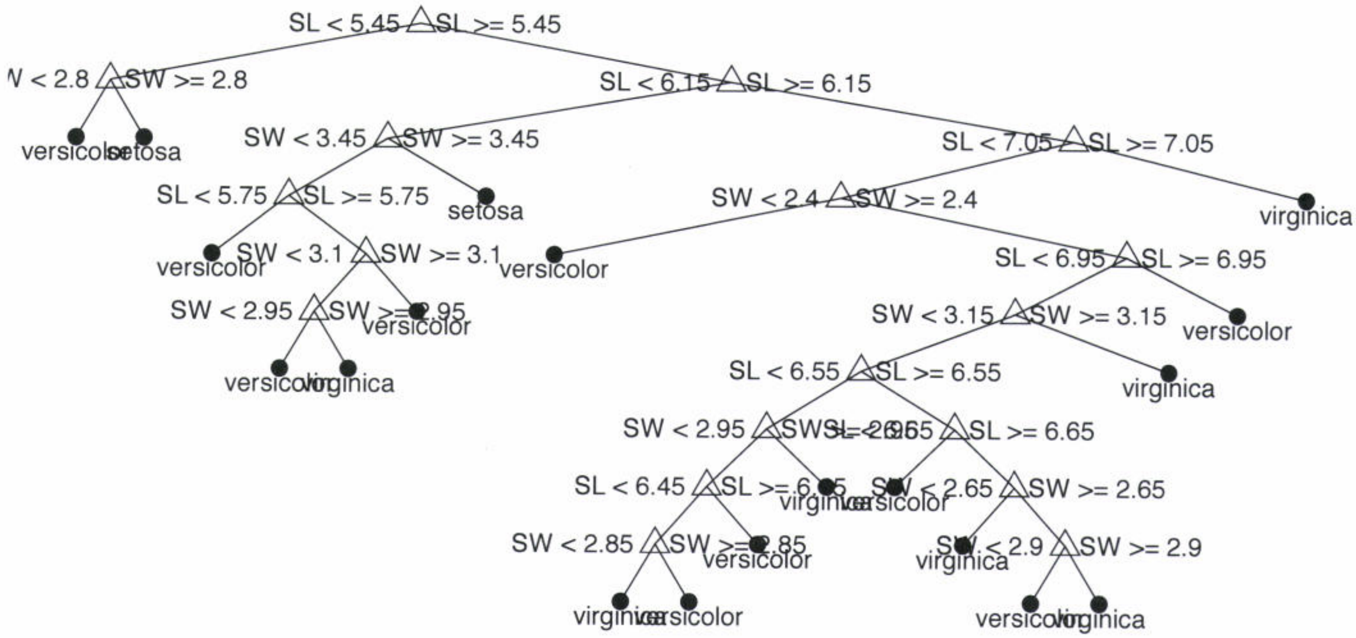
C = 3 CLASSES



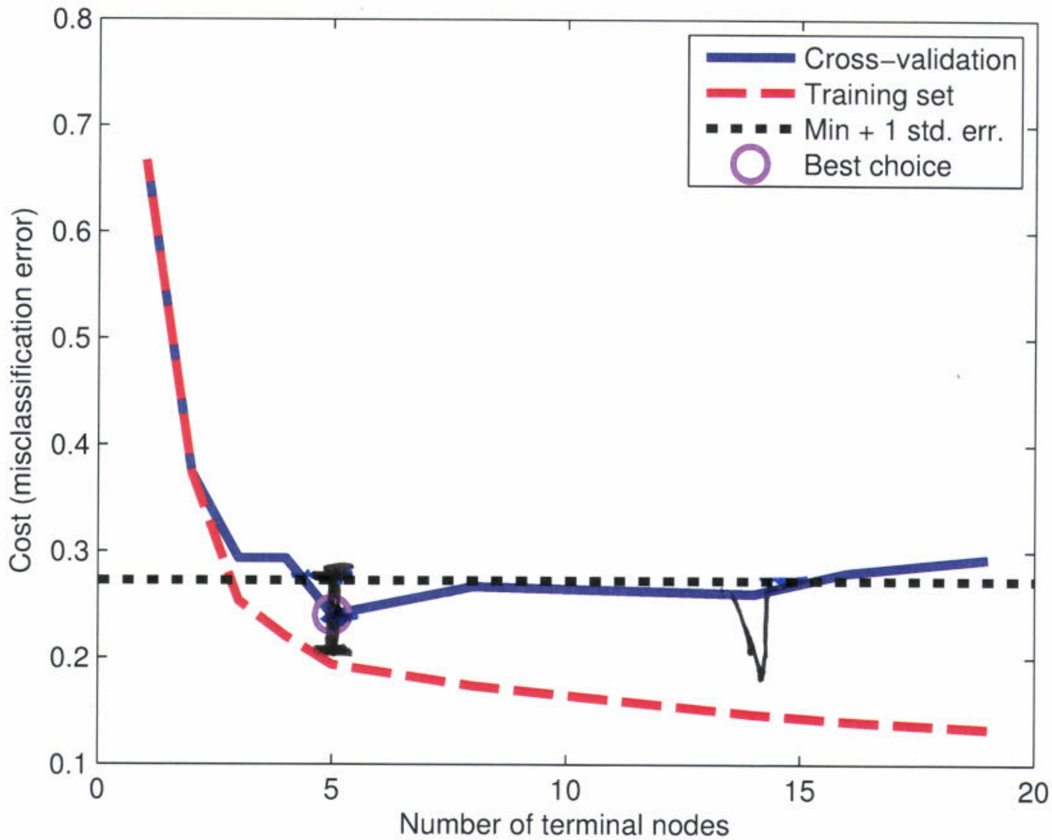
Murphy Fig. 16.4

$t_1 = 5.45$





