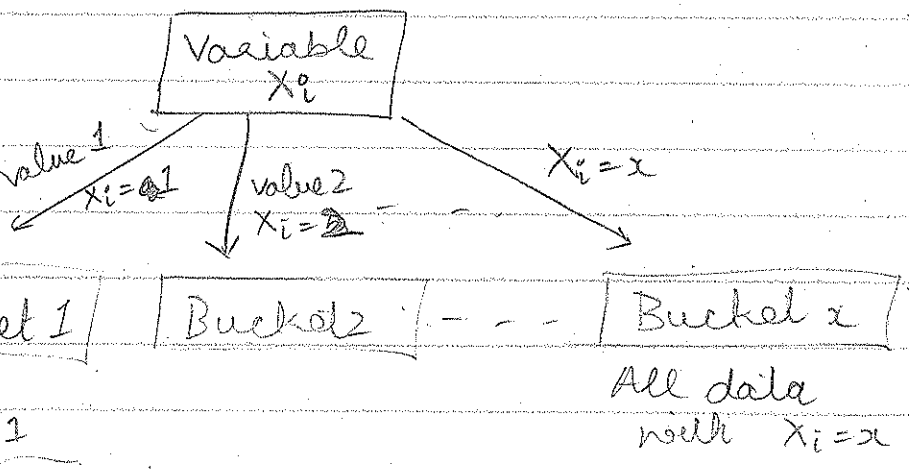
10/24/13

DECISION TREES

(1) Decision Stump



(2) Learning DTrees

learning: OPT (ie min depth) tree that makes least mistakes on training data is NP-hard.

We'll use ID3 [Iterative Dichotomizer v3]
→ greedily pick the "best" attribute/feature

→ What's the best feature?
One that reduces uncertainty!
How do we measure uncertainty?
Entropy!

Entropy:

→ Recall $H(Y) = -E[\log_2 P(Y)] = -\sum_y P(Y=y) \log P(Y=y)$

not that important could be e.

"How uncertain are we about Y"

→ Recall: Specific Conditional Entropy

$$H(Y|X=x) = -\sum_y P(Y=y|X=x) \log P(Y=y|X=x)$$

"How uncertain are we about Y if I tell you" X takes state x

→ Recall: Conditional Entropy

$$H(Y|X) = \sum_x P(X=x) H(Y|X=x)$$

"How uncertain are we about Y if I tell you some random value of X; ie if I conduct an experiment who's outcome is X=x, what will your confusion about Y be"

→ Recall Mutual Information $I(Y,X) = H(Y) - H(Y|X)$

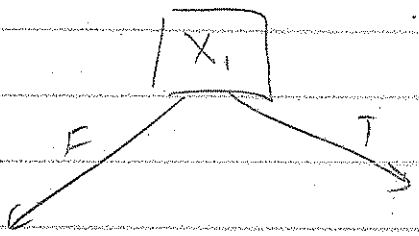
In the context of DT, sometimes people call Mutual information → Information Gain
Technically, that's not correct.
Technically $IG(X=x) = H(Y) - H(Y|X=x)$

(3) Example

X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F
F	T	F
F	F	F

$$H(Y) = -\frac{5}{8}\log\frac{5}{8} - \frac{3}{8}\log\frac{3}{8}$$

Choice 1 Split on X_1



3F 1T | 0F 4T

$H(Y|X_1 = F)$ | $H(Y|X_1 = T)$

$= -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}$ | $= -0\log 0 - 1\log 1$

$$H(Y|X_1) = \frac{1}{2}(\text{H}) + \frac{1}{2}(0)$$

$$\arg\max_X I(X) = \arg\min_X H(Y|X)$$

$$H(Y|X_1) < H(Y|X_2)$$

so split on X_1 first

(4)