Classification and Clustering

Eng Teong Cheah
MVP Visual Studio &
Development Technologies

Agenda

Using classification algorithms
Clustering techniques
Selecting algorithms

Using classification algorithms

Classification

Classification is a machine learning method that uses data to determine the category, type, or class of an item or row of data.

Classification

For example, you can use classification to:

- Classify email filters as spam, junk, or good.
- Determine whether a patient's lab sample is cancerous.
- Categorize customers by their propensity to respond to a sales campaign.
- Identify sentiment as positive or negative.

Create a classification model

To create a classification model, or classifier, first, select an appropriate algorithm. Consider these factors:

- How many classes or different outcomes do you want to predict?
- What is the distribution of the data?
- How much time can you allow for training?

Create a classification model

Machine Learning Studio provides multiple classification algorithms. When you use One-Vs-All algorithm, you can even apply a binary classifier to a multiclass problem.

Create a classification model

After you choose an algorithm and set the parameters by using the modules in this section, train the model on labeled data. Classification is a supervised machine learning method. It always requires labeled training data.

Create a classification model

When training is finished, you can evaluate and tune the model. When you're satisfied with the model, use the trained model for scoring with new data.

Clustering techniques

Clustering

Clustering, in machine learning, is a method of grouping data points into similar clusters. It is also called segmentation.

Clustering

Over the years, many clustering algorithms have been developed. Almost all clustering algorithms use the features of individual items to find similar items.

For example, you might apply clustering to find similar people by demographics. You might use clustering with text analysis to group sentences with similar topics or sentiment.

Clustering

Clustering is called a non-supervised learning technique because it can be used in unlabeled data.

Indeed, clustering is a useful first step for discovering new patterns, and requires little prior knowledge about how the data might be structured or how items are related. Clustering is often sued for exploration of data prior to analysis with other more predictive algorithms.

Clustering

Clustering is called a non-supervised learning technique because it can be used in unlabeled data.

Indeed, clustering is a useful first step for discovering new patterns, and requires little prior knowledge about how the data might be structured or how items are related. Clustering is often sued for exploration of data prior to analysis with other more predictive algorithms.

How to create a clustering model

In Machine Learning Studio, you can use clustering with either labeled or unlabeled data.

- In unlabeled data, the clustering algorithm determined which data points are closest together, and created clusters around a central point, or centroid. You can use the cluster ID as a temporary label for the group of data.

How to create a clustering model

- If the data has labels, you can use the label to drive the number of clusters, or use the label as just another feature.

After you have configured the clustering algorithm, you train it on data by using either the Train Clustering Model or Sweep Clustering modules.

How to create a clustering model

When the model is trained, use it to predict cluster membership for new data points.

For example, if you have used clustering to group customers by purchasing behavior, you can use the model to predict the purchasing behavior of new customers.

Selecting algorithms

How to choose algorithms

The answer to the question "**What machine learning algorithm should I use**?" is always "**It depends**"

