Class, here is the promised info on the final exam.

1) The final will have two parts. Part I covers the material from the midterm and Part II the material after the midterm. Part 1 counts for 10% of your final grade, Part 2 counts for 18% of your final grade. As indicated on the syllabus the final counts for 28% of your final grade. I had always said the exam is comprehensive but the material after the midterm will count for more since it was not tested previously. Everyone must complete both parts and do so in the two hour exam period. There is not separate time for each part. — should I redo the midterm?? Probably , yes I am going to say yes

2) If you do better on Part I then the midterm, your grade will be positively impacted. There is nothing that you need to do.

3) The distribution of questions is approximately as follows:

Part 1: 7 short answer, 2 multiple choice, 6 T/F and 1 decision tree splitting question.

Part II: 8 short answer, 5 multiple choice, 10 T/F, 1 Naive Bayes long, 1 association rule mining long

About 20% of Part I is the one long answer question and 1/3 of the second part is from the long answer questions.

As before, this is closed book and no calculators etc are allowed or needed.

OK so last night I did the two practice finals. They were useful.

I have tried the long questions it probably would not hurt to do them again along with.

Ok I think I have talked myself through the association analysis and naive bayes

Support == item set/ total # of rows

Start with one item, check for support

Move one to two. Order of sets does not matter so just start at one and combine it with everything below repeat till done

And continue until done, once you are combining two or more items only there last item should differ

Once you get to three, after combing make sure you check that all the subsets exist, apriori rule, before counting support.. It saves you time, and makes that algorithm can run

Confidence so when done you check confident

All items unioned support / LHS support (before arrow)

So larger number set of items support (generraly smaller but never bigger) / before the arrow support, less items so usually bigger and never smaller

Naive bayes

P(A|C) probability of attribute given condition which is simply# of rows A & C/ # of rows C

SOme mostly remember the condition is on the bottom, attribute and condition on top

I did attribute on bottom for the homework, pretend the | is a \ and p(A|C) becomes easier

What we want in the end is P(C|A) =[ P(A|C) P(C)] / P(A) but there can be a lot of attributes, and we just pretend they are all independent and get the total by multiplying them also since we are going to be comparing to this to another example with the same attributes we forget about P(A) and we just see which P(A|C) P(C) is bigger

After that is you want you can divide the two to see how much more likely one is than another

That’s it

Gini – follow formula

1- probability of one class squared - probability of the other class squared probabibilty is liiteally just number of times one class occurs / all rows

Then once split is made, calculat gini for each split, multiple each gini by it’s proportion or total, ie is the split has two rows and there are a tidal of 10 rows it is 2/10 and then add them together that’s it, if the split is a little lower than you made progress if not don’t do it

Error rate, you can just count errors and divide it ny the tidal rows, even with a split, just sat this split got me three errors the other split got me ten errors add them up and divide bty total number of rows

Entropy - probability of class times log 2 of probability of class - probability of other class time log 2 of probability of that class… and can keep going for all the classes

Since probability is between 0 and 1 the log 2 will always be negative or 0 so you are essentially just adding up the entropies of all the classes

And so you get the total entropy of the stent then when you split it you get teh entropy of each split and multiple each of the entropies by the proportion of the split (like ginis) and them together, but the you subtract from the overal entropy to get information gain, so like ginis but a little different