Fast Forest Quantile Regression

Updated: July 25, 2015

Creates a quantile regression model

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

Module Overview

You can use the **Fast Forest Quantile Regression** module to create a regression model that can predict values for a specified number of quantiles. Quantile regression is useful if you want to understand more about the distribution of the predicted value, rather than get a single mean prediction value.

This method has many applications, including:

- Predicting prices
- Estimating student performance or applying growth charts to assess child development
- Discovering predictive relationships in cases where there is only a weak relationship between variables

This regression method is a supervised learning method, and therefore requires a *tagged dataset*, which includes a label column. The label column must contain numerical values.

You can train the model by providing the model and the tagged dataset as an input to 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). The trained model can then be used to predict values for the new input examples.

Understanding Quantile Regression

There are many different types of regression. In the most basic sense, regression means fitting a model to a target expressed as a numeric vector. However, statisticians have been developing increasingly advanced methods for regression.

The simplest definition of quantile is a value that divides a set of data into equal-sized groups, and the quantile values mark the boundaries between groups. Statistically, quantiles are values taken at regular intervals from the inverse of the cumulative distribution function (CDF) of a random variable.

Quantile regression is used if you want to understand more about the distribution of the predicted value. Whereas linear regression models attempt to predict the value of a numeric variable using instead of a single estimate, the *mean*, sometimes you need to predict the range or entire distribution of the target variable. To do this, you can use techniques such as Bayesian regression and quantile regression. Tree-based quantile regression models, such as the one used in this module, have the additional advantage that they can be used to predict non-parametric distributions.

For additional implementation details and resources, see the Technical Notes section.

How to Configure a Quantile Regression Model

1. Specify how you want the model to be trained, by setting the **Create trainer mode** option.

   When you create this model, you have two options for training the model: using a single set of parameters with Train Model (https://msdn.microsoft.com/en-us/library/azure/dn906044.aspx), or choosing a range of parameters and training the model with a parameter sweep.

   - **Single Parameter mode**

     If you know how you want to configure the regression model, you can provide a specific set of values as arguments. You might have learned these values by experimentation or received them as guidance.

   - **Sweep mode**

     If you are not sure of the best parameters, you can find the optimal parameters by using Sweep Parameters (https://msdn.microsoft.com/en-us/library/azure/dn905810.aspx) to train the model.

2. If you choose the **Single Parameter** option, type other values in the **Properties** pane that control the behavior of the regression model, such as leaves per tree and learning rate. See the section for details.Bookmark link 'bkmk_Options' is broken in topic '{"project_id":"37f8d135-1f1d-4e57-9b7d-b084770c6bf5","entity_id":"b9064dc3-2d69-4e06-b307-6cebf324686a","entity_type":"Article","locale":"en-US"}'. Rebuilding the topic '{"project_id":"37f8d135-1f1d-4e57-9b7d-b084770c6bf5","entity_id":"b9064dc3-2d69-4e06-b307-6cebf324686a","entity_type":"Article","locale":"en-US"}' may solve the problem.

   Then, connect the model to the Train Model (https://msdn.microsoft.com/en-us/library/azure/dn906044.aspx) module, along with a labeled training dataset, and run the experiment

3. If you choose the **Parameter Range** option, use the **Range Builder** to set an upper and lower range for each numeric parameter.

   Then, connect the model to the Sweep Parameters (https://msdn.microsoft.com/en-us/library/azure/dn905810.aspx) module, along with a labeled training dataset, and run the experiment.

   Sweep Parameters (https://msdn.microsoft.com/en-us/library/azure/dn905810.aspx) will iterate over all possible combinations of the settings you provided and determine the combination of settings that produces the optimal results. You can use the model trained using those parameters, or you can make a note of the parameter settings to use when configuring a learner.

Options

To create a regression model that can predict quantiles, specify the following parameters:

Create trainer mode

Choose the method used for configuring and training the model:

- ***Single Parameter***

  Select this option to configure and train the model with a single set of parameter values that you supply.

