Christoph F. Eick

Homework2 COSC 6368 Fall 2017

Deadline: Friday, December 1, 11p

1) Information Gain, Entropy and Learning from Examples in General [10]

a) Assume we have a classification problem involving 3 classes: professors, students, and staff members. There are 700 students, 200 staff members and 100 professors. All professors have blond hair, 50 staff members have blond hair, and 350 students have blond hair. Compute the information gain of the test *“hair-color=’blond’”* that returns true or false. Give the formula you used to compute the information gain as well as the actual value! Use H as the entropy function in your formula (e.g. H(1/3,1/6,1/2) is the entropy that 1/3 of the examples belong to class1 1/6 of the examples belong to class 2, and half of the examples belong to class 3). [7]

The test splits the (700, 200, 100) class distribution into (350, 50, 100) and (350,150,0)

H(0.7,0.2,0.1)

log2(1/0.7)+0.2\*log2(5)+….)=*compute the exact value*

b) What is the “intuitive idea” that underlies ID3’s information gain heuristic --- what attributes does it prefer?

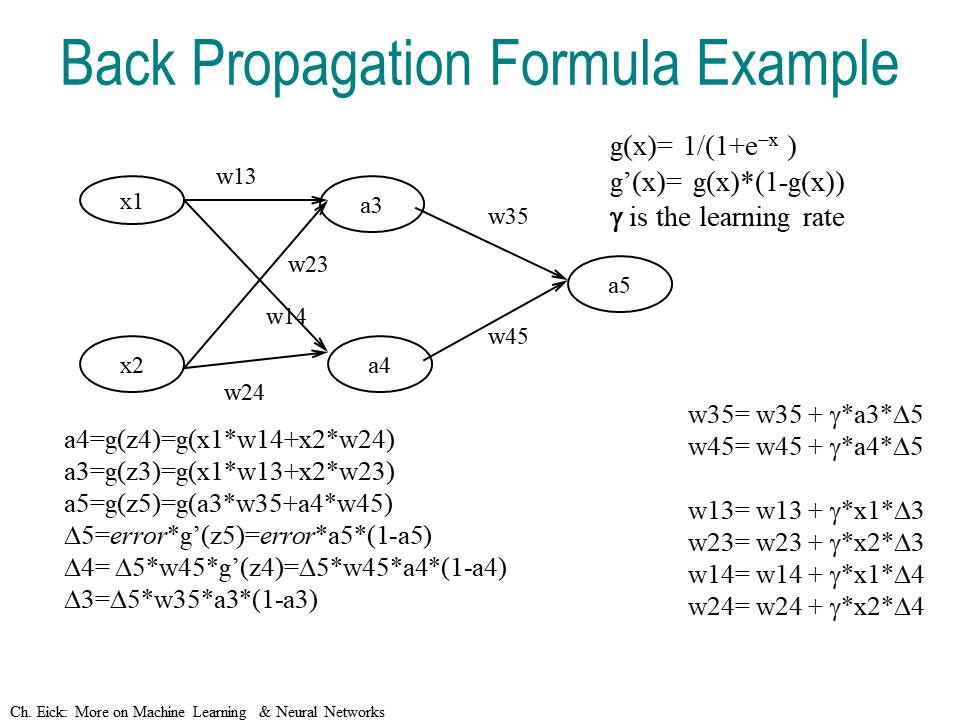
Attributes that reduce the entropy the most; that is, attributes that make the class distribution more uneven, more pure after the split; ideally, after the split the examples should all belong to the same class. [2]

c) Assume you have learnt a decision tree from a training set and you observe overfitting; what measures could you take to reduce overfitting? [1]

prune the tree reducing model complexity; increase the number of training examples.

2) Backpropagation [12]

Apply the Backpropagation algorithm to the neural network, given below:



assuming that all weights of the depicted NN are 0.5, except w14 is 0.1, and that the learning rate is 0.5, and the training example is (x1=1,x2=1;a5=1), and g is the sigmoid function. How are the weights updated by the backpropagation algorithm for this training example and what value for a5 is obtained for input x1=1 and x2=1 after the weight update?

