Diving Deeper into Azure Machine Learning

Jerry Kurata
CONSULTANT

@jerrykur   www.insteptech.com

Module Overview

Adding Data into Azure ML

Exploring and Pre-Processing Data

Selecting the Correct Algorithm

Incorporating R and Python Code

New Experiment – Loan Prediction

50-80% of a ML project
is spent
getting, cleaning, and
organizing data

Getting Data

Local files

- Static CSV, text, etc files uploaded to Azure

Other sources

- Web site, SQL, Hadoop, Document DB, BLOBs
- Can be dynamic:
  - Query last week’s data
  - Pull latest data from a website
  - Produced by ETL processes
  - Good for automating retraining on latest data

Demo

Adding local file as a dataset

German Credit data

- From UCI repository
- 2 files
  - Data – german.data.csv
  - Documentation – german.doc

Data Exploration

Get data

Review data

Plan changes

Learn relationships between features

Data Pre-Processing (Part 1)

Make data types useful

Remove extraneous data

Tidy Data

Tidy datasets are easy to manipulate, model and visualize, and have a specific structure:

each variable(feature) is a column,

each observation is a row,

each type of observational unit is a table.

Hadley Wickham

Data Pre-Processing (Part 2)

Need to handle 5 times as “bad” bias

Cannot alter algorithm’s code

Bias data instead

Make 5 copies of each “bad” credit risk

Doing in R (or Python) is fastest

Before or after Split Data module?

