

Assignment 2

March 5, 2020

Nicholas Alexeev

1 Problem 1

- 1.1 a. Using variable elimination, calculate the probability that a student who did well on the exam understands the material. $P(+u \mid +e)$. 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	$P(+i \mid +t) * P(+t) = 0.8 * 0.7 = 0.56$
-i	+t	$P(-i \mid +t) * P(+t) = 0.5 * 0.3 = 0.16$

$$\text{sum} = 0.56 + 0.16 = 0.72$$

Eliminating I

$P(t)$

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

Inspecting h

$P(h)$

h	Probability
-h	0.4

h	Probability
+h	0.6

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

u	i	h	Probability
+u	+i	+h	$0.9 \cdot 0.6 \cdot 0.7 = 0.378$
+u	+i	-h	$0.3 \cdot 0.7 \cdot 0.4 = 0.084$
+u	-i	+h	$0.7 \cdot 0.3 \cdot 0.6 = 0.126$
+u	-i	-h	$0.3 \cdot 0.3 \cdot 0.4 = 0.036$

$$\text{sum} = 0.378 + 0.084 + 0.126 + 0.036 = 0.624$$

u	Probability
+u	sum = 0.624
-u	$1 - \text{sum} = 0.376$

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

e	t	u	Probability
+e	+t	+u	$0.9 \cdot 0.72 \cdot 0.624 = 0.4043$
+e	+t	-u	$0.5 \cdot 0.72 \cdot 0.376 = 0.1354$
+e	-t	+u	$0.7 \cdot 0.28 \cdot 0.624 = 0.1223$
+e	-t	-u	$0.3 \cdot 0.28 \cdot 0.376 = 0.0316$

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	$P(+e,+u)/P(+u) = 0.5266/0.624 = 0.844$

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 in order to avoid bias in training and test set. The test set and training set likely have overlapping entries so that may cancel 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	0.494094
1	infant deaths	0.035939
2	Alcohol	0.169684
3	percentage expenditure	0.152284
4	Hepatitis B	0.055635
5	Measles	0.020728
6	BMI	0.339650
7	under-five deaths	0.046399
8	Polio	0.210744
9	Total expenditure	0.054937
10	Diphtheria	0.228632
11	HIV/AIDS	0.309518
12	GDP	0.215826
13	Population	0.000113
14	thinness 1-19 years	0.222816
15	thinness 5-9 years	0.218850
16	Income composition of resources	0.528687
17	Schooling	0.570626

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 $P(\text{Life Expectancy} \mid \text{Country})$?

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	Papua New Guinea	1.185477
1	Georgia	1.254267
2	Fiji	1.906205
3	Senegal	1.069974
4	Jordan	1.736303
..
120	Niger	0.000000
121	Portugal	0.000000
122	Latvia	0.000000
123	Colombia	0.000000
124	Belgium	0.000000

[125 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	-1.553457e-02
1	Year	9.766789e-02
2	Status	-4.693548e-02
3	Life expectancy	3.870736e-04
4	Adult Mortality	-8.171959e-03
5	infant deaths	-3.208095e-06
6	Alcohol	3.717799e-02
7	percentage expenditure	-7.325692e-02
8	Hepatitis B	8.516647e-03
9	Measles	1.091697e-01
10	BMI	1.576249e-02
11	under-five deaths	-4.638410e-01
12	Polio	1.270497e-05
13	Total expenditure	-1.290545e-09
14	Diphtheria	7.254550e-02

```

15             HIV/AIDS -1.296751e-01
16             GDP      9.644589e+00
17             Population 8.839114e-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, 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

In order 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	Papua New Guinea	1.639707
1	Georgia	1.401410
2	Fiji	2.168880
3	Senegal	1.387507
4	Jordan	2.052006
..
120	Niger	0.000000
121	Portugal	0.000000
122	Latvia	0.000000
123	Colombia	0.000000
124	Belgium	0.000000

[125 rows x 2 columns]

	Axis Name	Coefficient
0	Country	-1.690781e-02
1	Year	1.012318e-01
2	Status	0.000000e+00
3	Life expectancy	3.132767e-04
4	Adult Mortality	-8.734212e-03
5	infant deaths	-2.142607e-06
6	Alcohol	4.508840e-02

7	percentage expenditure	-7.535524e-02
8	Hepatitis B	9.142101e-03
9	Measles	7.891185e-02
10	BMI	2.088033e-02
11	under-five deaths	-4.660090e-01
12	Polio	3.393966e-05
13	Total expenditure	-1.378691e-09
14	Diphtheria	0.000000e+00
15	HIV/AIDS	-7.888213e-02
16	GDP	1.724474e+00
17	Population	1.138699e+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	Papua New Guinea	1.188786
1	Georgia	1.255276
2	Fiji	1.907968
3	Senegal	1.070356
4	Jordan	1.738262
..
120	Niger	0.000000
121	Portugal	0.000000
122	Latvia	0.000000
123	Colombia	0.000000
124	Belgium	0.000000

[125 rows x 2 columns]

	Axis Name	Coefficient
0	Country	-1.554255e-02
1	Year	9.772697e-02
2	Status	-4.658249e-02
3	Life expectancy	3.865412e-04
4	Adult Mortality	-8.180836e-03
5	infant deaths	-3.210338e-06
6	Alcohol	3.722757e-02
7	percentage expenditure	-7.329555e-02
8	Hepatitis B	8.518385e-03
9	Measles	1.090833e-01
10	BMI	1.580024e-02
11	under-five deaths	-4.638936e-01
12	Polio	1.284291e-05
13	Total expenditure	-1.293523e-09

```

14          Diphtheria    7.241506e-02
15          HIV/AIDS    -1.296930e-01
16          GDP          9.585529e+00
17          Population    8.859988e-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  0.818246          0.012269
1      Ridge  0.828392          0.011688
2  Linear Regression  0.828375          0.011698

```