Decision Forest Regression

Updated: July 8, 2015

Creates a regression model using the decision forest algorithm

Category: Machine Learning / Initialize Model / Regression (https://msdn.microsoft.com/en-us/library/azure/dn905922.aspx)

Module Overview

You can use the **Decision Forest Regression** module to create a regression model using an ensemble of decision trees.

After you have configured the model, you must train the model using a tagged dataset and the Train Model (https://msdn.microsoft.com/en-us/library/azure/dn906044.aspx) module. The trained model can then be used to make predictions. Alternatively, the untrained model can be passed to Cross-Validate Model (https://msdn.microsoft.com/en-us/library/azure/dn905852.aspx) for cross-validation against a labeled data set.

Understanding Decision Forests for Regression

Decision trees are non-parametric models that perform a sequence of simple tests for each instance, traversing a binary tree data structure until a leaf node (decision) is reached.

Decision trees have these advantages:

- They are efficient in both computation and memory usage during training and prediction.
- They can represent non-linear decision boundaries.
- They perform integrated feature selection and classification and are resilient in the presence of noisy features.

This regression model consists of an ensemble of decision trees. Each tree in a regression decision forest outputs a Gaussian distribution by way of prediction. An aggregation is performed over the ensemble of trees to find a Gaussian distribution closest to the combined distribution for all trees in the model.

How to Configure a Decision Forest Regression Model

1. Drag the **Decision Forest Regression** module into your experiment.

2. Set model .Bookmark link 'bkmk_Options' is broken in topic '{"project_id":"37f8d135-1f1d-4e57-9b7d-b084770c6bf5","entity_id":"562988b2-e740-4e3a-8131-358391bad755","entity_type":"Article","locale":"en-US"}'. Rebuilding the topic '{"project_id":"37f8d135-1f1d-4e57-9b7d-b084770c6bf5","entity_id":"562988b2-e740-4e3a-8131-358391bad755","entity_type":"Article","locale":"en-US"}' may solve the problem.

3. Connect an instance of Train Model (https://msdn.microsoft.com/en-us/library/azure/dn906044.aspx) or Sweep Parameters (https://msdn.microsoft.com/en-us/library/azure/dn905810.aspx) to train the model using a tagged dataset. *Tagged* means that the dataset contains a label column that contains the regression model’s expected output variable.

4. The trained model can then be used to make predictions. Alternatively, the untrained model can be passed to Cross-Validate Model (https://msdn.microsoft.com/en-us/library/azure/dn905852.aspx) for cross-validation against a labeled data set.

Options

You can configure a regression model using the same parameters that apply to other decision forest models:

Resampling method

You can specify which method should be used to combine the results of individual trees.

You can choose from *bagging* or *replication*.

- ***Bagging***.

  Select this option to use bagging, also called *bootstrap aggregating*. Each tree in a regression decision forest outputs a Gaussian distribution by way of prediction. The aggregation is to find a Gaussian whose first two moments match the moments of the mixture of Gaussians given by combining all Gaussians returned by individual trees.

- ***Replicate***. In replication, each tree is trained on exactly the same input data.

Number of decision trees

Specify the total number of decision trees that are created in the ensemble.

By creating more decision trees, you can potentially get better coverage, but training time will increase.

Maximum depth of the decision trees

Specify the maximum depth of each tree in the ensemble.

Increasing the depth of a tree has the effect of increasing precision at the risk of some overfitting and increased training time.

Number of random splits per node

Specify the number of splits to use when building each node of the tree.

At each split, features in each level of the tree (node) are randomly divided.

Minimum number of samples per leaf node

Type a value that defines the number of samples or case required to create a decision tree leaf node.

By increasing this value, you increase the threshold for creating new rules. For example, with the default value of 1, even a single case can cause a new rule to be created. If you increase the value to 5, the training data would have to contain at least 5 cases that meet the same conditions.

Allow unknown values for categorical features

When this option is selected, the model will create a grouping for Unknown values.

If you deselect this option, the model can accept only the values contained in the training data. In the former case, the model might be less precise on known values but provide better predictions for new (unknown) values.

Recommendations

If you have limited data or want to minimize the time spent training the model, try these settings:

Limited training set. If the training set contains a limited number of instances:

- Create the decision forest using a large number of decision trees (for example, more than 20)
- Use the **Bagging** option for resampling
- Specify a large number of random splits per node (for example, more than 1000)

Limited training time. If the training set contains a large number of instances and training time is limited:

- Create the decision forest using fewer decision trees (for example, 5-10)
- Use the **Replicate** option for resampling
- Specify a small number of random splits per node (for example, less than 100)

Examples

For examples of regression models, see these sample experiments in the Model Gallery (http://gallery.azureml.net/):

- The Compare Regression Models sample (http://go.microsoft.com/fwlink/?LinkId=525731) contrasts several different kinds of regression models.
- The sentiment analysis sample (http://go.microsoft.com/fwlink/?LinkId=525274) uses several different regression models to generate predicted ratings.

Technical Notes

Module Parameters

Name	Range	Type	Default	Description
Resampling method	any	ResamplingMethod	Bagging	Choose a resampling method
Number of decision trees	>=1	Integer	8	Specify the number of decision trees to create in the ensemble
Maximum depth of the decision trees	>=1	Integer	32	Specify the maximum depth of any decision tree that can be created in the ensemble
Number of random splits per node	>=1	Integer	128	Specify the number of splits generated per node, from which the optimal split is selected
Minimum number of samples per leaf node	>=1	Integer	1	Specify the minimum number of training samples required to generate a leaf node
Allow unknown values for categorical features	any	Boolean	true	Indicate whether unknown values of existing categorical features can be mapped to a new, additional feature

Outputs

Name	Type	Description
Untrained model	ILearner interface (https://msdn.microsoft.com/en-us/library/azure/dn905938.aspx)	An untrained regression model

See Also

Machine Learning / Initialize Model / Regression (https://msdn.microsoft.com/en-us/library/azure/dn905922.aspx)
A-Z List of Machine Learning Studio Modules (https://msdn.microsoft.com/en-us/library/azure/dn906033.aspx)