3. Bayes’ Theorem [8]



Thomas Bayes ≈1740

1. Assume we have 3 symptoms S1, S2, S3 and the following probabilities: P(D)=0.02 P(S1)=P(S2)=P(S3)=0.01; P(S1|D)=0.1; P(S2|D)=0.02; P(S3|D)=0.002. How would a naïve Bayesian system compute the following probability [2]?

P(D|S1,S2,S3)= P(D)\*10\*2\*0.2=P(D)\*4=0.08

b) Now assume the following additional knowledge has become available: P(S1,S2)=0.0002; P(S3|S1,S2)=0.08; P(S1,S2,S3|D)=0.000032; how would you use this information to make a “better” prediction of P(D|S1,S2,S3)? [4]

P(D|S1,S2,S3) = P(D,S1,S2,S3)[1/P(S1,S2,S3)]

P(D,S1,S2,S3) = P(D)\*P(S1,S2,S3|D))

P(S1,S2,S3) = P(S3|S1,S2) P(S1,S2)

P(D|S1,S2,S3) = 0.000032 x 0.02 / (0.08 x 0.0002) = 0.04

c) How can the discrepancy in the prediction in the cases a) and b) be explained? Why are the numbers you obtain different? What does this discrepancy tell you about naïve Bayesian systems in general? [2]

In the example, we overestimate the exact probability by 0.04.This discrepancy is due to the assumption of independence assumptions made by Naïve Bayesian approaches. Making these assumptions simplifies computing the condition probabilities and reduces knowledge acquisition, but we can “over-count”—as we do in the particular example— or “undercount” the contribution of each Si, and the result could be far from the true value. With extra information, as in (b), we do not need to make any assumptions. Full knowledge of the joint distribution will guarantee the exact computation of any conditional probability without encountering any prediction errors.

4) Computations in Belief Networks [13]

Assume that the following Belief Network is given that consists of nodes A, B, C, D, and E that can take values of true and false.

a) Using the given probabilities of the probability tables of the above belief network (D|C,E; C|A,B; A; B; E) give a formula to compute P(D|A). Justify all nontrivial steps you used to obtain the formula! [8]

P(D|A)=P(D,A)/P(A) *Definition of conditional probability.*

As P(A) is given we still need to compute P(D,A)

P(D,A)=P(D,A,C,B)+P(D,A,~C,B)+P(D,A,C,~B)+P(D,A,~C,~D)

P(D,A,C,B)=P(A,B)\*P(C|A,B)\*P(D|C,A,B) *Definition of “And”*

As A and B are d-separable (pattern 3) and D is d-separable from A given evidence C, and D is d-separable from B given evidence C (based on pattern 1a/1b); that is A and B can be dropped from the conditional probability formula (P(D|C)=P(D|A,B,C)); therefore we can simplify the computation of P(D,A,B,C) as follows:

P(D,A,C,B)=P(A)\*P(B)\*P(C|A,B)\*P(D|C) As all these probabilities can be found in the probability tables of the above belief network, we are done!

For the same reasons

P(D,A,~C,B)=P(A)\*P(B)\*P(~C|A,B)\*P(D|~C)

And the remaining 2 probabilities can be computed similarly.

b) Are C and E independent; is C|∅ and E|∅ d-separable? Give a reason for your answer! ∅ denotes “no evidence given” [2]

yes

There is only one path C-D-E; as D is not in evidence this path is blocked in node D due to pattern 3

c) Is E|CD d-separable from A|CD? Give a reason for your answer! [3]

yes

The is only one path A-C-D-E; as C is in evidence this path is blocked in node C[[1]](#footnote-1) due to pattern 1a/1b

1. It is not blocked in D, but this is not relevant because the path needs only be blocked in one node of the particular path. [↑](#footnote-ref-1)