Change after splitting data

Including R and Python

Leverages language strengths

Incorporates previously written code

Selecting an Algorithm

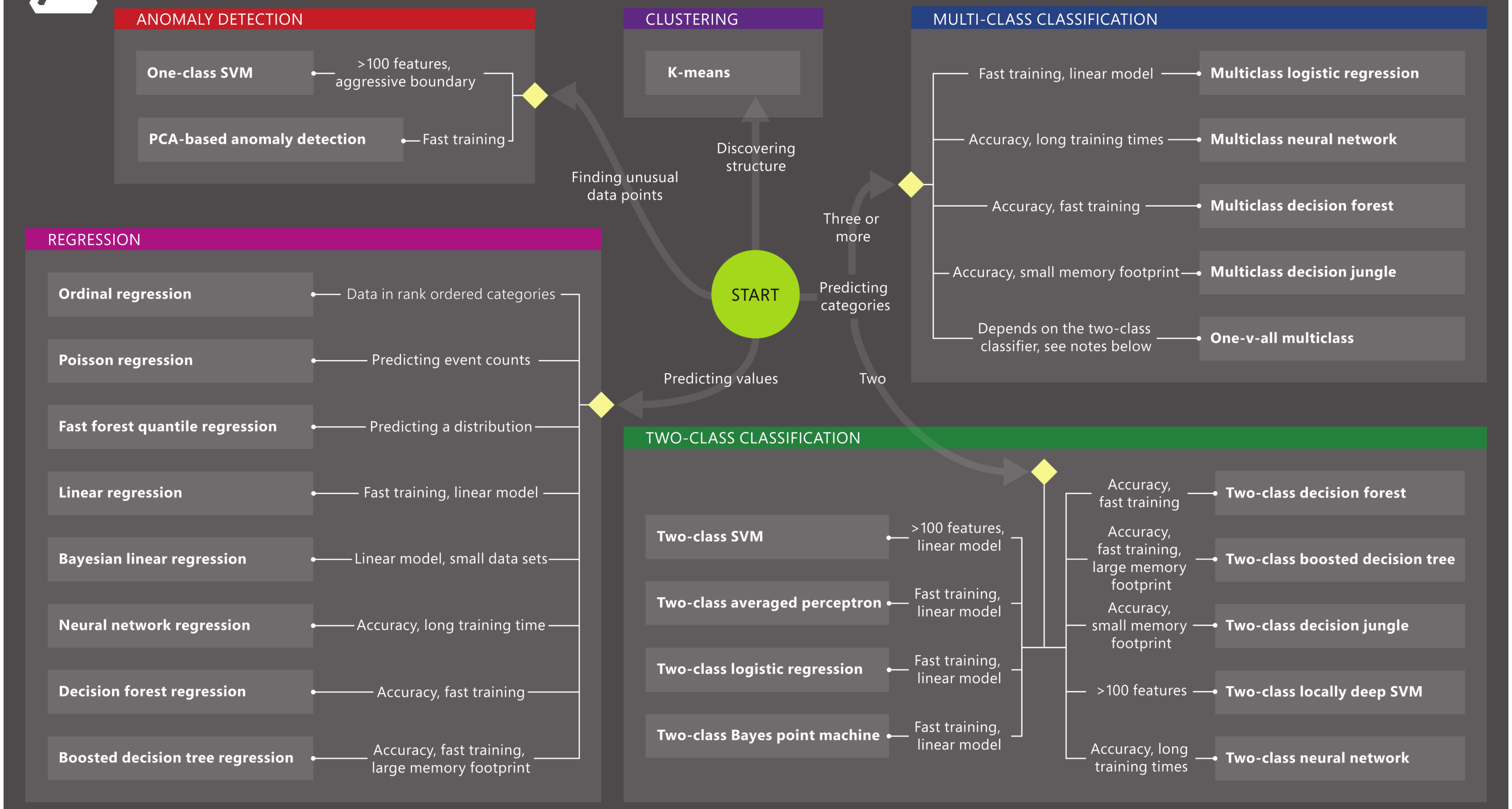
Use Microsoft ML algorithm cheat sheet

https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet/

Azure ML algorithms are similar to standard algorithms

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.



START

Finding unusual data points

Discovering structure

Three or more

Predicting categories

Predicting values

Two

ANOMALY DETECTION

One-class SVM	>100 features, aggressive boundary
PCA-based anomaly detection	Fast training

CLUSTERING

K-means

MULTI-CLASS CLASSIFICATION

Fast training, linear model	Multiclass logistic regression
Accuracy, long training times	Multiclass neural network
Accuracy, fast training	Multiclass decision forest
Accuracy, small memory footprint	Multiclass decision jungle
Depends on the two-class classifier, see notes below	One-v-all multiclass

REGRESSION

Ordinal regression	Data in rank ordered categories
Poisson regression	Predicting event counts
Fast forest quantile regression	Predicting a distribution
Linear regression	Fast training, linear model
Bayesian linear regression	Linear model, small data sets
Neural network regression	Accuracy, long training time
Decision forest regression	Accuracy, fast training
Boosted decision tree regression	Accuracy, fast training, large memory footprint

TWO-CLASS CLASSIFICATION

Two-class SVM	>100 features, linear model
Two-class averaged perceptron	Fast training, linear model
Two-class logistic regression	Fast training, linear model
Two-class Bayes point machine	Fast training, linear model

Accuracy, fast training	Two-class decision forest
Accuracy, fast training, large memory footprint	Two-class boosted decision tree
Accuracy, small memory footprint	Two-class decision jungle
>100 features	Two-class locally deep SVM
Accuracy, long training times	Two-class neural network

