Machine learning with Azure machine learning with R extension

Dr. Uros Godnov

Before we start

- Machine learning is suddenly very popular
- All non-scientist and non-statisticians are now data scientist
- Very easy to accomplish something
- No knowledge needed for „something“ to do „something“ that returns „something“

Machine learning

- Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.
- Similar to data mining, but there is a difference!
- „Hierarchy“:
  - **Statistics** *quantifies* numbers
  - **Data Mining** *explains* patterns
  - **Machine Learning** *predicts* with models
  - **Artificial Intelligence** *behaves* and *reasons*

Machine learning

- Facebook's News Feed uses machine learning to personalize each member's feed.
- WolframAlpha engine

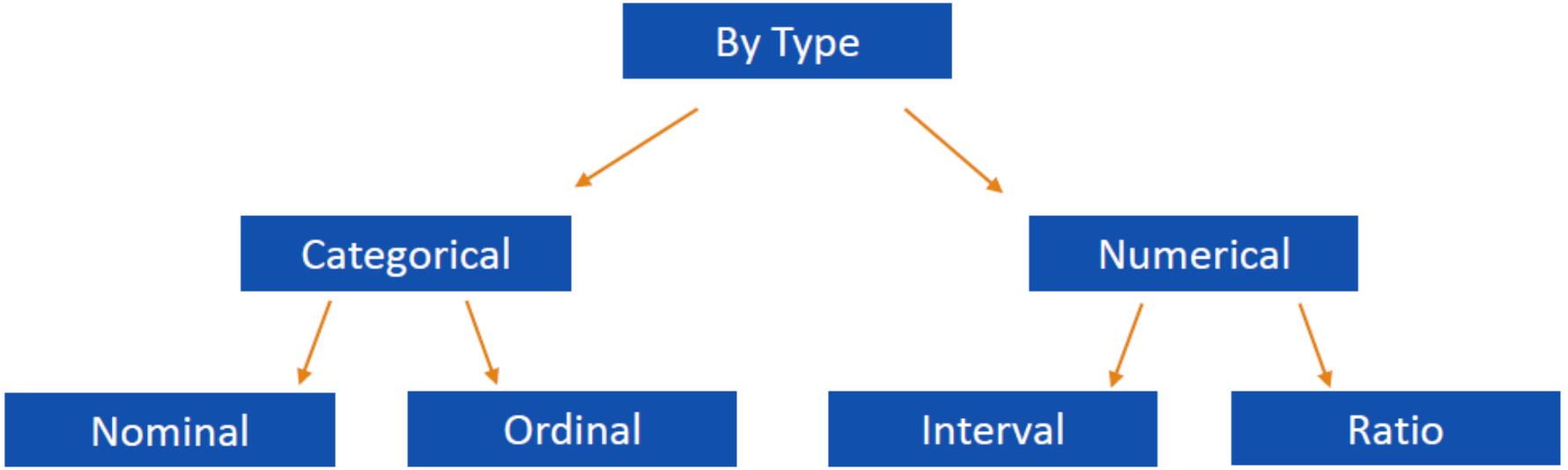
Azure machine learning (free version)

	FREE	STANDARD
Azure Subscription	Not Required	Required
Max Number of Modules per Experiment	100	Unlimited
Max Experiment Duration	1 hour per experiment	Up to 7 days per experiment with a maximum of 24 hours per module
Max Storage Space	10 GB	Unlimited - BYO
Execution / Performance	Single node	Multiple nodes
Production Web API	No	Yes
SLA	No	Yes

Types of data

- Plain text (.txt)
- Comma-separated values (CSV) with a header (.csv) or without (.nh.csv)
- Tab-separated values (TSV) with a header (.tsv) or without (.nh.tsv)
- Hive table
- SQL database table
- OData values
- SVMLight data (.svmlight)
- Attribute Relation File Format (ARFF) data (.arff)
- Zip file (.zip)
- R object or workspace file (.RData)

Variables



gender (male/ Female),
eye colour (blue, green, black,...),
marital status (S, M, D,V,...),
Buyer | Nonbuyer (0 | 1)

No inherited order; all values are the „same“ and all values equal in representation

education (primary, secondary, high school,...),
Salary (<101€,101-200€, 201-300€,..)

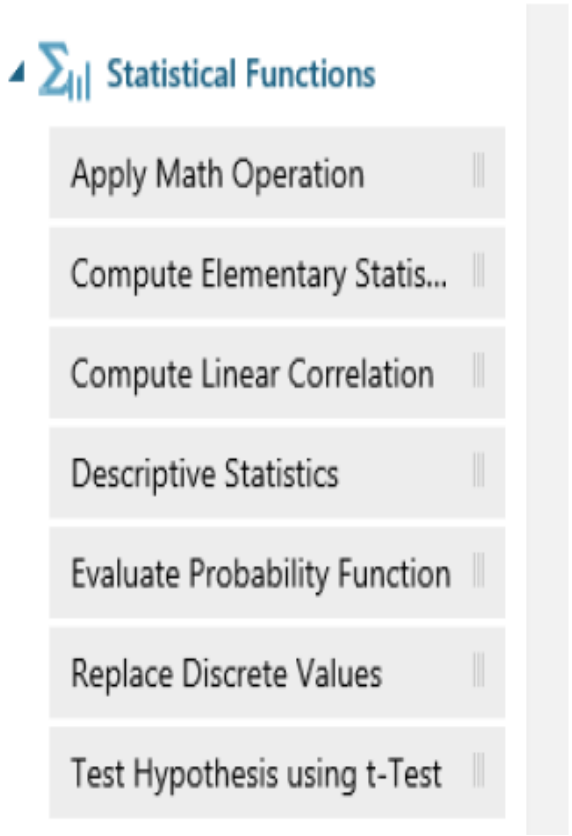
Ordering can be applied; values can be compared with =,>,< and recoded with numbers / classes.

