CS 4780/5780 Machine Learning

Solutions 2: Decision Trees and Hypothesis Testing

Instructor: Thorsten Joachims

Problem 1: Model Selection and Validation [20 points]

(a)

Train	20/24	83.3%
Test 1	8/10	80%
Test 2	16/20	80%

(b)

Train	24/24	100%
Test 1	8/10	80%
Test 2	15/20	75%

(c)

	Lin right	Lin wrong
DT right	7	1
DT wrong	1	1

So χ^2 input is:

$$\frac{(b-c)^2}{b+c} = \frac{0}{2} = 0$$

and output (chance of accepting null) is then 100%, so 0% conf one's better.

(d)

	Lin right	Lin wrong
DT right	13	2
DT wrong	3	2

So χ^2 input is:

$$\frac{(b-c)^2}{b+c} = \frac{1}{5} = 0$$

and output (chance of accepting null) is then 80.7%, so 19.3% conf one's better.

Problem 2: Model Averaging with Decision Trees [40 points]

(a) *Single Decision Tree*

Training a full decision tree (no early stopping) on the provided data (with integer splits only) should results in a similar decision boundary (below). Note that there might be some variation based on how you broke ties.

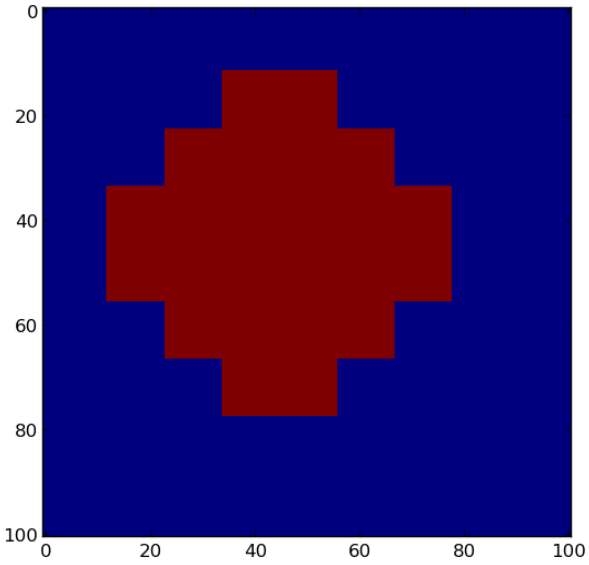


Figure 1: Part a. Full decision tree decision boundary (x,y scale/10)

(b) *Individual Trees*

Below are two examples of trees that clearly overfit the training data. Some of the trees produced (among the 101 trees in the ensemble) are actually quite good approximation of the true function, and don't constitute examples of overfitting. In your submission we were looking for clear examples of trees that overfit. Each submission would have different examples, since the training subsets were drawn randomly.

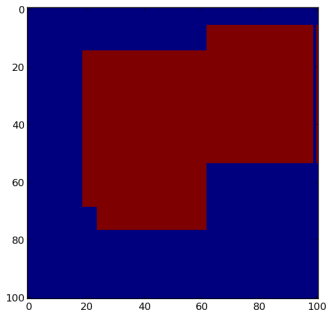


Figure 2: Part b. Overfitting example

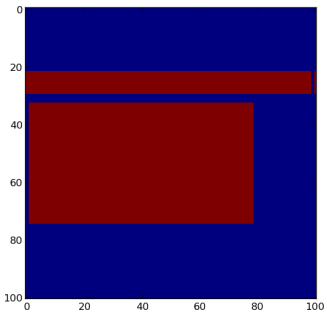


Figure 3: Part b. Overfitting example

(c) *Combined (Ensemble) Tree*

The decision boundary of the combined classifier is below. Your boundary may differ, however, the resulting boundary should be better than any of the individual trees produced.

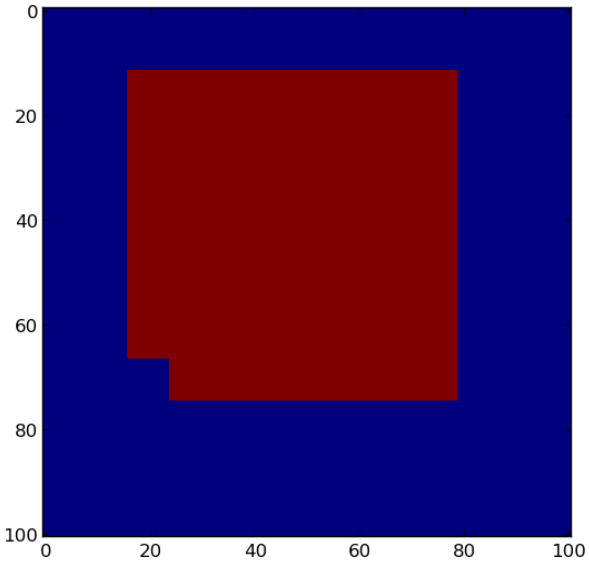


Figure 4: Part c. Combined trees (ensemble) decision boundary.

