Group 23

Yuhao Cao
yuc121@pitt.edu
4301151

Linlu Liu
lil131@pitt.edu
4302031

Dingming Feng
dif24@pitt.edu
4301973

Progress Report

20th March 2019

PROBLEM DESCRIPTION

We want to predict whether it will rain or not tomorrow by training a binary classification model on the observations in Australia. Observations were drawn from numerous weather stations. The focus of the current stage is the replacement of missing values. We have several features with missing values of more than 30%. Those features are considered important in predicting the outcome that we don't want to drop them. Now we are trying different methods to generate a complete dataset for model fitting.

LITERATURE SUMMARY

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski, David Stillwell, and Thore Graepel

PNAS April 9, 2013 110 (15) 5802-5805;
https://doi.org/10.1073/pnas.1218772110

The paper describes how people's personal attributes can be predicted through analyzing their traces left on Facebook (likes, comments, etc), which is valuable for the improvements of websites (sales, recommendations, etc).

We can learn several methods useful for machine learning: use SVD to reduce the dimensions of data; use AUC to examine the accuracy of prediction; use Pearson Correlation Coefficient to check the accuracy of numeric variables predictions.

Predicting consumer behavior with Web search

Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts

PNAS October 12, 2010 107 (41) 17486-17490;
https://doi.org/10.1073/pnas.1005962107

The paper indicates how search queries can provide a useful guide to make predictions for the present and near future when lacking enough data sources or urgently needing small improvements in predictive performance. It studies on the predictions of four main objects: movies' box-office, video games' sales, songs' ranks, and flu. They dig into the data and offer different factors that could cause the wide variability in the predictive power of search among the different domains.

What we can learn from the paper to fortify our machine learning skills: use baseline predictions' performance as the metric to compare other machine learning algorithm against; to account for the highly skewed distribution of data, we can use log-transformation; importing features from other datasets with the same prediction target to the current one may increase the predictive power.

Predictive modeling of U.S. healthcare spending in late life

Liran Einav, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer

Science 29 Jun 2018 360 (6396) 1462-1465;
DOI: 10.1126/science.aar5045

This paper focuses on proving the statement, that one-quarter of Medicare spending in the United States occurs in the last year of life is commonly interpreted as waste because that it was spent on those who eventually died, wrong. They come to the conclusion that although spending on the ex-post dead is very high, they find out that there are only a few individuals for whom, ex-ante, death is near-certain.

What's useful for machine learning: certain labels that depend on their individual conditions and the mechanisms of themselves can't be predicted; it is wrong to draw conclusions only based on the result; different measures of collecting the same kind of data could lead to different predictions.

Large-Scale Machine Learning with Stochastic Gradient Descent

Léon Bottou

Proceedings of COMPSTAT 22-27 Aug 2010 177-186;
https://doi.org/10.1007/978-3-7908-2604-3_16

In this paper, the author describes a way to process the large-scale data that we use in machine learning. Stochastic Gradient Descent is a drastic simplification of gradient descent. It won't compute the gradient exactly. The algorithm picks examples randomly and then estimate the gradient descent.

This paper is useful because it provides a method to us on how to deal with a large-scale of the dataset.

Predicting weather forecast uncertainty with machine learning

Sebastian Scher, and Gabriele Messori

RMetS October 2018 Part B 144 (717) 2830-2841;
https://doi-org.pitt.idm.oclc.org/10.1002/qj.3410

The authors propose a method based on deep learning with artificial convolutional neural networks that are trained on past weather forecasts. Their results show that it is possible to use machine learning in order to estimate future forecast uncertainty from past forecasts.

This paper gives us a new clue to investigate this problem by using concepts of the neural network. We will give it a try to apply methods of deep learning in our exploration.

Data-Driven Machine Learning Approach for Predicting Missing Values in Large Data Sets: A Comparison Study

Ogerta Elezaj. Sule Yildirim. Edlira Kalemi

DOIhttps://doi.org/10.1007/978-3-319-72926-8_23
Print ISBN978-3-319-72925-1
Online ISBN978-3-319-72926-8

In this paper, the authors introduce several methods to impute missing data, include Group Mode, K-means, Hot-Deck, Bayesian network, Decision tree(C4.5) and some other methods. They talk about the benefit and the limitation of each method. They also use a dataset to compare the accuracy of each method.

How this paper helps us: after reading this paper, we found that the missing pattern of our dataset is kind of the same as the dataset that the author used. So we decide to use the K-means method to impute our missing data.

Using the TensorFlow Deep Neural Network to Classify Mainland China Visitor Behaviours in Hong Kong from Check-in Data

Shanshan Han; Fu Ren; Chao Wu.

ISPRS International Journal of Geo-Information, 01/2018, Volume 7, Issue 4

In this paper, the author shows a method to use tensorflow in classifying different orient customers who are from China mainland to Hong Kong. The features include words, pictures and videos from Weibo.

How this paper helps us: tensorflow is a powerful tool in machine learning. Although our project is not unsupervised learning, this paper provides us some idea in applying tensorflow in our project.

PROGRESS

We have run a thorough exploration of the dataset. After dummying all the categorical variables, we have 118 features in total. The dataset is too big to run on our own computers, so to improve our computing ability, we utilize Google Cloud to run our code.

According to the results in Figure 1, the missing values are randomly distributed. After reading the paper: Data-Driven Machine Learning Approach for Predicting Missing Values in Large Data Sets: A Comparison Study, we found methods for replacing different kinds of missing values.