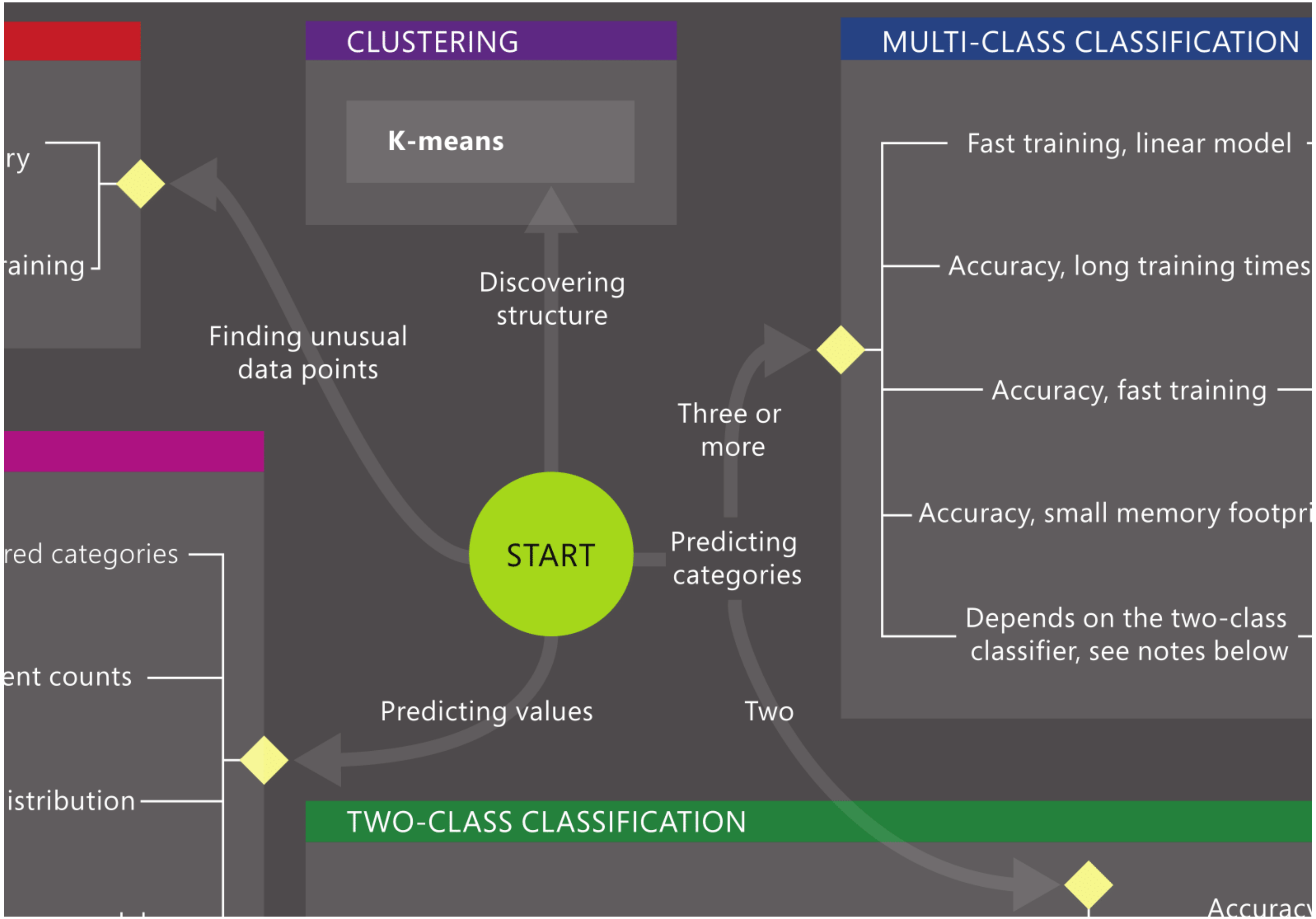
Created by the Azure Machine Learning Team