Height (173cm, 196cm),
Invoice Value (24.4€, 42.6€,..)

Ordering can be applied; values can be compared with =,>,< and expressed 2 is 2x times bigger..

Temperature: Celsius vs. Fahrenheit

General statistics



Apply Math Operation -> Applies a mathematical operation to column values
Compute Elementary Statistics -> Calculates specified summary statistics for selected dataset columns
Compute Linear Correlation -> Calculates the linear correlation between column values in a dataset
Descriptive Statistics -> Generates a basic descriptive statistics report for the columns in a dataset
Evaluate Probability Function -> Fits a specified probability distribution function to a dataset
Replace Discrete Values -> Replaces discrete values from one column with numeric values based on another column
Test Hypothesis Using t-Test -> Compares means from two datasets using a t-test

Demo 1

- Automobile price data
- Project column
- Elementary statistics
- Descriptive statistics
- Compute linear correlation

Extensions – data preparation phase

- Pyhton extension
- R extension:
  - ggplot2
  - Preparing data
- print(rownames(installed.packages()))

Back to previous example

- Add Execute R script task
- Inside task add the following:
  - library(PerformanceAnalytics)
  - chart.Correlation(dataset1)

R (visualization) – ggplot2

- ggplot2 → golden standard for plots in R
- Vizualizing using a „grammar“:
  - Data
  - Chart type
  - Smoothing curve
  - Facets
- Calculated columns with function within

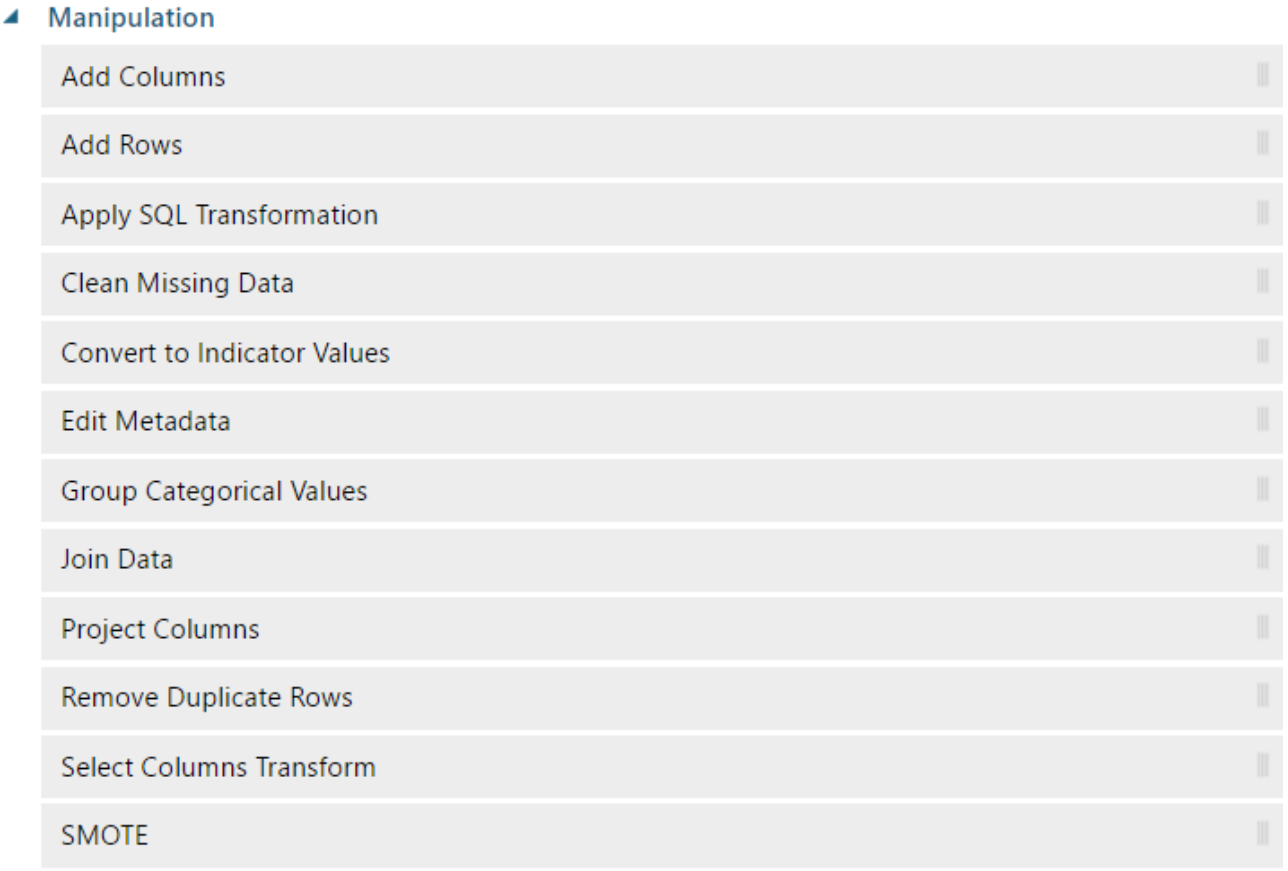
R – ggplot2

- install packages ggplot2 and reshape2
- Show „tips“ data set
- Add calculated column ratio
- Show scatterplot(total_bill,ratio)
- Expand basic graph with sex/time
- Add smooth linear curve

