Notebook

March 5, 2020

1 Problem 1

1.1 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.

First grabbing i variable
P(i)

I	Probability
-i	0.3
+i	0.7

Next Joining With T
P(i,t)

I	T	Probability
+i	+t	0.8*0.7 = 0.56
-i	+t	0.5*0.3 = 0.16

sum = 0.56+0.16 = 0.72
Eliminiating I
P(t)

T	Probability
+t	sum = 0.72
+t	1-sum = 0.28

Inspecting h
P(h)

h	Probability
-h	0.4
+h	0.6

Joining with h and i
P(u,i,h)

u	i	h	Probability
+u	+i	+h	0.90.60.7 = 0.378
+u	+i	-h	0.30.70.4 = 0.084
+u	-i	+h	0.70.30.6 = 0.126
+u	-i	-h	0.30.30.4 = 0.036

sum = 0.378+0.084+0.126+0.036 = 0.624

u	Probability
+u	sum = 0.624
-u	1 - sum = 0.376

Joining e with u and t
P(e,t,u)

e	t	u	Probability
+e	+t	+u	0.90.720.624 = 0.4043
+e	+t	-u	0.50.720.376 = 0.1354
+e	-t	+u	0.70.280.624 = 0.1223
+e	-t	-u	0.30.280.376 = 0.0316

Eliminating t

e	u	Probability
+e	+u	P(+e,+t,+u)+P(+e,-t,+u) = (0.4043 + 0.1223) = 0.5266
+e	-u	P(+e,+t,-u)+P(+e,-t,-u) = (0.1354+0.0316) = 0.167

Normalizing e

e	Probability
+e	+u

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

T and U are not independent because they both depend on i.

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

I and H are not conditionally independent because I depends on H.

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

E and H are not conditionally independent given U because H is a parent of U and E is a parent of U.

1.5 e. Are T and H independent? Why

```
T and H are independent because they do not share a parent.
```

2 Problem 2

2.1 a. Divide the data into a training set and a testing set

I am dividing the training set and testing set by randomly selecting with replacement 80% of the data for training (with replacement) and 20% as the test data set. I am randomly selecting data inorder to avoid bias in training and test set. The test set and training set likely have overlapping entries so that may concel overfitting that may occur.

2.2 b. Using your training set, determine which variables are actually affecting the life expectancy.How did you arrive at that conclusion?

Out[3]:

	measure_name	r squared
0	Adult Mortality	0.483276
1	infant deaths	0.036888
2	Alcohol	0.159903
3	percentage expenditure	0.146038
4	Hepatitis B	0.061029
5	Measles	0.026495
6	BMI	0.312592
7	under-five deaths	0.047289
8	Polio	0.217014
9	Total expenditure	0.051020
10	Diphtheria	0.230742
11	HIV/AIDS	0.315490
12	GDP	0.214742
13	Population	0.000813
14	thinness 1-19 years	0.214244
15	thinness 5-9 years	0.212311
16	Income composition of resources	0.530858
17	Schooling	0.570264

Infant mortality, Hepatitis B, Measles, under-five deaths, Total expenditure and Population have low correlation with life expectancy. Therefore they will be removed from the analysis. I arrived at that conclusion by finding the linear correlation between the variable and life expectancy. The r^2 values are shown above. The measures with low r^2 values were dropped.

2.3 c. Evaluate how well does your model predict life expectancy

2.3.1 a. Does it do better or worse depending on the country, i.e P(Life Expectancy | Country)?

The table below shows the standard deviation of **Predicted Life Expectancy** - **Actual Life expectancy** The standard deviation varies between countries indicating that the model's performance depends on the country in question.

	Country	Standard Deviation
0	Thailand	1.329488
1	Russian Federation	0.312178
2	France	0.965511
3	Georgia	0.474709
4	Philippines	0.081492
..	...	...
116	El Salvador	0.000000
117	Azerbaijan	0.000000
118	Botswana	1.806099
119	Cameroon	0.022660
120	Sweden	0.000000

[121 rows x 2 columns]

2.3.2 b. Which variables did you include in your model, which ones did you drop?

I dropped Infant mortality, Hepatitis B, Measles, under-five deaths, Total expenditure and Population have low correlation with life expectancy. Therefore they were dropped from the dataset.

2.3.3 c. Identify the coefficients of each of the variables in your best model.

The table below shows the coefficients of each variable.

	Axis Name	Coefficient
0	Country	-1.602004e-02
1	Year	9.408674e-02
2	Status	-5.964036e-02
3	Life expectancy	3.530955e-04
4	Adult Mortality	-5.818786e-03
5	infant deaths	-9.752282e-06
6	Alcohol	3.233709e-02
7	percentage expenditure	-7.081043e-02
8	Hepatitis B	1.202629e-02
9	Measles	1.123416e-01
10	BMI	6.969993e-03
11	under-five deaths	-4.521062e-01
12	Polio	1.443341e-05
13	Total expenditure	-4.377885e-10
14	Diphtheria	-5.094472e-03

15	HIV/AIDS	-5.943678e-02
16	GDP	9.578850e+00
17	Population	9.143196e-01

2.3.4 d. explain what the results mean.

The results show the error per country of the model and the weights used inside of the linear regression.

2.3.5 d. Scikit-learn offers two other types of regression, Ridge and Lasso, whichhelp with reducing the magnitude of the coefficientsand 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

Inorder to find the optimal alpha I am iterating through alpha values ranging from 0.1 to 10. The score for each alpha is then computed. The alpha that results in the score is saved and the model produced by the best alpha value is also saved.
Max Lasso Alpha: 0.1
Max Ridge Alpha: 0.1

2.3.6 Lasso Evaluation

The tables show the standard deviation of error (as in 2a) per country for lasso regression.

	Country	Standard Deviation
0	Thailand	1.458138
1	Russian Federation	0.450272
2	France	0.935485
3	Georgia	0.650166
4	Philippines	0.147861
..	...	...
116	El Salvador	0.000000
117	Azerbaijan	0.000000
118	Botswana	1.440814
119	Cameroon	0.255762
120	Sweden	0.000000

[121 rows x 2 columns]

	Axis Name	Coefficient
0	Country	-1.746863e-02
1	Year	9.970860e-02
2	Status	-0.000000e+00
3	Life expectancy	2.716419e-04
4	Adult Mortality	-6.470066e-03
5	infant deaths	-9.542699e-06
6	Alcohol	3.970592e-02

7	percentage expenditure	-7.454338e-02
8	Hepatitis B	1.213517e-02
9	Measles	7.849139e-02
10	BMI	1.319146e-02
11	under-five deaths	-4.556120e-01
12	Polio	3.690258e-05
13	Total expenditure	-4.701398e-10
14	Diphtheria	-2.258098e-02
15	HIV/AIDS	-6.013211e-02
16	GDP	9.142401e-01
17	Population	1.197560e+00

2.3.7 Ridge Evaluation

The tables show the standard deviation of error (as in 2a) per country for ridge regression.

	Country	Standard Deviation
0	Thailand	1.330209
1	Russian Federation	0.313235
2	France	0.965681
3	Georgia	0.475434
4	Philippines	0.082036
..	...	...
116	El Salvador	0.000000
117	Azerbaijan	0.000000
118	Botswana	1.803476
119	Cameroon	0.020895
120	Sweden	0.000000

[121 rows x 2 columns]

	Axis Name	Coefficient
0	Country	-1.602857e-02
1	Year	9.414719e-02
2	Status	-5.931574e-02
3	Life expectancy	3.525364e-04
4	Adult Mortality	-5.828289e-03
5	infant deaths	-9.754444e-06
6	Alcohol	3.238079e-02
7	percentage expenditure	-7.085039e-02
8	Hepatitis B	1.202679e-02
9	Measles	1.122384e-01
10	BMI	7.017110e-03
11	under-five deaths	-4.521658e-01
12	Polio	1.458661e-05
13	Total expenditure	-4.394947e-10

```
14             Diphtheria  -5.252116e-03
15               HIV/AIDS -5.948042e-02
16                    GDP  9.513838e+00
17             Population  9.166668e-01
```

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

Below I am showing the mean and standard deviation of the scores of each model for cross validation. The lasso regression score had a lower mean score and higher standard deviation then ridge and linear models. This indicates that it is a worse model for the workload and given dataset.

Out[11]:

	Model	Mean	Standard Deviation
0	Lasso	0.809671	0.008868
1	Ridge	0.820377	0.009085
2	Linear Regression	0.820361	0.009112