**Azure Machine Learning Studio**

<http://studio.azureml.net/>

<https://docs.microsoft.com/en-us/azure/machine-learning/>

<https://docs.microsoft.com/zh-cn/azure/machine-learning/>

Machine Learning Studio: Algorithm and Module Help

From <<https://msdn.microsoft.com/en-us/library/dn905974.aspx>>

Text Analytics

From <<https://msdn.microsoft.com/library/en-us/Dn905886.aspx>>

**Cortana Intelligence Gallery 试验库**

<https://gallery.cortanaintelligence.com/browse>

帮大家找了一些和文本分析相关的Lab

* Predict Book Reviews
  + <https://gallery.cortanaintelligence.com/Experiment/Predict-Book-Reviews-1>
  + <https://gallery.cortanaintelligence.com/Experiment/Predict-Book-Reviews-Predictive-Experiment-1>
* The [News Categorization](http://go.microsoft.com/fwlink/?LinkId=525167) sample uses feature hashing to classify articles into a predefined list of categories.
* The [Find similar companies](http://go.microsoft.com/fwlink/?LinkId=525164) sample uses the text of Wikipedia articles to categorize companies.
* In the five-part [Text Classification](http://go.microsoft.com/fwlink/?LinkId=525957) sample, text from Twitter messages is used to perform sentiment analysis.
* [Text Classification: Step 1 of 5, data preprocessing](http://gallery.azureml.net/Details/f43e79f47d8a4219bf8613d271ea2c45)
* [Text Classification: Step 2 of 5, text preprocessing](http://gallery.azureml.net/Details/464eb78e197d4440a332a129d8d523eb)
* [Text Classification: Step 3A of 5, n-grams TF feature extraction](http://gallery.azureml.net/Details/cf65bf129fee4190b6f48a53e599a755)
* [Text Classification: Step 3B of 5, unigrams TF-IDF feature extraction](http://gallery.azureml.net/Details/7a5a38a13fa34e2b847b519629da1a59)
* [Text Classification: Step 4 of 5, train and evaluate models](http://gallery.azureml.net/Details/28437611ee1a42df8efb8f4b12a7aa88)
* [Text Classification: Step 5A of 5, deploy web service with n-grams TF model](http://gallery.azureml.net/Details/ecaa60e30c19443e9313f53155bc4367)
* [Text Classification: Step 5B of 5, deploy web service with unigrams TF-IDF model](http://gallery.azureml.net/Details/e98cccbbec1c4739b3691848d01b1b56)

1.  数据如何导入及数据的类型。

对于所要用于训练及验证模型的数据，需要导入到AML的Studio中。目前支持的数据导入方式如下：

•本地文件上传

•Azure BLOB storage, table

•Azure SQL database

•Hadoop using HiveQL

•A web URL using HTTP

•A data feed provider（OData）

支持的数据类型如下：

• txt文本文件；

• CSV 文件，包括.csv和.nh.csv；

• TSV文件，包括.tsv和.nh.tsv；

• Hadoop Hive table

• SQL database table

• OData values

• SVMLight data (.svmlight) (具体描述见链接：http://svmlight.joachims.org/ )

• Attribute Relation File Format (ARFF) data (.arff) (具体描述见链接：http://weka.wikispaces.com/ARFF )

• Zip file (.zip)

• R object or workspace file (.RData)

 2. 内置的算法

概括的说，微软Azure机器学习内置了基于监督学习和非监督学习的分类、回归、聚类等的20多种算法，详细的算法描述详见链接：https://msdn.microsoft.com/en-us/library/azure/dn905812.aspx 。我也会在后面的博文中陆续向大家介绍。除了算法之外， AML还集成了400+多个R语言的程序包。

对于选择什么样的算法，无论是对于初学者还是有经验的数据科学家，其实都是很让人费脑筋的事情。微软也提供了很多资料帮助大家判断应该选择哪些算法。如下是几个非常有用的链接（英文）：

•Microsoft Azure Machine Learning Algorithm Cheat Sheet - https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet/

•Choosing a Learning Algorithm in Azure Machine Learning - http://blogs.technet.com/b/machinelearning/archive/2015/05/20/choosing-a-learning-algorithm-in-azure-ml.aspx

•Choosing a Machine Learning Classifier - http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/

•Choosing the right estimator - http://scikit-learn.org/stable/tutorial/machine\_learning\_map/

        3. 内置的应用模块

为了方便微软Azure机器学习让更多的人很容易的上手和使用，AML原生内置了很多业务场景的原始数据和机器学习模块及API。用户可直接使用它们，或者做少量的修改为自己所用。主要的业务场景包括但不限于如下（还在持续增加中）。初学者可以先从这些已有的模块理解和掌握机器学习的使用。可以从登录首页上方菜单的“Gallery”中找到这些已经构建好的模型。

• 文本分析；

• 客户流失预测；

• 推荐系统；

• 预测性维护；

• 欺诈监测

Azure machine learning 机器学习算法的选择

异常数据点检测：

One-class SVM

PCA-based anomaly detection

聚类：

K-means

预测值：

数据分类是排序的：Ordinal regression

预测事件次数：Poisson regression

预测分布：Fast forest quantile regression

快速训练，线性模型：Linear regression

线性模型，小数据集：Bayesian linear regression

精确，但训练时间长：Neural network regression

精确，快速训练：Decision forest regression

精确，快速学习，占用多内存：Boosted decision tree regression

预测两个分类（二元分类）：

大于100个特征，线性模型：Two-class SVM

快速训练，线性模型：Two-class averaged perceptron

快速训练，线性模型：Two-class logistic regression

快速训练，线性模型：Two-class Bayes point machine

精确，快速训练：Two-class decision forest

精确，快速训练，占用多内存：Two-class boosted decision tree

精确，占用内存少：Two-class decision jungle

大于100个特征：Two-class locally deep SVM

精确，训练时间长：Two-class neural network

预测多个分类（多元分类）：

快速训练，线性模型：Multiclass logistic regression

精确，训练时间长：Multiclass neural network

精确，快速训练：Multiclass decision forest

精确，占用内存少：Multiclass decision jungle

依赖于二元分类器：One-v-all multiclass

在实际应用中，先分清问题类别（值预测，分类预测，聚类），然后一一尝试该问题分类中的每种算法，最后再做决定。

参考链接 : <https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet/>