Math 390.4 Lec 24 ~~5/8/19~~

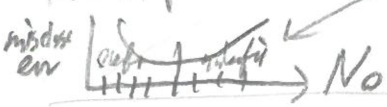
1

(I) Try all split rules. Calc Gini for each node. Select split rule with lowest weighted Gini avg:

$$\frac{n_L G_L + n_R G_R}{n_L + n_R}$$

(II) Repeat step I recursively until node has N_0. Default $N_0 = 1$.
Assign $\hat{y} = \text{Mode}[\{y_i \text{ in node}\}]$

How to pick N_0? opposite!



RF? Yes! [demo] [demo]
You use $N_0 = 1$ because overfitting is good in bagging!

Is misclass. err rule ok greedy? No...

New topic: asymmetric cost models.

There are two errors, two ways s.t. $\hat{y} \neq y$

(I) If $\hat{y} = 0$ when $y = 1$ $\Rightarrow$ False negative (FN)
(II) If $\hat{y} = 1$ when $y = 0$ $\Rightarrow$ False positive (FP)

all options can be summed up in a 2x2 table called a confusion table!

	$\hat{y}=0$	$\hat{y}=1$		
$y=0$	TN	FP	#N	num neg.
$y=1$	FN	TP	#P	num pos.
	#PN	#PP	n	sample size
	num pred neg	num pred. pos.		

(TN = true negatives, TP = true pos)

$$err := \frac{FP + FN}{n}$$

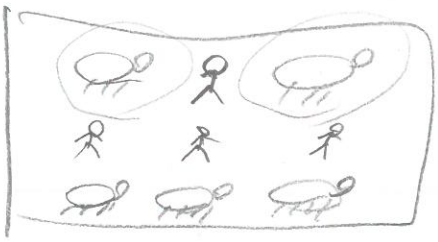
$$ACC = \frac{TP + TN}{n} = 1 - \frac{FP + FN}{n} = 1 - err$$

Many other metrics

precision $= \frac{TP}{\#PP}$ — what prop of those predicted pos. were actually pos?

recall $= \frac{TP}{\#P}$ — what prop of positives did you locate?

Task: identify bugs



$$\text{Precision} = \frac{\#TP}{\#PP} = \frac{2}{2} = 100\% \quad \text{good!}$$

$$\text{Recall} \quad \frac{\#TP}{\#P} = \frac{2}{5} = 40\% \quad \text{bad!}$$

y \ $\hat{y}$	0	1	
0	4	0	4
1	3	2	5
	7	2	9

Holistic measures:

$$\text{err} = \frac{FP+FN}{n} = \frac{0+3}{9} = 33\% \quad \text{bad!}$$

$$F_1 = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{precision}}} = \frac{2}{\frac{1}{0.4} + \frac{1}{1}} = .57 \quad \text{bad!}$$

More metrics:

false discovery rate $FDR = \frac{FP}{\#PP}$ if $\hat{y}=1$, what's your error rate?

false omission rate $FOR = \frac{FN}{\#PN}$ if $\hat{y}=0$ what's your error rate

these I find the most useful and most underrated!

All these can be used in-sample & oos.

Now... sometimes each error has different costs. Example?

$$C = c_{FP} FP + c_{FN} FN$$

(c_{FP}: cost of an FP; c_{FN}: cost of an FN)

Idea:

Instead of err, use C as the objective function when growing & splitting a tree

Another option: Stratified sampling: over or under sample the classes. This makes it easy to predict minority classes.