MACHINE LEARNING-LAB1

CSL 603

SEPTEMBER 4, 2017

SARTHAK GUPTA
2015CSB1029

Experiment -1

The decision tree was built using 1000 training samples and 5000 attributes taken on the base of polarity from imdbEx.feat. Decision Tree so formed has 89.9 accuracy on training dataset and 65.1% on test data set.

The following attributes were selected based on the best split and information gain. Also Decision tree is formed using continuous values for words rather than discretizing them.

Worst	Lame	Uninteresting
Avoid	Effect	White
Waste	Poor	Worse
Boring	Crap	My
Poorly	Wasted	Draw
Badly	Embarrassed	Inept
Ridiculous	Thus	Amateur
terrible	Pile	missed
Briefly	Tedious	Feature
Photos	Face	Mystery
a	awful	self
First	Costs	Question
insult	horrible	trash
Stupidity	Discernable	fantasies
Idiots	too	josh
Murderer	amnesia	Included
Now	stinker	Heard
acted	skip	parts
place	bother	common
you	things	context
lousy	snoozer	Time
every	latest	Amateurish
with	boggy	Good
extremely	lowest	Functions
relief	historical	Rip off
insomniac	nauseating	beginning

Experiment -2

Now I choose information gain as the stopping criteria. if the information gain is larger than the given threshold than no further subtree will formed.

Threshold	Train Accuracy in %	Test accuracy in %	Attributes Used
0.1	82.4	71.6	worst, avoid, waste, boring, poorly, badly, ridiculous, terrible, lame, effect, poor, crap, wasted, embarrassed, thus, pile, uninteresting, white, worse, my, draw, inept, amateur, missed, briefly, photos, a, tedious, face, awful, feature, mystery, self, first, insult, costs, horrible,

			question, trash, stupidity, idiots, discernible, two, fantasies, josh, murderer, amnesia, included, now, stinker, heard, acted, skip, parts, are, with, lady, which, for, it, beginning, nice, one, red, the, time, when, you, the
0.8	88.9	64.8	 worst, avoid, waste, boring, poorly, badly, ridiculous, terrible, lame, effect, poor, crap, wasted, embarrassed,

			thus, pile, uninteresting, white, worse, my, draw, inept, amateur, missed, briefly, photos, a, tedious, face, awful, feature, mystery, self, first, insult, costs, horrible, question, trash, stupidity, idiots, discernible, two, fantasies, josh, murderer, amnesia, included, now, stinker, heard, acted, skip, parts, are, with, lady, which, for, it, time,

			beginning, nice, one, red, the, from, unfunny, nauseating, dull, ending, comments, fights, help, does, way, insomniac, we, time, when, you, the, fan, good, ending, think, it, idiotic, good, it, way, the, nasally, ripoff, are, when, good, way, turkey, the, horrible, historical, relief, functions, from, way, nutters, the,

			it, extremely, way, lousy, time, things, one, my, good, boggy, for, part, with, with, my, amateurish, help, latest, every, time, the we a, snoozer, context, for things, you, the, common a, bo,ther it, stupidity, place
1.3	89.9	65.1	worst, avoid, waste, boring, poorly, badly, ridiculous terrible lame, effect, poor, crap,

			wasted, embarrassed, thus, pile, uninteresting, white, worse, my, draw, inept, amateur, missed, briefly, photos, a, tedious, face, awful feature, mystery, self, first, insult, costs horrible question, trash, stupidity, idiots discernible two fantasies josh murderer amnesia included, now, stinker, heard, acted, skip, parts, are, with, lady, which, for,

			it, time, beginning, nice, one, red, the, from, unfunny, nauseating, dull, ending, comments, fights, help, does, way, good, unintentional, 89.9 65.1
2	89.9	65.1	insomniac, we, however, time, when, you, the, fan, good, ending, think, it, it, idiotic, good, it, way, the, raise, nasally, ripoff, time, are, when, it, good,

			way, turkey, the, good, horrible, historical, relief, functions, time, from, way, lowest, nutters, the, it, it, extremely, way, lousy, time, things, one, my, good, it, boggy, for, part, one, with, with, worse, my, amateurish, velde, help, latest, every, time, the, we, a, snoozer, lousy, context, for, things,

			you, the, common, a, bother, it, poorly, stupidity, place,
2.3	89.9	65.1	insomniac, we, however, time, when, you, the, fan, good, ending, think, it, it, idiotic, good, it, way, the, raise, nasally, ripoff, time, are, when, it, good, way, turkey, the, good, horrible, historical, relief, functions, time, from, way,

			lowest, nutters, the, it, it, extremely, way, lousy, time, things, one, my, good, it, boggy, for, part, one, with, with, worse, my, amateurish, velde, help, latest, every, time, the, we, a, snoozer, lousy, context, for, things, you, the, common, a, bother, it, poorly, stupidity, place, 65.1 ----------

			avoid waste boring poorly badly ridiculous terrible lame effect poor crap wasted embarrassed thus pile uninteresting white worse my draw inept amateur missed briefly photos a tedious face awful feature mystery self first insult costs horrible question trash stupidity idiots discernible two fantasies josh murderer amnesia included

			now stinker heard acted skip parts are with lady which for it time beginning nice one red the from unfunny nauseating dull ending comments fights help does way good unintentional insomniac we however time when you the fan good ending think it it idiotic good it way

			the raise nasally ripoff time are when it good way turkey the good horrible historical relief functions time from way lowest nutters the it it extremely way lousy time things one my good it boggy for part one with with worse my amateurish velde help latest every

			time the we a snoozer lousy context for things you the common a bother it poorly stupidity place

EXPERIMENT 3

One benefit of adding noise to training data is to prevent overfit. Following results were obtained:

Noise	Test Accuracy
0.5%	65.1
1%	65.3
5%	64.1
10%	62.7

Observations:

The quality of tree doesn't alter much even on increasing the % of noise. There is no such trend. Small reduction in size is observed on adding noise

EXPERIMENT-4

Pruning is a technique that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. It reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting

Word	Test Accuracy
time	65.4
Unintentional	65.5
Beginning	65.6
ending	65.7
it	66.2
Fan	66.5
Rip off	66.6
Are	66.8
Discernable	66.9
Horrible	67.2
Cost	67.4
First	68
Way	68.2
Good	68.3
Nutters	68.4
The	68.6
Thing	68.7
Lousy	68.8
Good	69.2
For	70.1
Context	71.3
Things	71.1
For	72.6
Too early	72.7
Place	73.2
It	73.9

Final accuracy of test set after pruning is 73.9%

Experiment 5

Random forest is another way to avoid overfitting on the training data as it has multiple tree each built on different set of features. Another benefit of Random forest is the power of handle large data set with higher dimensionality (dimensionality reduction). In the experiment Random forest was built using Feature bagging

Number of trees	Accuracy in %
7	63.5
10	64.1
20	75.9
30	79.7

CONCLUSION

- Early stopping criteria w.r.t information gain shows that as we increase the threshold shows no particular trend out for accuracy of training and test set. comes constant for many of them
- As we increase the noise we observe no particular trend though it seems that accuracy first increases by very small amount and then decreases.
- We perform post-pruning and observe that accuracy for test set increase by 8.8%.
- In decision forest as the number of trees in decision forest increases accuracy for test set increases.