Murphy Fig. 16.5(a)



Murphy Fig. 16.5(b)

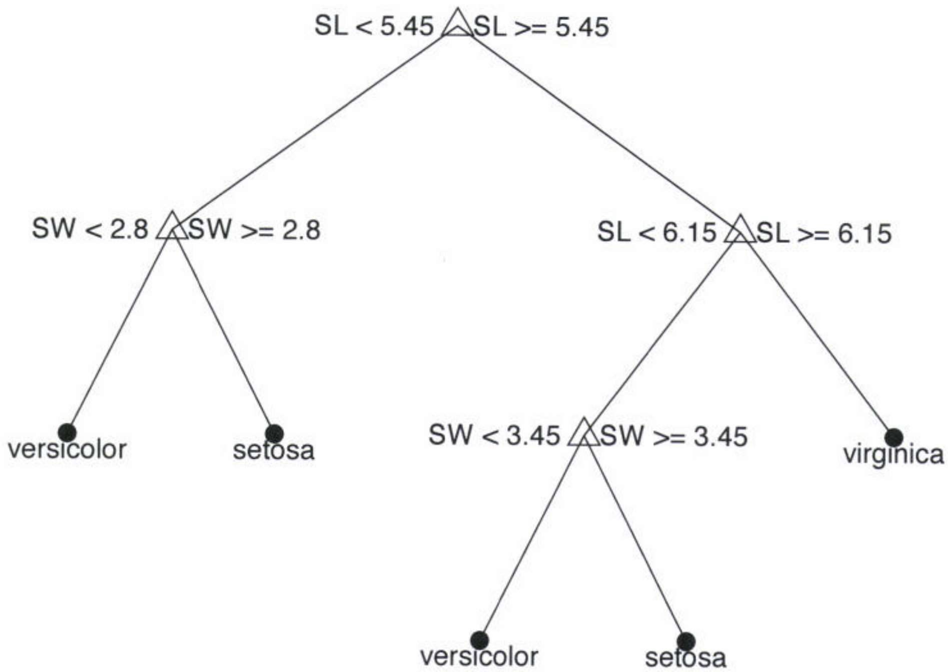
[MURPHY FIG. 16.4-16.5a]
[" " 16.5b]