Demo 1/a

- Select columns
- Edit metadata
- Execute R script

Data manipulation/transformation



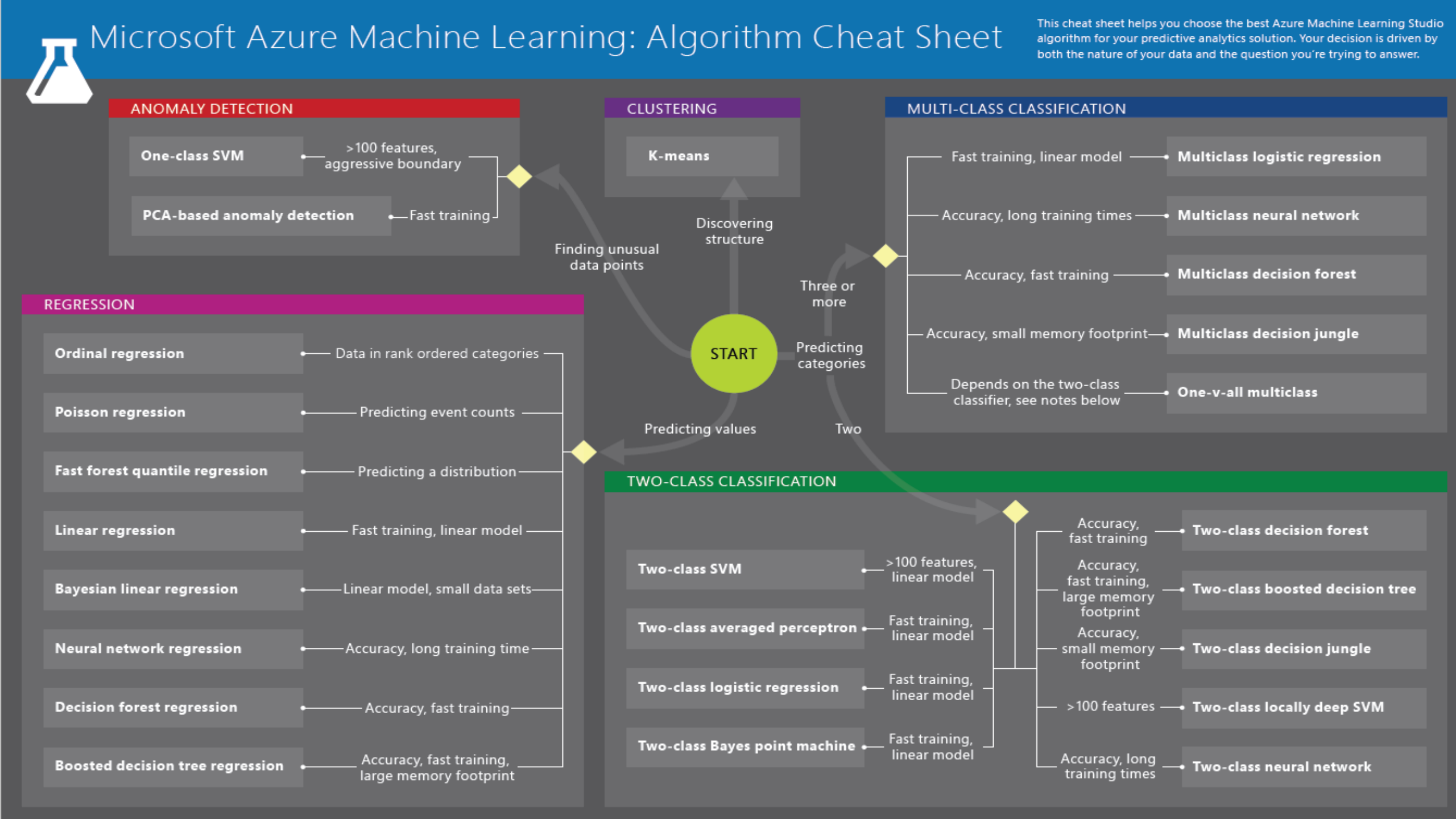
Demo 1/b

- Group into bins
- Split data:
  - random
  - stratified

Now, we are ready!

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.



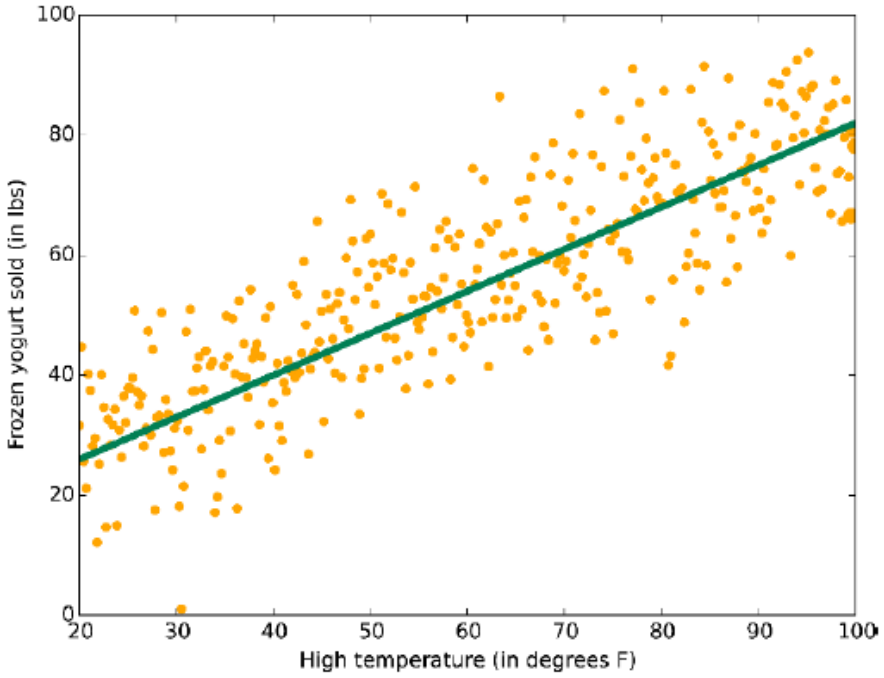
Making list of algorithms more transparent

