Group 23

Yuhao Cao	Linlu Liu	Dingming Feng
yuc121@pitt.edu	lil131@pitt.edu	dif24@pitt.edu
4301151	4302031	4301973

Project Proposal

11th February 2019

PROBLEM DESCRIPTION

We want to predict whether or not it will rain tomorrow by training a binary classification model on target Rain Tomorrow.

SPECIFICATIONS

We are planning to use different machine learning methods to produce a model with the highest accuracy we can get. First, the dataset has 24 attributes, which makes it a little complex to make the prediction. We need to use p-values and confidence intervals to do feature selection, so that we can decide which of the features, single, interaction or polynomial ones, are going to be used. We will try different ways to train the model and find the one with the best result. Different kinds of validation will be used to test the models.

DATA DESCRIPTION

We are planning to use the dataset on kaggle: Rain in Australia, to make the prediction of did it rain tomorrow. The dataset contains daily weather observations from numerous Australian weather stations. The target variable RainTomorrow means: Did it rain the next day? Yes or No.

Observations were drawn from numerous weather stations. The daily observations are available from http://www.bom.gov.au/climate/data. Copyright Commonwealth of Australia 2010, Bureau of Meteorology.

Definitions adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtm

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.

RESPONSIBILITIES

We will do one model at a time, while Dingming Feng for feature selection, Linlu Liu for model fitting and Yuhao Cao for validation.