Email: AzurePoster@microsoft.com

Download this poster: http://aka.ms/MLCheatSheet

Microsoft

Selecting the Algorithm



CLUSTERING

K-means

MULTI-CLASS CLASSIFICATION

Fast training, linear model

Accuracy, long training times

Accuracy, fast training

Accuracy, small memory footpri

Depends on the two-class classifier, see notes below

ry

aining

Finding unusual data points

Discovering structure

Three or more

START

Predicting categories

red categories

ent counts

Predicting values

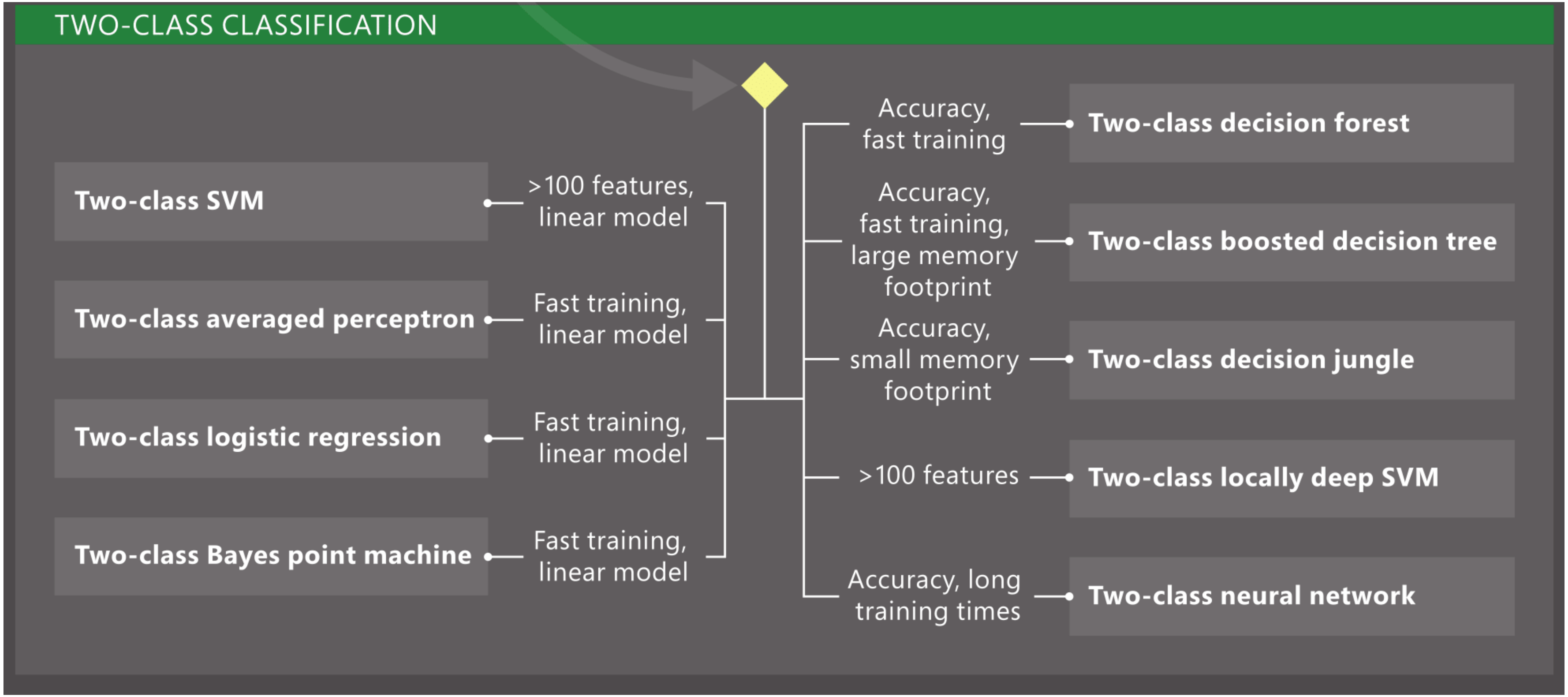
Two

istribution

TWO-CLASS CLASSIFICATION

Accuracy

Which Two-class Algorithm?



TWO-CLASS CLASSIFICATION

Two-class SVM — >100 features, linear model

Two-class averaged perceptron — Fast training, linear model

Two-class logistic regression — Fast training, linear model

Two-class Bayes point machine — Fast training, linear model

Accuracy, fast training — Two-class decision forest

Accuracy, fast training, large memory footprint — Two-class boosted decision tree

Accuracy, small memory footprint — Two-class decision jungle

>100 features — Two-class locally deep SVM

Accuracy, long training times — Two-class neural network

Training Model

Create Train models

Score models

Understanding model evaluation

Adjusting Performance

Adjust parameters

Add new model

Compare performance between models

Summary

Getting Data

Pre-Processing Data

Training Multiple Models

Comparing Multiple Models