	Regresssion	Classification	
		Two-class	Multiclass
Average Perceptron		✓	
Bayes Point Machine		✓	
Decision Forest	✓	✓	✓
Decision Jungle		✓	✓
Decision Tree	✓	✓	
Fast Forest	✓		
Linear Regression	✓		
Bayes Linear Regression	✓		
Log Regression		✓	✓
Neural Network	✓	✓	✓
Ordinal Regression	✓		
Poisson Regression	✓		
SVM		✓	
SVM Deep Support		✓	

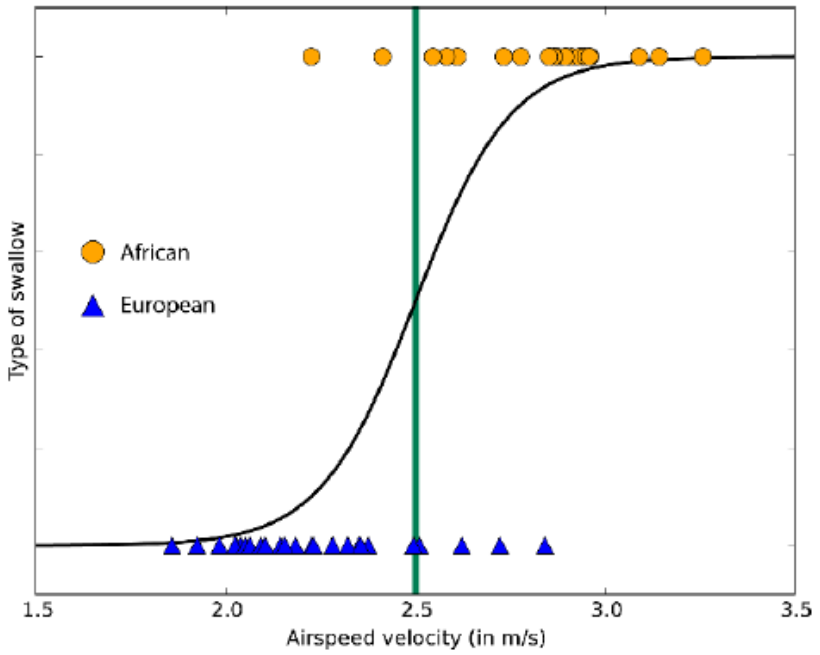
Algorithms in Theory

Regression

Linear and Logistic



Azure ML: Linear Regression

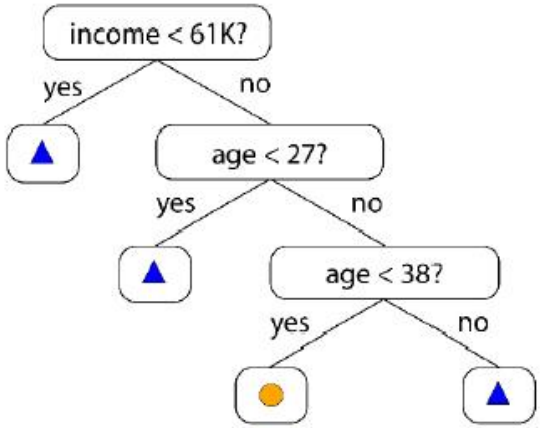


Azure ML: Two-class Classification Logistic Regression
Multiclass classification Logistic Regression