  If you choose this option, you should train the model by using the Train Model (https://msdn.microsoft.com/en-us/library/azure/dn906044.aspx) module.

- ***Parameter Range***

  Select this option to use the Range Builder and specify a range of possible values. You then train the model using a parameter sweep, to find the optimum configuration.

> **Warning**
>
> - If you pass a parameter range to Train Model (https://msdn.microsoft.com/en-us/library/azure/dn906044.aspx), it will use only the first value in the parameter range list.
> - If you pass a single set of parameter values to the Sweep Parameters (https://msdn.microsoft.com/en-us/library/azure/dn905810.aspx) module, when it expects a range of settings for each parameter, it ignores the values and using the default values for the learner.
> - If you select the **Parameter Range** option and enter a single value for any parameter, that single value you specified will be used throughout the sweep, even if other parameters change across a range of values.

Random number seed

Specify a value to seed the random number generator used by the model. Use of a seed is useful in order to reproduce duplicate runs.

The default is 0, meaning a random seed is chosen.

Allow unknown categorical levels

Select this option to create an additional level for each categorical column that is provided as training input to the model.

Single Parameter options

When **Create trainer mode** is set to **Single Parameter**, the following parameters are available:

Number of trees

Specify the total number of trees that you want to construct.

If you create more trees, it generally leads to greater accuracy, but at the cost of longer training time.

Number of leaves

Specify the total number of leaves, or terminal nodes, to allow in each tree.

Minimum number of training instances required to form a leaf

Specify the number of sample cases needed to create a leaf node.

Bagging fraction

Specify the fraction of samples to use when building each tree. Cases are always chosen randomly, with replacement.

Feature fraction

Specify the fraction of total features to use when building any particular tree. Features are always chosen randomly.

Split fraction

Specify the fraction of features to use in each split of the tree. The features used are always chosen randomly.

Quantile sample count
Specify the number of predicted labels to be sampled in each leaf node

Quantiles to be estimated
Provide a comma-separated list of the quantiles on which you want the model to train and create predictions.

For example, if you want to build a model that estimates for quartiles, you would type **0.25, 0.5, 0.75**.

Parameter Range options

When **Create trainer mode** is set to **Parameter Range**, the following parameters are available:

Maximum number of leaves per tree
Specify the total number of leaves, or terminal nodes, to allow in each tree.

Number of trees constructed
Specify the total number of trees that you want to construct.

If you create more trees, it generally leads to greater accuracy, but at the cost of longer training time.

Minimum number of sample per leaf node
Specify the number of sample cases needed to create a leaf node.

Range for bagging fraction
Specify the fraction of samples to use when building each group of quantiles. Samples are chosen randomly, with replacement.

Range for feature fraction
Specify the fraction of total features to use when building each group of quantiles. Features are chosen randomly.

Range for split fraction
Specify the fraction of features to use in each group of quantiles. The features used are chosen randomly.

Sample count used to estimate the quantiles
Specify the number of samples to evaluate when estimating the quantiles.

Required quantile values
Provide a comma-separated list of the quantiles on which you want the model to train and create predictions.

For example, if you want to build a model that estimates for quartiles, you would type **0.25, 0.5, 0.75**.

Examples

For examples of how to use this module, see these sample experiments in the Model Gallery (http://gallery.azureml.net/):

- The Quantile Regression (http://go.microsoft.com/fwlink/?LinkId=525769) sample demonstrates how to build and interpret a quantile regression model, using the auto price dataset.

Technical Notes

The **Fast Forest Quantile Regression** module in Azure Machine Learning is an implementation of random forest quantile regression using decision trees. Random forests can be helpful to avoid overfitting that can occur with decision trees. A decision tree is a binary tree-like flow chart, where at every interior node, one decides which of the two child nodes to continue to, based on the value of one of the features of the input.