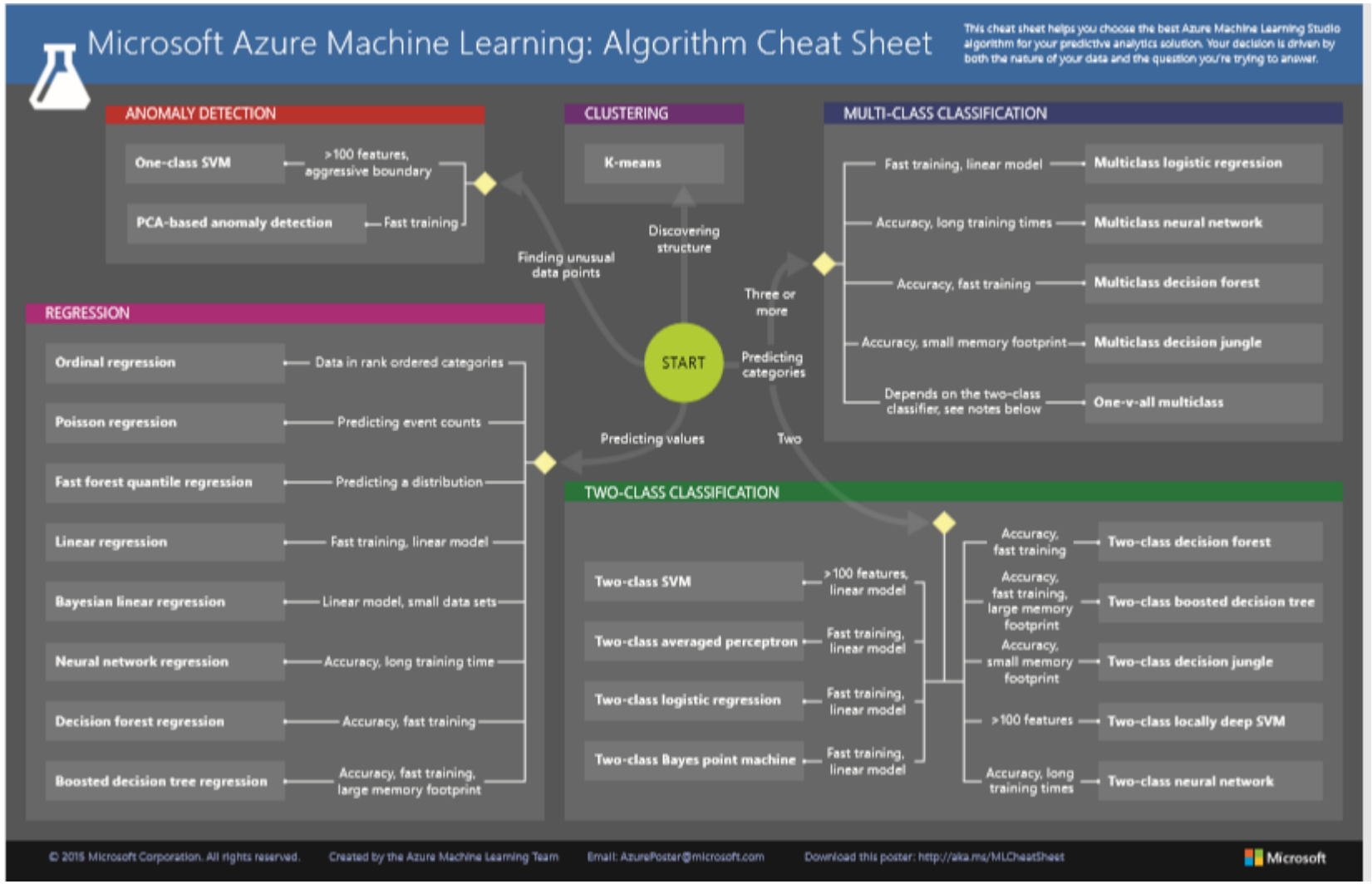
It depends on the size, quality, and nature of the data.

It depends on what you want to do with the answer.

It depends on how the math of the algorithm was translated in instructions for the computer you are using.

It depends on how much time you have.

The Algorithm Cheat Sheet



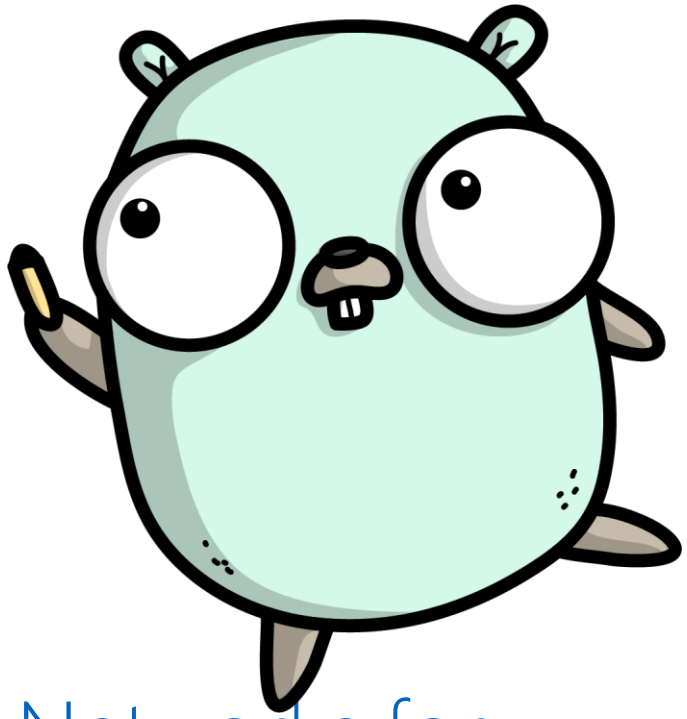
Download this poster:

http://aka.ms/MLCheatSheet

Demo

Using clustering in Azure ML Studio

Resources



TutorialsPoint

Microsoft Docs

Lecture Collection | Convolutional Neural Networks for Visual Recognition(Spring 2017)

Python Numpy Tutorial

Image Credits: @ashleymcnamara

Thank you

Eng Teong Cheah
Microsoft MVP Visual Studio & Development Technologies
Twitter: @walkercet
Github: https://github.com/ceteongvanness
Blog: https://ceteongvanness.wordpress.com/
Youtube: http://bit.ly/etyoutubechannel