Christoph F. Eick

Homework2 COSC 6368 Fall 2017

Deadline: Friday, December 1, 11p

1) Information Gain, Entropy and Learning from Examples in General w=3

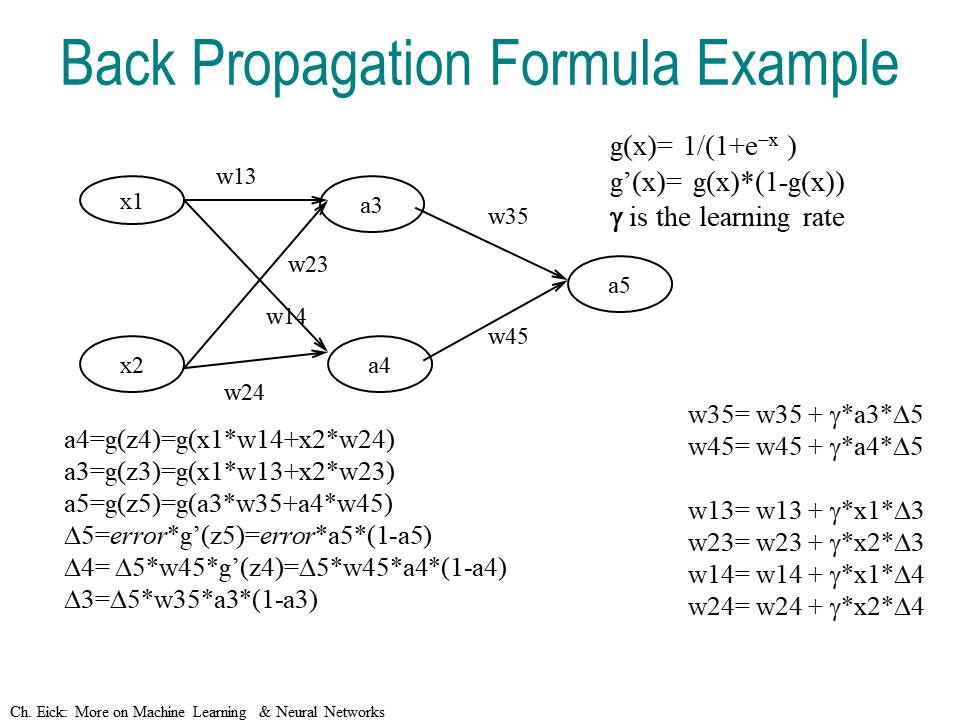
a) Assume we have a classification problem involving 3 classes: professors, students, and staff members. There are 700 students, 200 staff members and 100 professors. All professors have blond hair, 50 staff members have blond hair, and 350 students have blond hair. Compute the information gain of the 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).

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

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

2) Backpropagation

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



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)=…

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)? [3]

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?

4) Computations in Belief Networks

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! [7]

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

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