EXPERIMENT (/BROWSE?CATEGORIES=["1"])

Evaluating and Parameter Tuning a Decision Tree Model

DD Data Science Dojo . (/Home/Author?authorId=BBC43A8E9E414428484BE49D979CE6FE474178886CCFCC7922CEC62541946AE3) • November 15, 2016

♡ 7 likes

ody=Check%20out%20this%20link%3A%20https%3A%2F%2Fgallery.azure.ai%2FExperiment%2FEvaluating-

Open in Studio (//studio.azureml.net/community/unpack?packageUri=https%3a%2f%2fstorage.azureml.net%2fdirectories%2f2d824a436634447c8a9904dd1cd2f8a0%2fitems...

and-parameter-tuning-a-decision-tree-model-1&entityId=Evaluating-and-Parameter-Tuning-a-Decision-Tree-Model-1)

+ Add to Collection (/Home/SignIn)

13302 views

3079 downloads

RELATED ITEMS

Cross Validating a Classification Model (/Experiment/Cross-Validating-a-Classification-Model-1)

EXPERIMENT by Data Science Dojo . (/Home/Author?authorId=BBC43A8E9E414428484BE49D979CE6FE474178886CCFCC7922CEC62541946AE3)

Model Parameter Optimization : Sweep parameters (/Experiment/Model-Parameter-Optimization-Sweep-parameters-2)

EXPERIMENT by Microsoft (/Home/Author?authorId=72f988bf86f141af91ab2d7cd011db47)

Building a Decision Tree Classifier Model (/Experiment/Building-a-Decision-Tree-Classifier-Model-1)

EXPERIMENT by Data Science Dojo . (/Home/Author?authorId=BBC43A8E9E414428484BE49D979CE6FE474178886CCFCC7922CEC62541946AE3)

See all related items (/browse?s=Evaluating and Parameter Tuning a Decision Tree Model)

ALGORITHMS

Two-Class Decision Forest (/browse?algorithms=["Two-Class Decision Forest"])

TAGS

data cleansing (/browse/?tags=["data cleansing"])

preprocessing (/browse/?tags=["preprocessing"])

decision tree (/browse/?tags=["decision tree"])

evaluation (/browse/?tags=["evaluation"])

parameter tuning (/browse/?tags=["parameter tuning"])

Report Abuse

Summary

Learn how to optimize a decision tree learning algorithm's parameters to better fit the data.

Description

Data

This version of the Titanic dataset can be retrieved from the Kaggle (https://www.kaggle.com/c/titanic-gettingStarted/data) website, specifically their “train” data (59.76 kb). The train Titanic data ships with 891 rows, each one pertaining to an occupant of the RMS Titanic on the night of its sinking. Demo: Interact with the user interface of a model deployed as service (http://demos.datasciencedojo.com/demo/titanic/)

The dataset also has 12 columns that each record an attribute about each occupant’s circumstances and demographic. For this particular experiment we will build a classification model that can predict whether or not someone would survive the Titanic disaster given the same circumstances and demographic.

Model

First, some preprocessing. It is highly recommended that you read the detailed tutorial (http://datasciencedojo.com/dojo/building-and-deploying-a-classification-model-in-azure-ml/) to understand the rationale behind each step:

- Drop the columns that do not add immediate value for data mining or hold too many missing categorical values to be a reliable predictor attribute. The following columns were dropped using the **select columns in dataset** module:
    - PassengerID, Name, Ticket, Cabin
- Identify categorical attributes and cast them into categorical features using the **edit metadata** module. The following attributes were cast into categorical values:
    - Survived, Pclass, Sex, Embarked
- Scrub the missing values from the following columns using the **clean missing data** module:
    - All missing values associated with numeric columns were replaced with the median value of the entire column
- All missing values associated with categorical columns were replaced with the mode value of the entire column
- Randomly split and partition the data into 70% training and 30% scoring using the **split** module.

Algorithm Selection

In this gallery experiment we show that how to build a single decision tree in Azure ML, much like that of the rpart package in R programming. We will take the **two-class decision forest** as the learning algorithm and set the number of trees to one. Then we train the model using the **train model** module. We use the **score model** module to get predictions from our model on the 30% test set from the **split data** module. Evaluation metrics are given in the **evaluate model** module.

Parameter Tuning

- We begin by running the model on default parameters to get a baseline. The model starts off with 79% accuracy.
- Min-sample-per-leaf node was set to 1 by default, which would naturally make the tree over-fit and learn from the all the data points, including outliers. We increase it to about ~1% of the data points to stop the tree from prematurely classifying these outliers. Accuracy saw an improvement.
- A decision tree depth of 32 is too large for a data set with only 7 predictors. We want to create a situation where almost all features have been given a chance to participate in becoming a decision node, but not too much so that we start splitting on arbitrary numeric cut off in numeric columns. Maximum tree depth was reduced to 6, and accuracy saw an improvement.
- Number of random splits per node matters a lot more in the context of a decision forest vs a decision tree. This controls how similar the trees will look toward one another. Reducing this will have marginal impact on the performance of the model, however will dramatically increase model build times. This number needs to be not so large that a true greedy approach is applied when learning, but not so small that good features are always excluded.

Related

1. Detailed Tutorial: Building and deploying a classification model in Azure Machine Learning Studio (http://datasciencedojo.com/dojo/building-and-deploying-a-classification-model-in-azure-ml/)
2. Demo: Interact with the user interface of a model deployed as service (http://demos.datasciencedojo.com/demo/titanic/)
3. Tutorial: Creating a random forest regression model in R and using it for scoring (https://gallery.azureml.net/Details/b729c21014a34955b20fa94dc13390e5)
4. Tutorial: Obtaining feature importance using variable importance plots (https://gallery.azureml.net/Details/964dfc4151e24511aa5f78159cab0485)

1 Comment Cortana Intelligence Gallery 1 Login

Recommend Tweet Share Sort by Best

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Shubham Sonu • 2 years ago

How to get plots of a sample tree of a random forest model ?

• Reply • Share ›

Subscribe Add Disqus to your siteAdd DisqusAdd Disqus' Privacy PolicyPrivacy PolicyPrivacy

Microsoft

FAQ (http://azure.microsoft.com/en-us/documentation/articles/machine-learning-faq/) Privacy and Cookies (http://www.microsoft.com/privacystatement/en-us/core/default.aspx) Terms of Use (http://aka.ms/gallery-termsofuse) © Microsoft