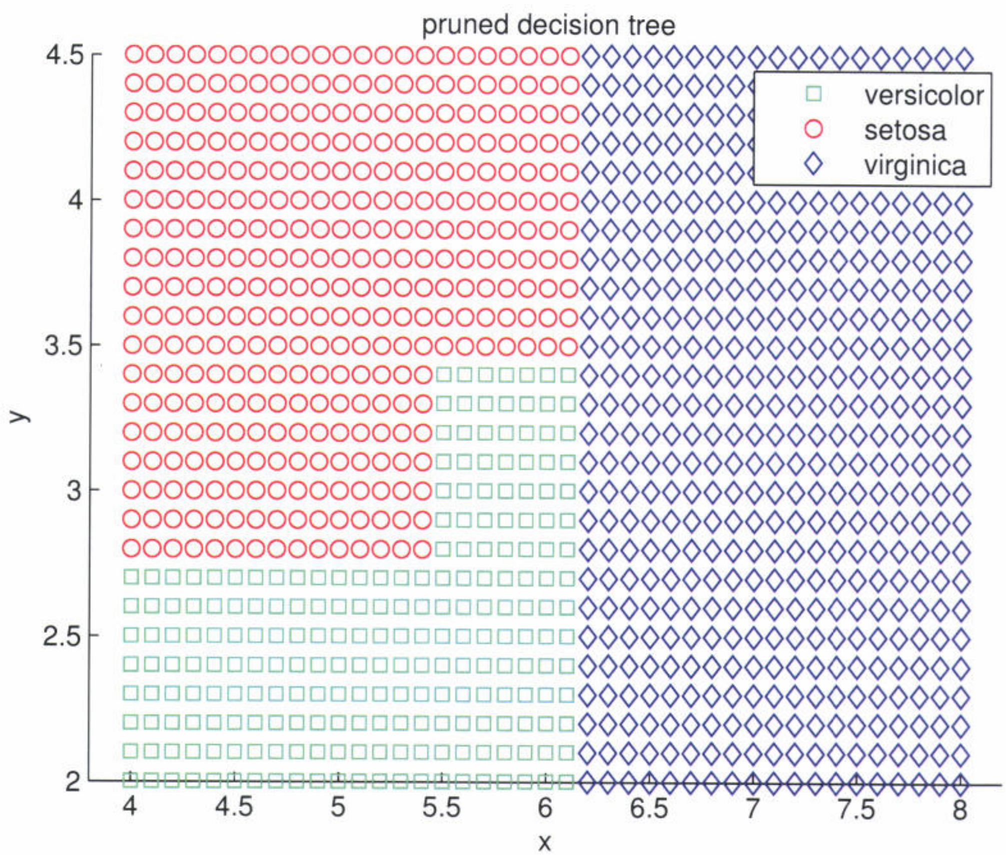
BECAUSE CART IS A GREEDY ALGORITHM, GROWING THE TREE UNTIL THE OPTIMAL STOPPING POINT TYPICALLY DOESN'T YIELD THE BEST RESULTS. USUALLY IT IS RUN PAST THIS POINT, TO YIELD A TREE THAT OVERFITS. THEN TREE IS PRUNED.

"WEAKEST LINK PRUNING":

- COLLAPSE THE INTERNAL NODE THAT GIVES THE SMALLEST INCREASE IN COST FCN.; ITERATE.
- USE CROSS-VALIDATION TO HALT WHEN MIN. VALIDATION ERROR IS REACHED. (WITHIN 1σ).

[MURPHY FIG. 16.5b, 16.6].