(d) We have to show that a prediction $\hat{y} \in \{-1, 1\}$ of a decision tree T on test instance $\mathbf{x}_{test}$ can be expressed as follows:

$$\hat{y} = \text{sign}\left(\sum_{i}^{n} y_i K(\mathbf{x}_{test}, \mathbf{x}_i)\right)$$

Consider a prediction of decision tree T on instance x_{test}. T will output a label $\hat{y}$ for instance x_{test} based on the majority of the training instances x_i that ended up in the same leaf as x_{test} (TDIDT, C4.5). By definition of K, weights for all training instances not in the same leaf as x_{test} are zero. All remaining $y_i K(x_i, x_{test})$ must sum to a value whose sign represents the class $\hat{y}$ with the majority of training instances in the leaf of x_{test}.

(e) For the case of multiple trees, we want to show that the prediction for the ensemble given by:

$$\hat{y} = \text{sign}\left(\frac{1}{M}\sum_{j}^{M}\sum_{i}^{n} y_i K_j(\mathbf{x}_{test}, \mathbf{x}_i)\right)$$

can be written as the linear combination:

$$\hat{y} = \text{sign}\left(\sum_{i}^{n} y_i \tilde{K}(\mathbf{x}_{test}, \mathbf{x}_i)\right)$$

It is sufficient to see that if we switch the order of summation:

$$\hat{y} = \text{sign}\left(\sum_i^n y_i \left\{\sum_j^M \frac{1}{M} K_j(\mathbf{x}_{test}, \mathbf{x}_i)\right\}\right)$$

where we define $\tilde{K} = \sum_j^M \frac{1}{M} K_j(\mathbf{x}_{test}, \mathbf{x}_i)$. This quantity is the *average weight* across different trees, assigned to training instance i, on test instance x_{test}. Note that the $1/M$ is irrelevant for the case of classification, however, it plays an important role in the case of regression trees.

Problem 3: Text Categorization with Decision Trees [40 points]

(a) Test error: 21.9%

(b) Top 2 layers: god, car, windows

Bot 2 by depth: for, nntp-posting-host, you

Bot 2 by prox. to leaf: child, consider, whose, bike, under, worth, every, special, companies, math, enjoy, direct, likely, even, errors, christ, hate, usenet, above, frank, ever, told, widget, here, kids, i'm, punishment, univ, would, call, tell, phone, excellent, must, me, md, work, mi, mr, my, want, keep, eng, mouse, how, law, appreciate, order, before, then, one, each, series, re, got, little, wanted, days, onto, nhl, tom, john, took, seen, germany, dod, appreciated, do, de, new, said, bitnet, we, never, against, players, com, three, been, life, rutgers, it, if, suggest, just, lcs, for, post, available, fight, ford, na, let's, did, turns, team, guy, nntp-posting-host, baseball, car, cwru, didn't, means, years, he's, window, not, year, cars, besides, org, that, saw, slow, review, article, case, these, running, early, using, server, on, or, there, file, you, very, tek, jesus, christian, bible, reply-to, that's, it's, held, pens, cross, double, writes, pretty, his, him, please, probably, edu>, church, else, look, while, some, up, us, edu, win, private, use, usa, biblical, high, something, university, ancient, have, who, at, thought, book, e-mail, i've, hi, corporation, christians

The words at the top level are likely to generalize well to data from the same distribution, as these words are very good discriminators of the article's group. Words on the bottom levels, however, are for the most part words that could appear in any of the categories, including stopwords.

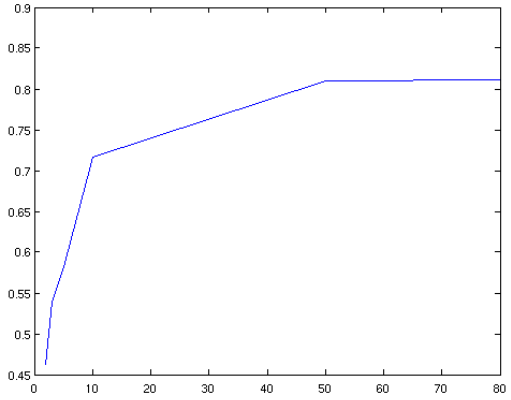
(c) Train error: 24.8%
Test error: 27.4% (yes, goes up)

(d) Unlimited makes 113 unique errors, limited to 10 makes 224

So McNemar's is $\frac{(113-224)^2}{113+224} = 36.6$, which gives almost 100% confidence models are different.

(e) T-Test accepts null hypothesis with near 100% probability. So, doesn't beat 95%

(f) Example plot:



Different splits may give different results, but one of the largest 2 sizes should perform best, and curve should look mostly the same (curves up and levels off)