		C4.5		Bayesian		K means		Hot-Deck		Group Mode	
Missing ratio	Balance approach	Accuracy %	RMSE	Accuracy %	RMSE	Accuracy %	RMSE	Accuracy %	RMSE	Accuracy %	RMSE
25%	no sampling	93.8552	0.2264	93.4441	0.2306	95.92721	0.2017	86.88	0.3622	83.84	0.2481
	down sampling	92.2622	0.2442	92.2322	0.2469	96.8758	0.1919	86.99	0.3606	83.77	0.2451
	up sampling	90.0247	0.2813	89.9771	0.2872	96.3174	0.1965	85.45	0.3814	82.99	0.2647
10%	no sampling	93.9672	0.2259	93.4229	0.2312	96.4695	0.1878	87.25	0.3570	87.93	0.2463
0.50%	no sampling	93.8526	0.2269	93.29	0.2348	91.5057	0.2914	88.91	0.3330	93.54	0.2541

Figure 2. Accuracy of different kinds of missing value replacement methods

As shown in Figure 2, for a low percentage of nan values, group mode is the best, while k-means for a medium and high percentage. We are now still working on generating complete datasets. The workload is pretty heavy as for the k-means must use complete tuples for clustering, so we have to fill all the null parts before running that. Correlation between features is also used in replacement: some features are highly correlated with others, so we can use regression to predict them, e.g., MaxTemp and Temp3pm have 98% correlation. In this case, we replace the missing values of them with each other.

One of the problems so far is that: some correlations between features are too high for keeping both of them. We are struggling with which ones to keep and how to keep them (maybe in an interaction way or a high power version). Another problem is how to evaluate our replacement. We are thinking of using the accuracy of the models to define the performance of our replacement.

Future plan

After filling all the void, we are going to fit a couple of models on the data to find out the accuracy. When choosing parameters, we will consider using gradient descent. Then we will evaluate the performance and decide whether to generate a new dataset or not.

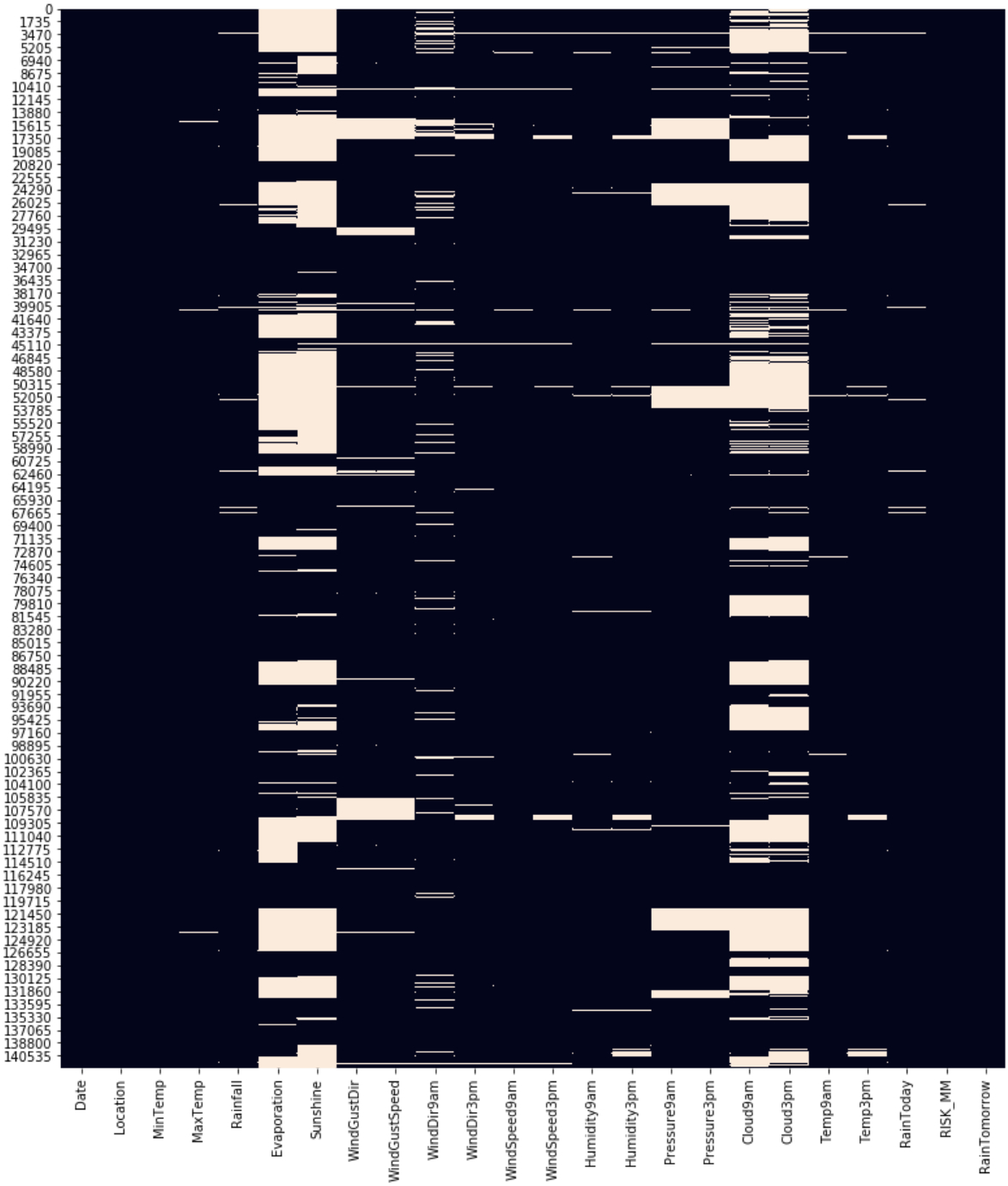

Figure 1. Missing value distribution visualization