Assignment 2 Regression and Bayes Net (75 points)

CSCE A415 and CS F480: Machine Learning

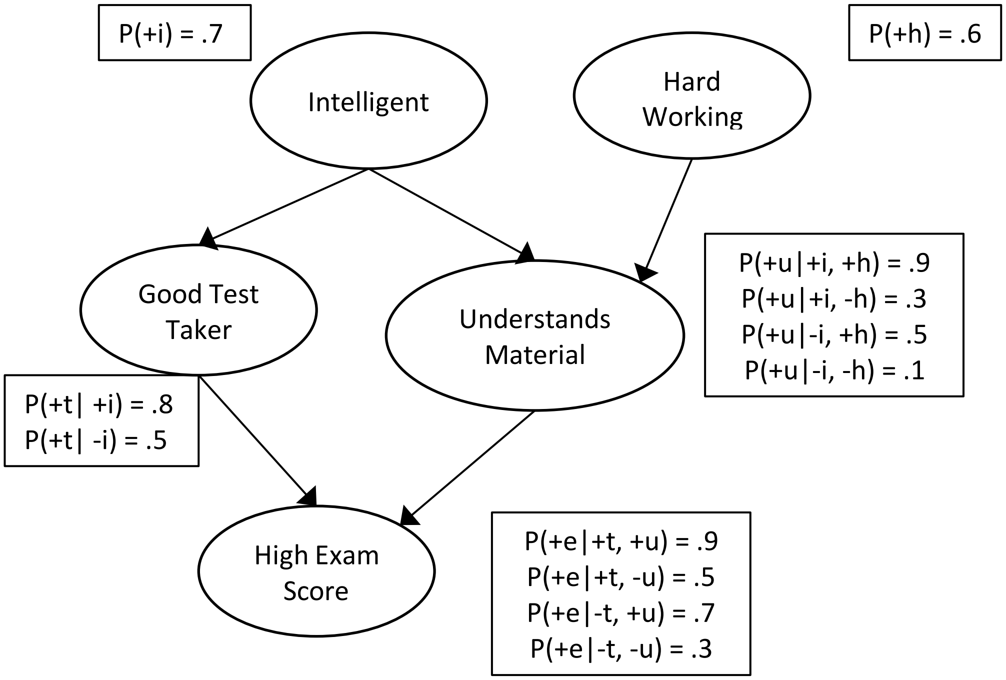
CSCE A490 Applications of PDC

Spring 2020

Due: 5 March 2020

Submission Instructions: Turn in both assignments in a zip or tar file through Blackboard. The zip or tar file should be of the form *last_name_only.zip* Each problem should be in a separate folder before zipping/tarring the two folders.

Problem 1 (25 points): Bayesian Nets



a. Using variable elimination, calculate the probability that a student who did well on the exam understands the material. P(+u | +e). Show your work.

b. Given the Bayesian network, are T and U independent? Why

c. Are I and H conditionally independent given E? Why

d. Are E and H conditionally independent given U? Why

e. Are T and H independent? Why

Problem 2 (50 points): Implementing a Regression Model

Reference: Hands-on Machine Learning with Scikit-Learn & Tensor Flow, Chapter 4

For this problem you have one file that has 21 columns of data which may be related to life expectancy in several countries over a 15-year period. Another file, LifeExpectancyMetaData.txt briefly describes the meaning of each of the columns. Using Scikit-learn or python 3.x, and the techniques discussed in lecture:

a. Divide the data into a training set and a testing set.
b. Using your training set:
    a. Determine which variables are actually affecting the life expectancy. How did you arrive at that conclusion?
c. Evaluate how well does your model predict life expectancy?
    a. Does it do better or worse depending on the country, i.e P(Life Expectancy | Country)?
    b. Which variables did you include in your model, which ones did you drop?
    c. Identify the coefficients of each of the variables in your best model.
    d. Explain what the results mean.
d. Scikit-learn offers two other types of regression, Ridge and Lasso, which help with

reducing the magnitude of the coefficients and reduces overfitting. Using *regularization*, determine if your model improves using the Ridge or Lasso regression. See which *alpha* values provide the best results. Describe your results.

e. Cross validate your model(you can use Scikit's cross validation feature). Describe how your cross validated performance compares with best model. What does the cross validation results tell you about your model.

Special Notes for problem 2: You should do all work, including data preparation using the Scikit, Numpy, Pandas, Matplotlib or Seaborn libraries. Turn in the Jupyter notebook file or python code. Ensure that your work on the data iis n the same directory as you are running the python or Jupyter code (otherwise it will not work when we test your code) . We will run your code or Jupyter Notebook file on the original data set, so don't change the data manually. Further submission instructions will be forthcoming. Please turn in your files showing your work in addition to your write up in a single zip or tar file, with your last name.