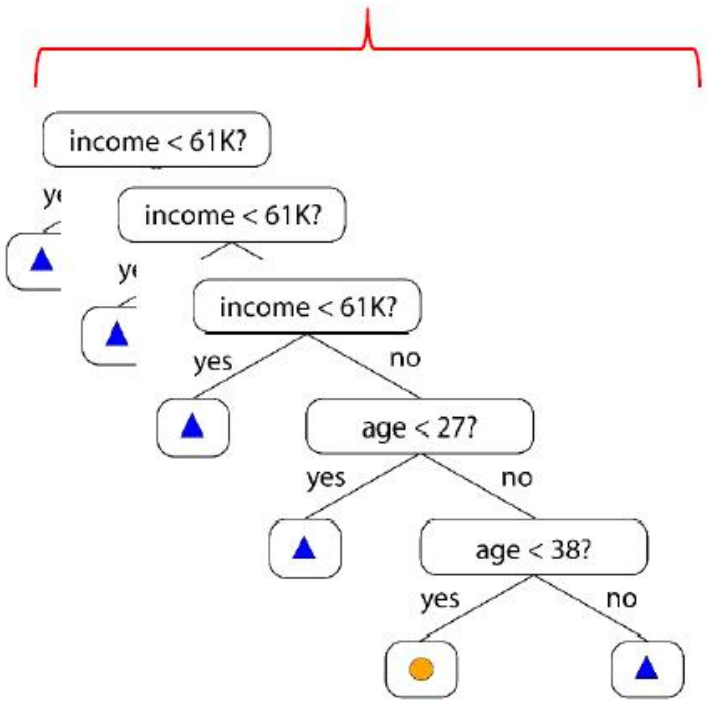
Algorithms in Theory

Decision Tree, Decision Forests, Decision Jungles

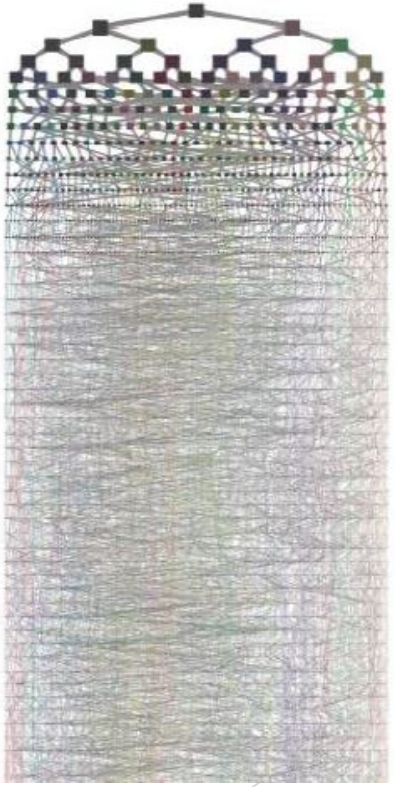
Decision tree



Decision Forest



Decision Jungle

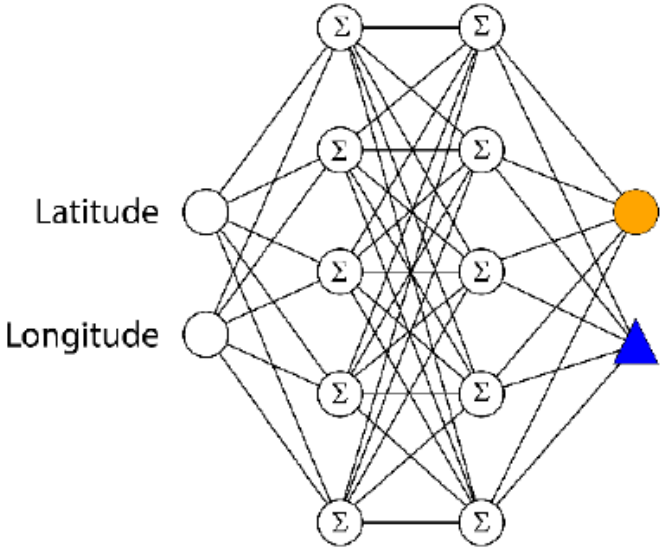


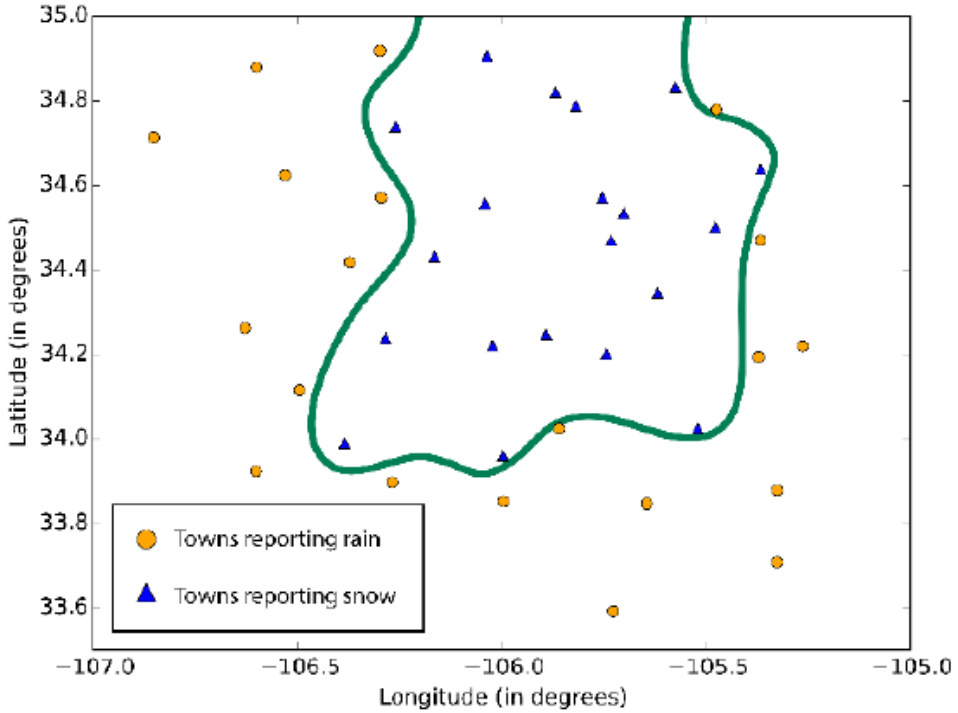
Azure ML: Regression boosted decision tree
Two-class classification boosted decision tree

Azure ML: Regression decision forrest

Algorithms in Theory

Neural networks and perceptrons

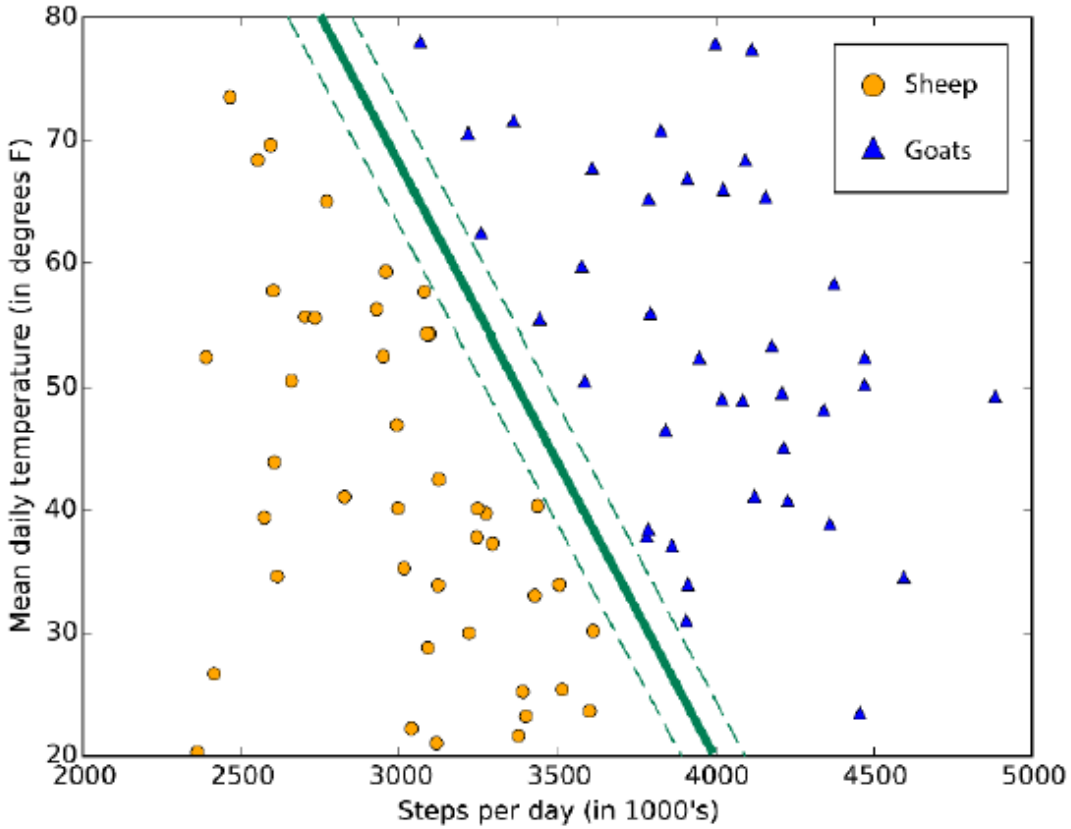




Azure ML: Regression Neural networks
Two Class classification Neural networks
Multi Class classification Neural networks

Algorithms in Theory

SVM



Azure ML: Two Class classification SVM
Two Class classification locally deep SVM

Anomaly detection SVM

Regression algorithms

- Regression is method for estimating relations among parameters/varibles.
- Linear vs. Logistic (linear combination of parameters vs. Logistic combination of parameters)
- Typical Problem would be predicting Y ; a numeric value.
- Typical Azure Algorithms
  - Boosted Decision Tree Regression
  - Decision Forest Regression
  - Linear Regression
  - Bayesian Linear Regression

Linear Regression

Solution method: Ordinary Least Squares

L2 regularization weight: 0.001

Include intercept term

Random number seed:

Allow unknown categ...

Bayesian Linear Regression

Regularization weight: 1

Allow unknown categ...

Decision Forest Regression

Resampling method: Bagging

Create trainer mode: Single Parameter

Number of decision trees: 8

Maximum depth of the d...: 32

Number of random splits...: 128

Minimum number of sam...: 1

Allow unknown value...

Boosted Decision Tree Regressi...

Create trainer mode: Single Parameter

Maximum number of leav...: 20

Minimum number of sam...: 10

Learning rate: 0.2

Total number of trees con...: 100

Random number seed:

Allow unknown categ...

Neural Network Regression

Create trainer mode: Single Parameter

Hidden layer specification: Fully-connected case

Number of hidden nodes: 100

Learning rate: 0.005

Number of learning iterat...: 100

The initial learning weight...: 0.1

The momentum: 0

The type of normalizer: Min-Max normalizer

Shuffle examples

Random number seed:

Evaluating regression algorithms

Mean Absolute Error: $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$

Metrics to measure how close predictions are to eventual outcomes

Root Mean Square Error: $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(f_i - y_i)^2}{n}}$

Metrics of differences between predicted values and actual values.

Relative square Error: $RSE = \frac{\sum_{i=1}^{n}(f_i - y_i)^2}{\sum_{i=1}^{n}(\overline{y} - y_i)^2}$

Coeff. of Determination: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Summarization of regression model how well fits a statistical model; R^2 = 1 model is perfect, respectively

Demo 1/c

- Regression model
- Score
- Evaluation

Comparison of Regression Algorithms

Regression Algorithm	Accuracy	Training time	Linearity	Customization	Predicting Variable	Type of independant variable(s)	Data Quantity
linear	Good	Fast	Excellent	Good	Interval	Any	small to big
Bayesian linear	Good	Fast	Excellent	Moderate	Interval	Any	big
decision forest	Excellent	Moderate	Good	Good	Interval	Any	
boosted decision tree	Excellent	Fast	Good	Good	Interval	Any	big
fast forest quantile	Excellent	Moderate	Moderate	Excellent	Distribution (Interval)	Any	
neural network	Excellent	Slow	Moderate	Excellent	Interval	Any	smaller
Poisson	Good	Moderate	Excellent* (log linear)	Good	Interval (counts)	Any	small to big
ordinal	Good	Moderate	Excellent	None	Ordinal (order)	Any	small to big

Scale:

Excellent	Good	Moderate
Fast	Moderate	Slow

Two-class clasification

- Creates classification estimates for label / prediction variable with dichotomious values
- Typical Problem would be predicting a binary class for label variable
- Typical Azure Algorithms
  - Boosted Decision Tree two-class
  - Decision Forest two-class
  - Decision Jungle two-class
  - Logistic Regression two-class
  - Neural Network two-class
  - Averaged Perceptron two-class
  - SVM two-class

Two-Class Boosted Decision Tree

Create trainer mode: Single Parameter

Maximum number of leav...: 20

Minimum number of sam...: 10

Learning rate: 0.2

Number of trees construc...: 100

Random number seed:

Two-Class Decision Jungle

Resampling method: Bagging

Create trainer mode: Single Parameter

Number of decision DAGs: 8

Maximum depth of the d...: 32

Maximum width of the de...: 128

Number of optimization s...: 2048

Two-Class Averaged Perceptron

Create trainer mode: Single Parameter

Learning rate: 1

Maximum number of iter...: 10

Random number seed:

Allow unknown categ...: ☑

Two-Class Bayes Point Machine

Number of training iterati...: 30

Include bias: ☑

Allow unknown value...: ☑

Two-Class Support Vector Mac...

Create trainer mode: Single Parameter

Number of iterations: 1

Lambda: 0.001

Normalize features: ☑

Project to the unit-sp...: ☐

Random number seed:

Allow unknown categ...: ☑

Two-Class Logistic Regression

Create trainer mode: Single Parameter

Optimization tolerance: 1E-07

L1 regularization weight: 1

L2 regularization weight: 1

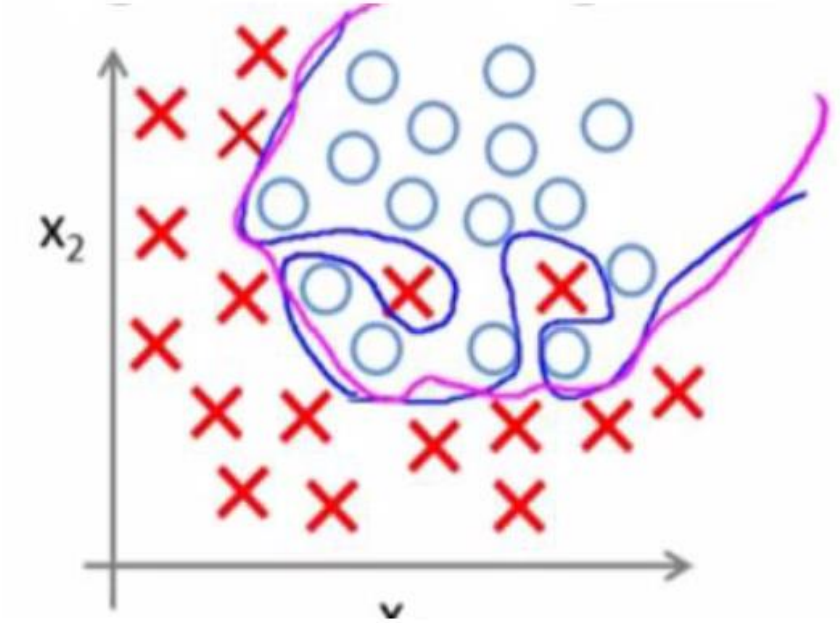
Memory size for L-BFGS: 20

Random number seed:

Allow unknown categ...: ☑

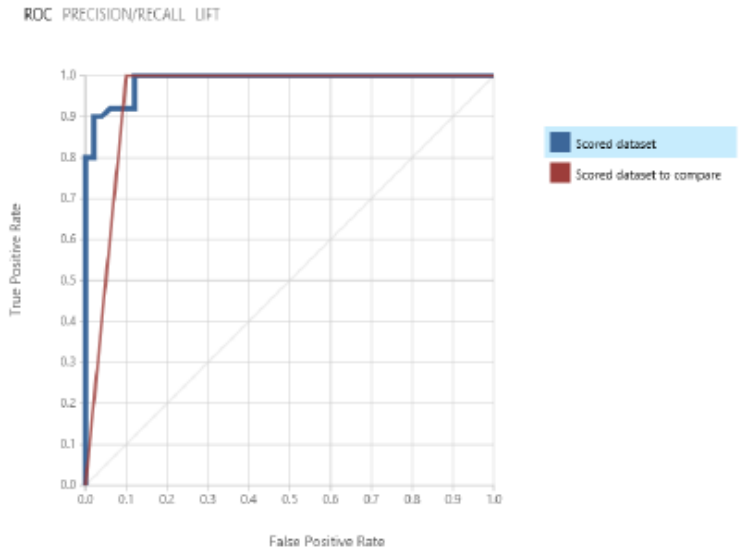
Regularization weight

- Used for avoiding overfitting.
- L1, L2 penalized estimation methods shrink the estimates of regre. coefficient towards zero in relation to maximize likelihood of estimates.
- L1 - for sparse, high-dimensional model
- L2 – for dense (or smaller) model and computationally efficient



Evaluating two-class Classification

ROC (AUC) Curve / Precision / Lift Chart



AUC/ROC:

<= 0.5 -- ☹☹

0.5 – 0.6 -- ☹

0.6 – 0.7 -- 😐

0.7 – 0.8 -- ☺

0.8 – 0.9 -- ☺☺

0.9 – 1 -- WTF?

Classification Matrix / Confussion matrix / Metrics

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
46	4	0.900	0.885	0.5	0.988
False Positive	True Negative	Recall	F1 Score		
6	44	0.920	0.902		
Positive Label	Negative Label				
1	0				

Evaluating two-class Classification

Metrics

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
46	4	0.900	0.885	0.5	0.988

False Positive	True Negative	Recall	F1 Score
6	44	0.920	0.902

Positive Label	Negative Label
1	0

True Positive (TP) – correctly identified: Buyes is classified as Buyer
False Positive (FP) – Incorrectly identified: Buyes is classified as non-buyer
True Negative (TN) - correctly identified: Non-buyes is classified as non-buyer
False Negative (FN) - Incorrectly identified: Non-buyes is classified as buyer

Accuracy (TP + TN) / (TP + TN + FP + FN) – Proportion of correctly classified
Precision TP / (TP+FP) – Proportion of positive cases classified correctly
Sensitivity* TP / (TP + TN) - Proportion of actual positive cases classified correctly
Score 2TP / (2TP + FP + FN) – Harmonic mean of precision and Sensitivity

Demo 2

- Adult Census Income Binary Classification dataset
- Download dataset and create ggplot in R
- Focus only on USA
- Omit fnlwgt, education-num, capital-gain, capital-loss

Comparison of Two-class Classification Algorithms

Two-class classification	Accuracy	Training time	Linearity	Customization	Predicting Variable	Type of independant variable(s)	Data Quantity
logistic regression	Good	Fast	Excellent	Good	dichotomous / binary	Any	small-big
decision forest	Excellent	Moderate	Good	Good	dichotomous / binary	Any	small-big
decision jungle	Excellent	Moderate	Good	Good	dichotomous / binary	Any	big
boosted decision tree	Excellent	Moderate	Good	Good	dichotomous / binary	Any	big
neural network	Excellent	Slow	Moderate	Excellent	dichotomous / binary	Any	
averaged perceptron	Good	Moderate	Excellent	Moderate	dichotomous / binary	Any	
support vector machine	Excellent	Moderate	Excellent	Good	dichotomous / binary	Any	big
locally deep support vector machine	Good	Slow	Good	Excellent	dichotomous / binary	Any	big
Bayes' point machine	Moderate	Moderate	Excellent	Moderate	dichotomous / binary	Any	

Scale:

Excellent	Good	Moderate
Fast	Moderate	Slow

Multi-class Classification

- Creates classification estimates for label / prediction variable with 2+ classes
- Decision trees vs. Logistic Regression vs. Neural Network
- Typical Problem would be predicting a class for label variable
- Typical Azure Algorithms
  - Decision Forest Multiclass
  - Decision Jungle Multiclass
  - Logistic Regression Multiclass
  - Neural Network Multiclass

Multiclass Decision Forest

Resampling method

Bagging

Create trainer mode

Single Parameter

Number of decision trees

8

Maximum depth of the d...

32

Number of random splits...

128

Minimum number of sam...

1

Allow unknown value...

Multiclass Decision Jungle

Resampling method

Bagging

Create trainer mode

Single Parameter

Number of decision DAGs

8

Maximum depth of the d...

32

Maximum width of the de...

128

Number of optimization s...

2048

Allow unknown value...

Multiclass Logistic Regression

Create trainer mode

Single Parameter

Optimization tolerance

1E-07

L1 regularization weight

1

L2 regularization weight

1

Memory size for L-BFGS

20

Random number seed

Allow unknown categ...

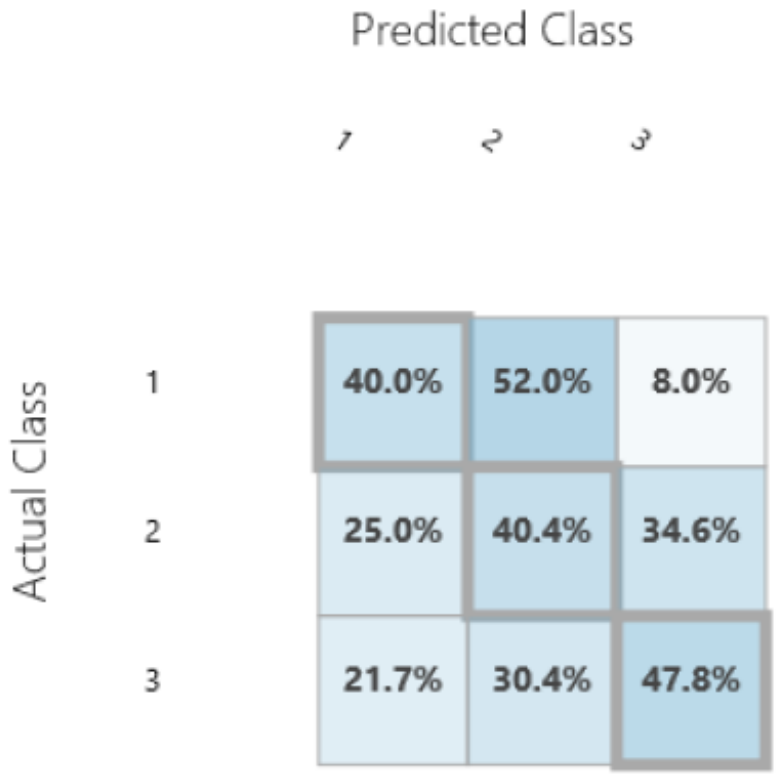
Evaluating multi-class Classification

Metrics

Metrics

Overall accuracy	0.42
Average accuracy	0.613333
Micro-averaged precision	0.42
Macro-averaged precision	0.408059
Micro-averaged recall	0.42
Macro-averaged recall	0.427369

Confusion Matrix



Demo 3

- Steel data

Comparison of Multi-class Classification Algorithms

Multi-class classification	Accuracy	Training time	Linearity	Customization	Predicting Variable	Type of independant variable(s)	Data Quantity
logistic regression	Good	Fast	Excellent	Good	Nominal / ordinal (with 2+ classes)	any	small-big
decision forest	Excellent	Moderate	Good	Good	Nominal / ordinal (with 2+ classes)	any	big
decision jungle	Excellent	Moderate	Good	Good	Nominal / ordinal (with 2+ classes)	any	big
neural network	Excellent	Slow	Moderate	Excellent	Nominal / ordinal (with 2+ classes)	any	small

Scale:

Excellent	Good	Moderate
Fast	Moderate	Slow

Good to know!

- Importing large dataset → zip it → upload
- When using it → use Unpack zipped datasets

Webservice

- Publish model as webservice
- Deploy it and connect with excel via Machine learning add-in

Jupiter Notebook

- Jupyter notebooks provide an interactive environment for exploring data and collaborating with other data scientists.
- Jupyter.org
- When running in Azure ML, we don‘t need to worry about security
- 50 kernells available

Jupyter Notebook

- 3 types of cells:
  - Code
  - Raw
  - Markdown
- In Azure ML – only csv files
- To run code (ctrl+enter)