In each leaf node, a value is returned. In the interior nodes, the decision is based on the test $x \leq v$, where x is the value of the feature in the input sample and v is one of the possible values of this feature. The functions that can be produced by a regression tree are all the piece-wise constant functions.

In a random forest, an ensemble of trees is created by using bagging to select a subset of random samples and features of the training data, and then fit a decision tree to each subset of data. Unlike the random forest algorithm, which averages out the output of the all the trees, **Fast Forest Quantile Regression** keeps all the predicted labels in trees specified by the parameter **Quantile sample count** and outputs the distribution, so that the user can view the quantile values for the given instance.

For more information see these books and articles:

- Quantile Regression Forests. Nicolai Meinshausen

  http://jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf (http://jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf)

- Random forests. Leo Breiman.

  http://rd.springer.com/article/10.1023%2FA%3A1010933404324 (http://rd.springer.com/article/10.1023/A:1010933404324)

Module Parameters

Name	Type	Range	Optional	Description	Default
Create trainer mode	CreateLearnerMode	List:Single Parameter|Parameter Range	Required	Single Parameter	Create advanced learner options
Number of Trees	Integer		mode:Single Parameter	100	Specify the number of trees to be constructed
Number of Leaves	Integer		mode:Single Parameter	20	Specify the maximum number of leaves per tree. The default number is 20

Minimum number of training instances required to form a leaf	Integer		mode:Single Parameter	10	Indicates the minimum number of training instances required to form a leaf
Bagging fraction	Float		mode:Single Parameter	0.7	Specifies the fraction of training data to use for each tree
Feature fraction	Float		mode:Single Parameter	0.7	Specifies the fraction of features (chosen randomly) to use for each tree
Split fraction	Float		mode:Single Parameter	0.7	Specifies the fraction of features (chosen randomly) to use for each split
Quantile sample count	Integer		mode:Single Parameter	100	Specifies number of instances used in each node to estimate quantiles
Quantiles to be estimated	String		mode:Single Parameter	"0.25;0.5;0.75"	Specifies the quantile to be estimated
Random number seed	Integer		Optional		Provide a seed for the random number generator used by the model. Leave blank for default.
Allow unknown categorical	Boolean		Required	true	If true, create an additional

levels					level for each categorical column. Levels in the test dataset not available in the training dataset are mapped to this additional level.
Maximum number of leaves per tree	ParameterRangeSettings	[16;128]	mode:ParameterRange	16; 32; 64	Specify range for the maximum number of leaves allowed per tree
Number of trees constructed	ParameterRangeSettings	[1;256]	mode:ParameterRange	16; 32; 64	Specify the range for the maximum number of trees that can be created during training
Minimum number of samples per leaf node	ParameterRangeSettings	[1;10]	mode:ParameterRange	1; 5; 10	Specify the range for the minimum number of cases required to form a leaf
Range for bagging fraction	ParameterRangeSettings	[0.25;1.0]	mode:ParameterRange	0.25; 0.5; 0.75	Specifies the range for fraction of training data to use for each tree
Range for feature fraction	ParameterRangeSettings	[0.25;1.0]	mode:ParameterRange	0.25; 0.5; 0.75	Specifies the range for fraction of features (chosen randomly) to use for each tree

Range for split fraction	ParameterRangeSettings	[0.25;1.0]	mode:ParameterRange	0.25; 0.5; 0.75	Specifies the range for fraction of features (chosen randomly) to use for each split
Sample count used to estimate the quantiles	Integer		mode:ParameterRange	100	Sample count used to estimate the quantiles
Required quantile values	String		mode:ParameterRange	"0.25;0.5;0.75"	Required quantile value used during parameter sweep

Outputs

Name	Type	Description
Untrained model	ILearner interface (https://msdn.microsoft.com/en-us/library/azure/dn905938.aspx)	An untrained quantile regression model that can be connected to the Train Generic Model or Cross Validate Model modules.

See Also

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