Murphy Fig. 16.6

CART SUMMARY [CM 16.2.4]

PROS

- ALGORITHM IS FAIRLY SIMPLE
- INCORPORATES LINEAR AND NONLINEAR BOUNDARIES ON ITS OWN.
- CAN INCLUDE SOME AUTOMATIC FEATURE SELECTION, OR RANK FEATURE IMPORTANCE.
- TYPICALLY ROBUST TO OUTLIERS.

CONS

- PREDICTIVE ACCURACY OFTEN ISN'T AS GOOD AS WITH SOME OTHER MODELS.
- CAN BE UNSTABLE TO SLIGHT CHANGES IN DATA.

$\Rightarrow$ HIGH VARIANCE.

VARIANCE OF AN AVERAGE

$$
(1)\begin{cases}
\operatorname{var}(\underline{x}) = \mathbb{E}_{\mathcal{D}}\left\{\left(\underbrace{h_g^{(\mathcal{D})}(\underline{x})} - \bar{h}_g(\underline{x})\right)^2\right\} \\[1em]
\tilde{h}_g^{(\mathcal{D})}(\underline{x}) \triangleq \dfrac{1}{B}\displaystyle\sum_{b=1}^{B} h_g^{(\mathcal{D}_b)}(\underline{x})
\end{cases}
$$

IN WHICH $\mathcal{D}_b$ IS ONE DATASET DRAWN (WITH REPLACEMENT) FROM $\mathcal{D}$, AND USED TO TO TRAIN TREE b.

NOTE THAT:

IF WE TAKE AVERAGE OF B iid RANDOM VARIABLES v_i, $i = 1, 2, \cdots, B$, EACH WITH VARIANCE σ^2, THE AVERAGE WILL HAVE VARIANCE:

$$\sigma_{AVE}^2 = \frac{\sigma^2}{B}.$$

IF, INSTEAD, THE R.V. ARE IDENTICALLY DISTRIBUTED BUT HAVE PAIRWISE CORRELATION $\rho \geq 0$, ONE CAN SHOW THAT:

$$\boxed{\sigma_{AVE}^2 = \rho\,\sigma^2 + (1-\rho)\frac{\sigma^2}{B}, \quad 0 \le \rho \le 1}$$

(2)

IN WHICH ρ = CORRELATION COEFFICIENT:

$$\rho \triangleq \frac{E\{v_i v_j\} - E\{v_i\}E\{v_j\}}{\sigma_{v_i}\,\sigma_{v_j}}$$

FOR US:

$$\boxed{\rho = \frac{E\{v_i v_j\} - \mu_v^2}{\sigma_v^2}}$$

$\Rightarrow$ WANT $h_g^{(D_b)}$ TO BE AS UNCORRELATED (BY ρ MEASURE) AS POSSIBLE, AND LARGE B.

RANDOM FORESTS

(a) DRAW MANY DATASETS FROM $\mathcal{D}_{Tr}$, WITH REPLACEMENT.

→ EACH GIVES RISE TO A TREE T_m AND EST. $\hat{f}_m(\underline{x})$.

AT EACH POINT $\underline{x}$, COULD TAKE AVERAGE RESULT (REGRESSION):

$$\hat{f}(\underline{x}) = \frac{1}{M} \sum_{m=1}^{M} \hat{f}_m(\underline{x})$$

OR COULD TAKE A VOTE (CLASSIFICATION).

→ THIS BY ITSELF IS CALLED <u>BAGGING</u> (FOR "BOOTSTRAP AGGREGATING").

BUT:

- DATASETS ARE (HIGHLY) CORRELATED
- REDUCES VAR SOME, BUT NOT A LOT.

(b) BEFORE SPLITTING EACH REGION R_m,
SELECT A RANDOM SUBSET OF d FEATURES
($d < D$ OR $d << D$),
THEN SELECT BEST FEATURE OUT OF THE
SUBSET TO THRESHOLD.

$\Rightarrow$ CORRELATION BETWEEN TREES IS
TYPICALLY MUCH SMALLER

$\Rightarrow$ REDUCES VARIANCE BY A LOT MORE.