

Problem assignment 7

Due: Friday, March 26, 2021

In this assignment we continue our investigation of the "Pima" dataset. As in the previous assignment, you can download the dataset (*pima.txt*) and its description (*pima_desc.txt*) from the course web page. In addition to the complete dataset *pima.txt*, you have *pima_train.txt* and *pima_test.txt* you will need to use for training and testing purposes. The dataset has been obtained from the UC Irvine machine learning repository:

<http://www1.ics.uci.edu/~mlearn/MLRepository.html>.

Problem 1. Decision trees

The decision tree approach is yet another classification method we covered in the course. The method builds a tree by recursively splitting the training set using one of the attributes by optimizing the gain with respect to some impurity measure.

- Part a. The script *run_DT.m* shows how to train, display, and apply the decision tree in Matlab. The script first builds a default tree with minimal restrictions on its size, and after that the tree obtained by restricting the number of nodes in the tree. Please run and familiarize yourself with the code. What do you think, which tree is better for prediction, the unrestricted or restricted tree? Why? Should we always try to backprune it?
- Part b. Experiment with the decision tree function *fitctree.m* and its optional parameters, modifying the algorithm and the tree built. Report the results of your investigations in the report by listing the settings used for the tree learning algorithm and obtained results. You can find the different settings in the matlab help documents.

Note: no code should be submitted for Problem 1.

Problem 2. Conditional independence

Random variables A, B are conditionally independent given C when:

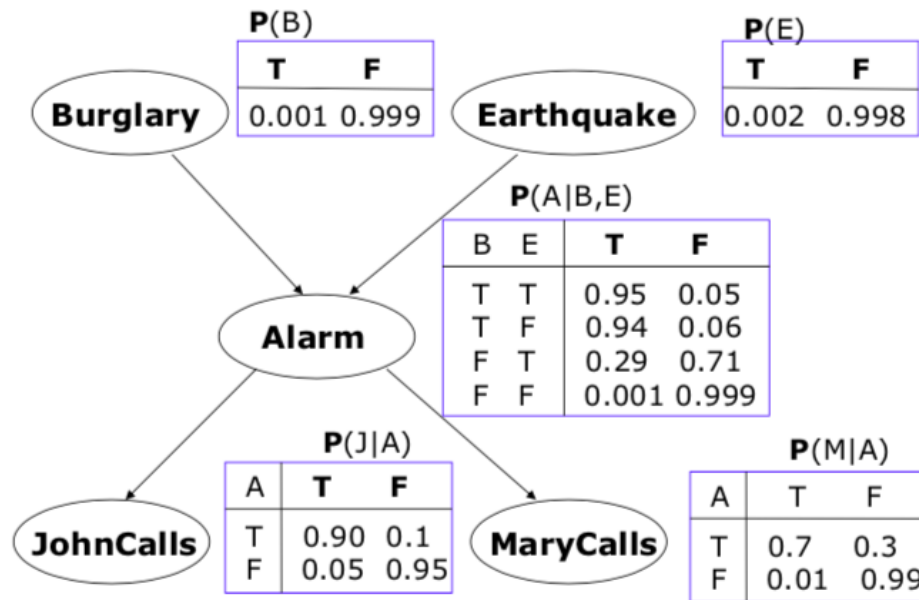
$$P(A, B|C) = P(A|C)P(B|C).$$

Prove that this implies:

$$P(A|B, C) = P(A|C).$$

Problem 3. Bayesian belief networks

Assume the Alarm network from Lecture 16 that is shown the figure below.



Part a. Express the joint probability $P(E = F, B = F, A = F, J = T, M = F)$ in terms of local conditional probabilities parameterizing the Bayesian belief network.

Part b. Calculate the probability $P(E = F, B = F, A = F, J = T, M = F)$ by substituting the local conditional probabilities to numbers.