HDS5230 Final Exam - programming - Miao Cai

May 6, 2019

HDS 5230 High Performance Computing
Final Exam - Progamming Part
Miao Cai

The big goal is to use the provided dataset on health insurance charges to create a model that predicts charges as accurately as possible, based on the patient traits of age, sex, bmi, children, smoker, and region. As you generate this model, you should perform and document initial data quality checks, exploratory data analysis, and all of the models you try to fit.

General components:

- Brief data summary for EDA summarizing the model input variables
- Univariate summary of the model output (cost)
- Pick a loss function and explain why you think it is a reasonable choice.
- Implement a cross validation scheme. Explain how you did this in your report.
- For the machine learning sections, implement at least the following models:
  - A few different models using H2O (random forest, gbm, regularized regression, AutoML)
  - At least 2 different architectures of neural networks using keras and tensorflow, implementing some form of regularization
- A summary of the training error versus generalization error to ensure you didn't overfit the data.
- Estimate the generalization error in your final model. Clearly state how you chose to estimate the generalization error, as well as what final expected value is. This is based on your choice of loss function

1 Introduction

In [ ]:

2 Methods summary

```
In [35]: import os
         import sys
         import pathlib
         from tableone import TableOne
         import pandas as pd
         import numpy as np

         print(sys.version)
         print("Pandas version: {0}".format(pd.__version__))
         print("Numpy version:{0}".format(np.__version__))
```

```
3.7.1 (default, Dec 14 2018, 13:28:58)
[Clang 4.0.1 (tags/RELEASE_401/final)]
Pandas version: 0.23.4
Numpy version:1.15.4
```

```
In [7]: d = pd.read_csv("insurance.csv")
        d.head()
```

Out[7]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

```
In [40]: d.dtypes
```

```
Out[40]: age           int64
         sex          object
         bmi         float64
         children      int64
         smoker       object
         region       object
         charges     float64
         dtype: object
```

```
In [52]: col_types = d.dtypes.to_dict()
         col_types['age'] = 'float64'
         d = pd.read_csv("insurance.csv", dtype=col_types)

         overall_table = TableOne(
             d, categorical = ['children', 'smoker', 'region'],
             groupby = 'sex', label_suffix=True, pval = True)
         overall_table
```

Out[52]:

		Grouped by sex			
		isnull	female	male	p
variable	level				
n			662	676	
age, mean (SD)		0	39.5 (14.1)	38.9 (14.1)	0.4
bmi, mean (SD)		0	30.4 (6.0)	30.9 (6.1)	0.0
children, n (%)	0	0	289 (43.7)	285 (42.2)	0.9
	1		158 (23.9)	166 (24.6)	
	2		119 (18.0)	121 (17.9)	
	3		77 (11.6)	80 (11.8)	
	4		11 (1.7)	14 (2.1)	
	5		8 (1.2)	10 (1.5)	
smoker, n (%)	no	0	547 (82.6)	517 (76.5)	0.0
	yes		115 (17.4)	159 (23.5)	
region, n (%)	northeast	0	161 (24.3)	163 (24.1)	0.9
	northwest		164 (24.8)	161 (23.8)	
	southeast		175 (26.4)	189 (28.0)	
	southwest		162 (24.5)	163 (24.1)	
charges, mean (SD)		0	12569.6 (11128.7)	13956.8 (12971.0)	0.0

[1] Warning, Hartigan's Dip Test reports possible multimodal distributions for: age.
[2] Warning, Tukey test indicates far outliers in: charges.
[3] Warning, test for normality reports non-normal distributions for: age, charges.

In [ ]:

3 Results

In [ ]:

4 Conclusion

In [ ]: