

[xwMOOC R Meetup 6회- 세션1 발표]

- H2O 활용 Kaggle Iris dataset 분석 사례 -

서강대 머신러닝 Lab 황문기 교수

*This document is confidential and is intended solely for the use and information of the client to whom it is addressed.
Sogang Machine Learning Lab*



2018. 01. 17



목 차

1. Kaggle 붓꽃(IRIS) Dataset 사례

2. H2o 분석

- 딥러닝

- GBM(Gradient Boosting Model)

3. Q&A

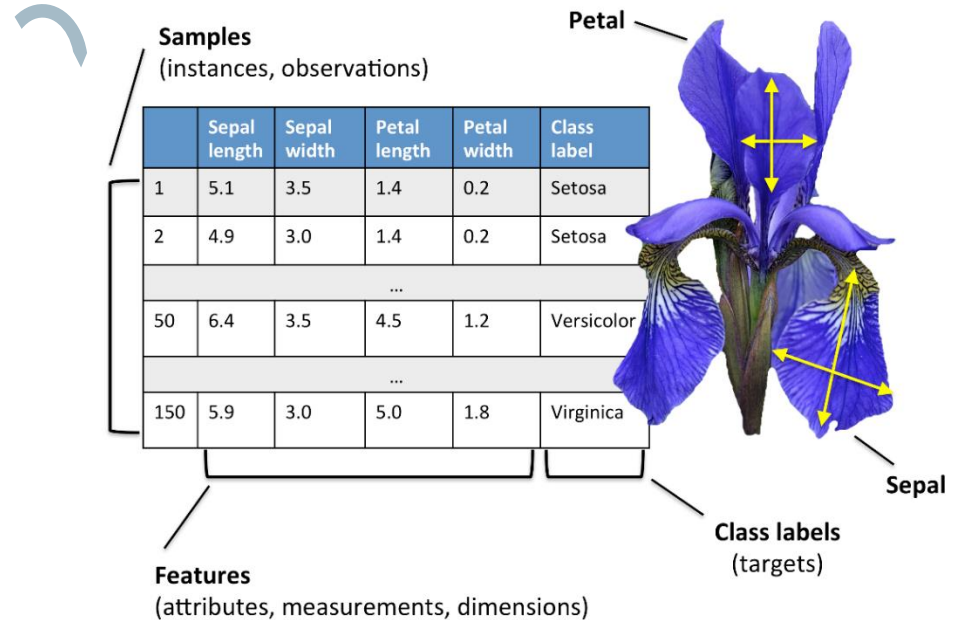
1. 붓꽃(IRIS) dataset ?

붓꽃(IRIS)은 3개의 종이 있으며, IRIS 분석 dataset은 150개 샘플데이터로 꽃잎 (Petal) 길이/폭 2개, 꽃받침(Sepal)의 길이/폭 2개 총 4개의 변수로 구성됨

(그림1) 붓꽃(Iris) 3종



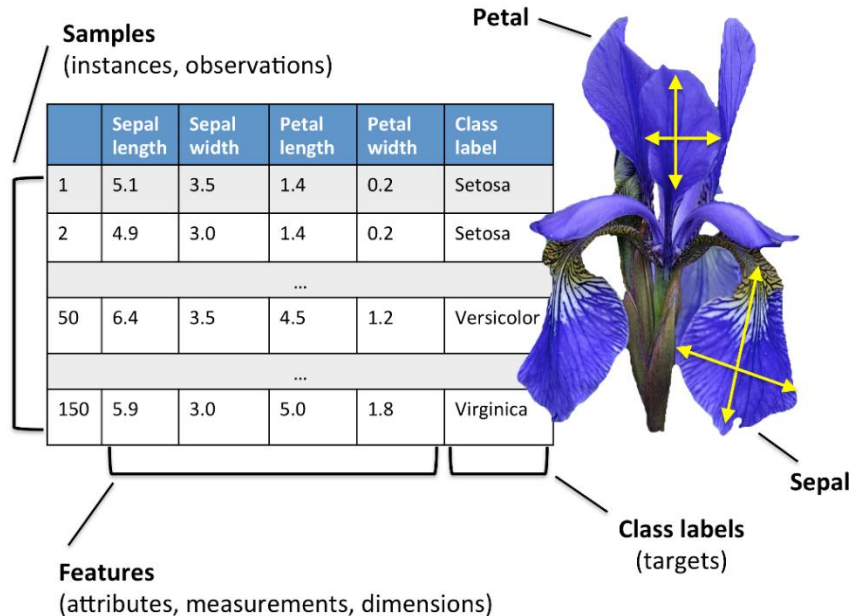
(그림2) Iris Data Set 구성



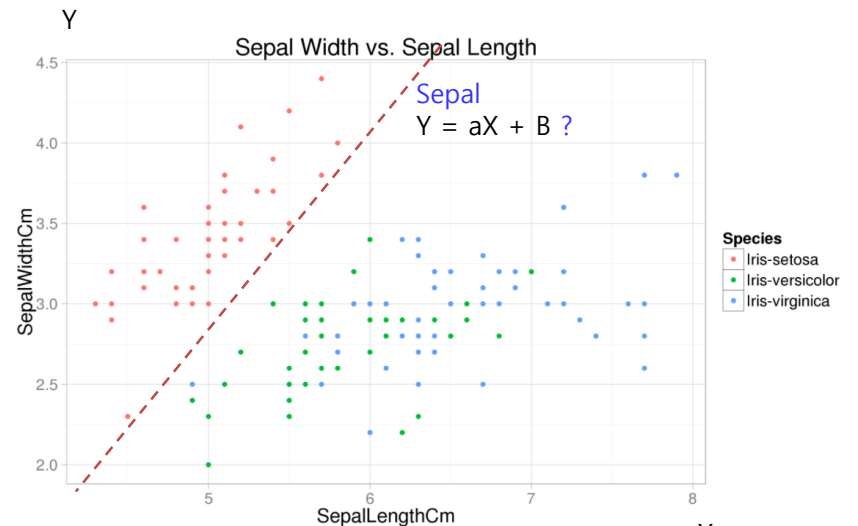
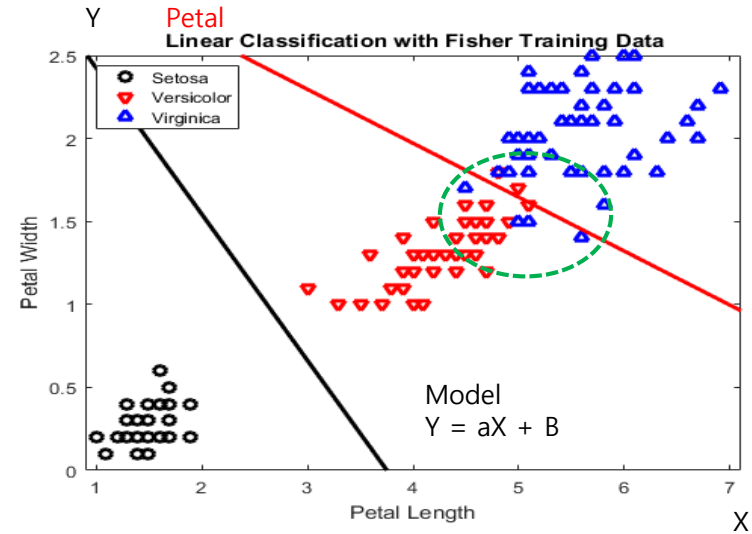
주) 150개 샘플, 꽃받침(Sepal) 폭/넓이, 꽃잎(Petal) 폭/넓이

2. 붓꽃(IRIS) dataset 사전분석

꽃잎(Petal) 길이/폭 2개, 꽃받침(Sepal) 길이/폭 2개 총 4개의 변수중에, Sepal 데이터는 선형적 분리가 어려울 것으로 판단되어 Petal을 먼저 사전탐색하기로 결정..



주) 150개 샘플, 꽃받침(Sepal) 폭/넓이, 꽃잎(Petal) 폭/넓이



3. Kaggle Iris Species dataset

Kaggle사이트의 Iris Species dataset 입수

The image shows a screenshot of the Kaggle website. The main focus is on the 'Iris Species' dataset page, which is a 'Reviewed Dataset'. The dataset description reads: 'Classify iris plants into three species in this classic dataset'. It is associated with the 'UCI Machine Learning' repository and was last updated a year ago. The dataset has 461 votes.

Below the dataset description, there are tabs for 'Overview', 'Data', 'Kernels', 'Discussion', and 'Activity'. The 'Data' tab is selected, showing a list of files: 'database.sqlite', 'Iris', and 'Iris.csv'. The 'Iris' file is highlighted.

On the right side of the 'Data' tab, there is a section for 'Iris Species' with a description: 'Classify iris plants into three species in this classic dataset'. Below this, there is a section 'About this Dataset' which states: 'The Iris dataset was used in R.A. Fisher's classic 1936 paper, [The Use of Multiple Measurements in Taxonomic Problems](#), and can also be found on the [UCI Machine Learning Repository](#). It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other. The columns in this dataset are: Id, SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, Species'.

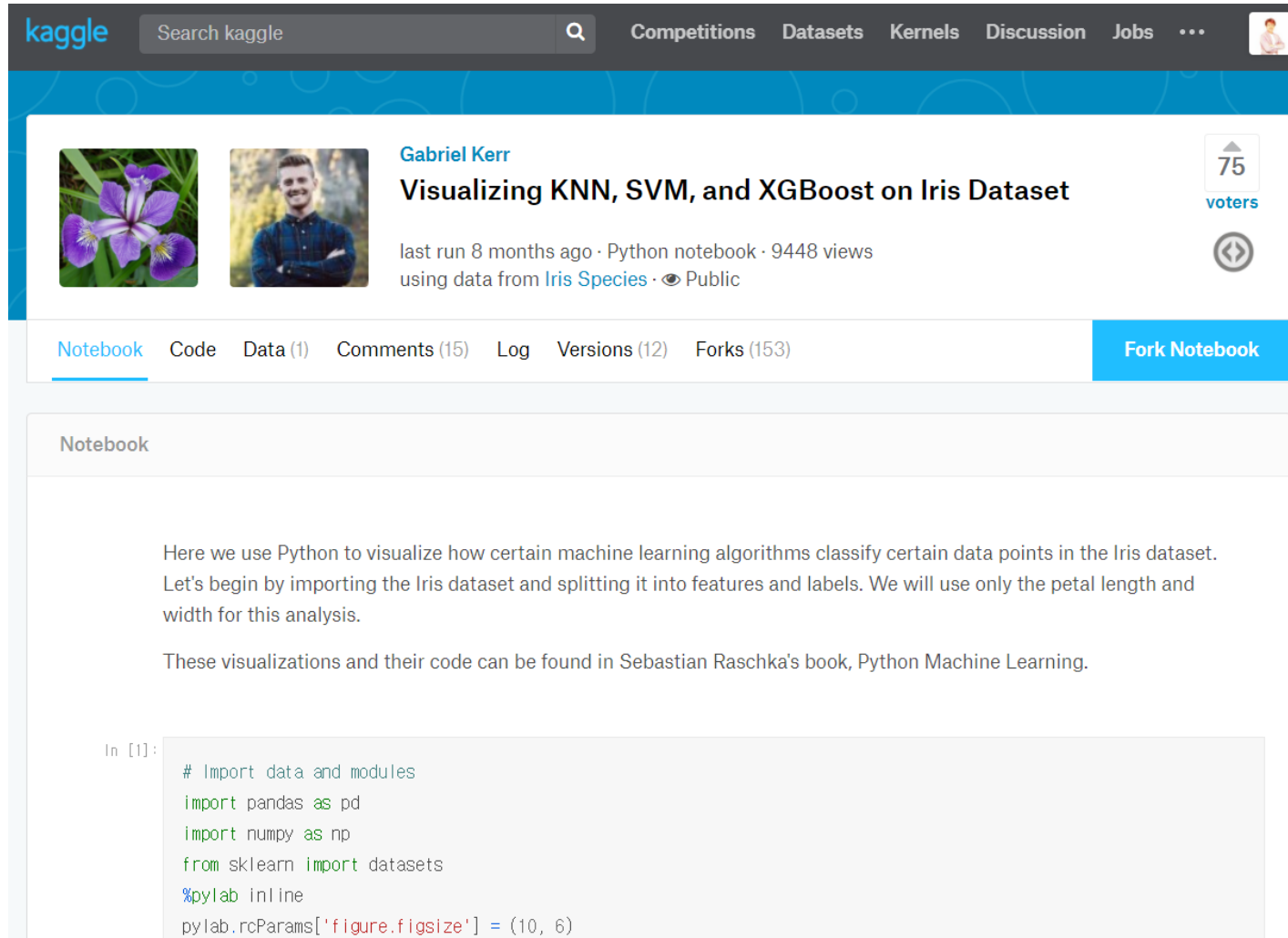
Overlaid on the dataset page is a browser window showing a Kaggle notebook titled 'Visualizing KNN, SVM, and XGBoost on Iris Dataset' by Gabriel Kerr. The notebook was last run 9 months ago, has 9870 views, and is using data from the 'Iris Species' dataset. It has 77 voters and is public. The notebook page includes tabs for 'Notebook', 'Code', 'Data (1)', 'Comments (15)', 'Log', 'Versions (12)', and 'Forks (153)'. A 'Fork Notebook' button is visible.

At the bottom of the screenshot, the Windows taskbar is visible, showing the system tray with the date and time: '오후 2:49 2018-01-17'.

<https://www.kaggle.com/mgabrielkerr/visualizing-knn-svm-and-xgboost-on-iris-dataset>

3. Kaggle Iris Species Case

Kaggle Iris Species dataset visualization 사례 – KNN, SVM, XGBoost



The screenshot shows a Kaggle notebook page. At the top, there's a navigation bar with 'kaggle' logo, a search bar, and links for 'Competitions', 'Datasets', 'Kernels', 'Discussion', and 'Jobs'. Below this, the notebook header features a profile picture of Gabriel Kerr, the title 'Visualizing KNN, SVM, and XGBoost on Iris Dataset', and a '75 voters' badge. The notebook content starts with a text block: 'Here we use Python to visualize how certain machine learning algorithms classify certain data points in the Iris dataset. Let's begin by importing the Iris dataset and splitting it into features and labels. We will use only the petal length and width for this analysis. These visualizations and their code can be found in Sebastian Raschka's book, Python Machine Learning.' Below the text is a code cell labeled 'In [1]:' containing the following Python code:

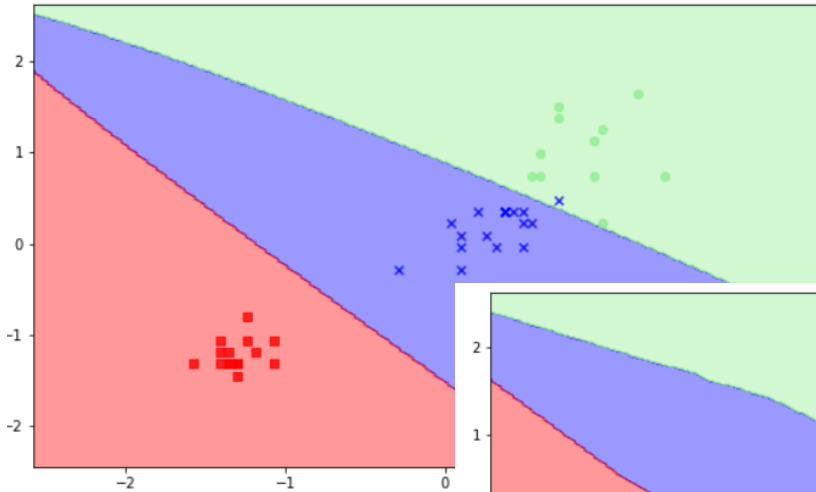
```
In [1]: # Import data and modules
import pandas as pd
import numpy as np
from sklearn import datasets
%pylab inline
pylab.rcParams['figure.figsize'] = (10, 6)
```

<https://www.kaggle.com/mgabrielkerr/visualizing-knn-svm-and-xgboost-on-iris-dataset>

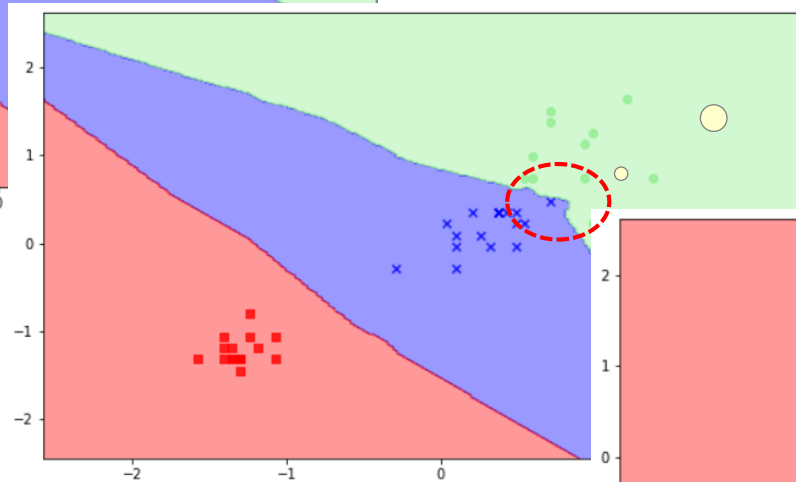
4. Kaggle Iris Species Case 결과

Kaggle Iris Visualization 사례에서, 꽃잎(Petal) 데이터 머신러닝 결과중에 KNN 결과가 가장 좋게 나왔으며, KNN 분석결과 150개 중에 1개의 Error를 기록

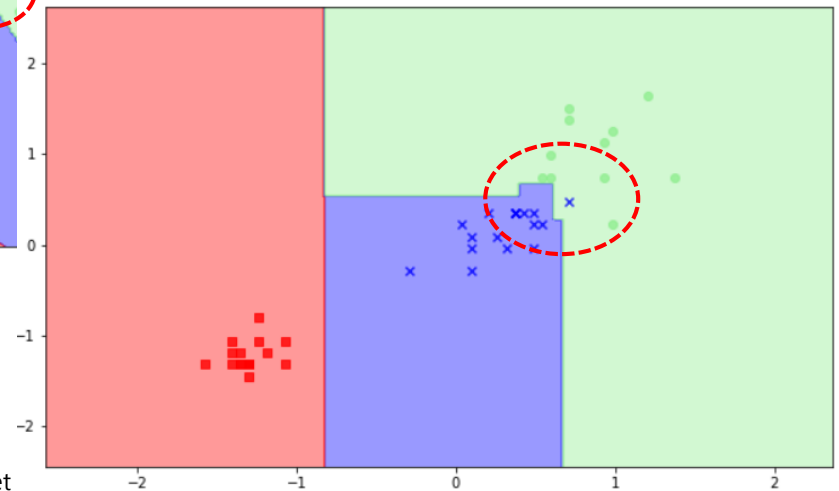
(그림4) Linear SVC



(그림5) KNN 결과



(그림6) XG Boost 결과



돌연변이?
네잎 클로버?

출처) <https://www.kaggle.com/mgabrielkerr/visualizing-knn-svm-and-xgboost-on-iris-dataset>

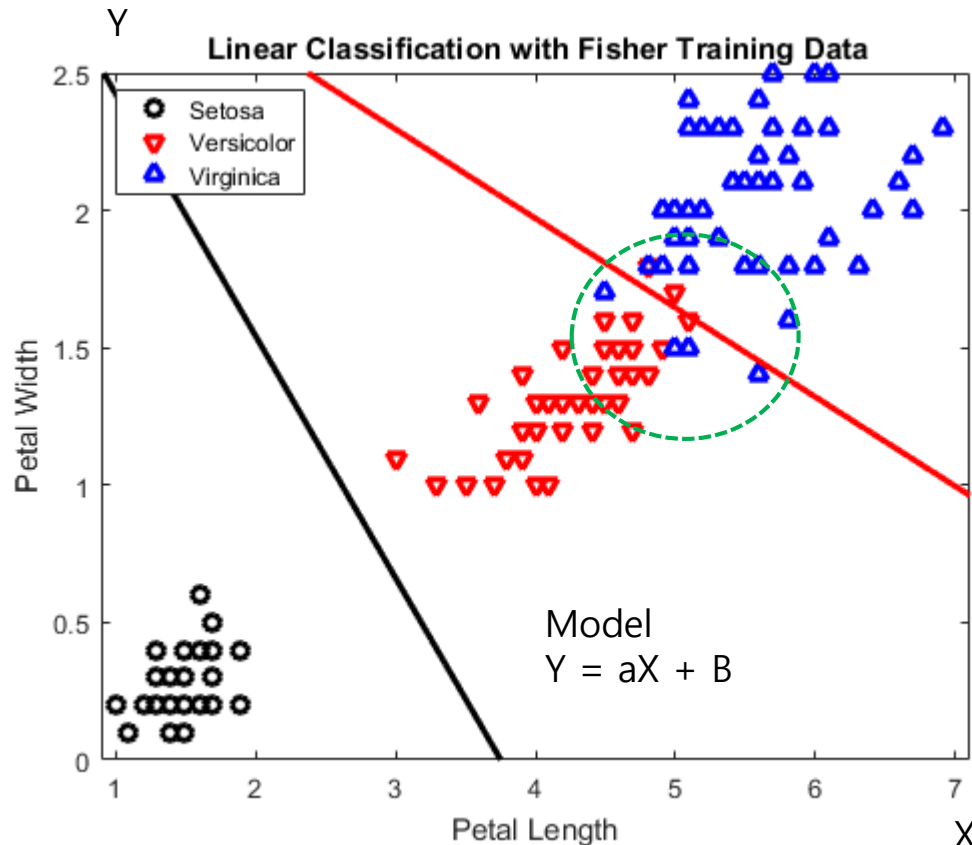
목 차

1. Kaggle IRIS(붓꽃) Dataset 사례
2. H2o 분석
 - 딥러닝
 - GBM(Gradient Boosting Model)
3. Q&A

1. H2O 분석 Approach

Petal(꽃잎)과 Sepal(꽃받침) dataset을 Deep Learning과 GBM(Gradient Boosting Machine)으로 분석하여 결과를 비교함

(그림3) 꽃잎(Petal) 데이터 분포



Iris 분석 Approach

- Dataset
 - X축 : 꽃잎(Petal)길이(length)
 - Y축 : 꽃잎(Petal)폭(width)
 - > Virginica와 Versicolor 중첩부분(6개)
- Deep Learning
 - 활성화함수, epochs 증가
 - > 개선효과
- GBM
 - ntrees 50 -> 500
 - > 증가할수록 성능 개선효과?

2. H2O 제공 Model

딥러닝, GBM 등 12개의 알고리즘 선택가능..

The screenshot shows the H2O Flow web interface. The browser tabs include ' SAINT', 'Booz Data Science', 'Wireless', '(57) XwMOOC R N', 'Download H2O 3.', 'H2O Flow', and 'Visualizing KNN, S'. The address bar shows 'localhost:54321/flow/index.html'. The interface has a menu bar with 'Flow', 'Cell', 'Data', 'Model', 'Score', 'Admin', and 'Help'. The main area is titled 'Untitled Flow' and contains a 'Build a Model' section. A dropdown menu for 'Select an algorithm:' is open, listing 12 algorithms: Gradient Boosting Machine, Aggregator, Deep Learning, Distributed Random Forest, Gradient Boosting Machine (highlighted), Generalized Linear Modelling, Generalized Low Rank Modeling, K-means, Naive Bayes, Principal Components Analysis, Stacked Ensemble, and Word2Vec. A large blue arrow points from the 'Gradient Boosting Machine' option in the dropdown to a larger, semi-transparent box that lists the same 12 algorithms. Below the dropdown, a table of parameters for the selected algorithm is visible, with a red dashed box around it. The parameters include 'ignore_const_cols', 'ntrees', 'max_depth', 'min_rows', 'nbins', 'seed', 'learn_rate', 'sample_rate', and 'col_sample_rate'. The right sidebar contains 'OUTLINE', 'FLOWS', 'CLIPS', and 'HELP' tabs, with 'HELP' selected. It includes a 'Help' section with 'Using Flow for the first time?' and 'Quickstart Videos', a 'STAR H2O ON GITHUB!' section with a star icon and '2,762', and a 'GENERAL' section with links to 'Flow Web UI...', 'Importing Data', 'Building Models', 'Making Predictions', 'Using Flows', and 'Troubleshooting Flow'. There is also an 'EXAMPLES' section and an 'H2O REST API' section with links to 'Routes' and 'Schemas'. The bottom status bar shows 'Ready' and 'Connections: 0 H2O'. The Windows taskbar at the bottom shows the system tray with the date '2018-01-17' and time '오후 12:21'.

Gradient Boosting Machine (Algorithm)

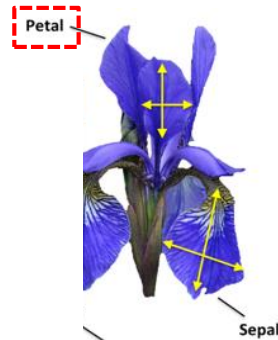
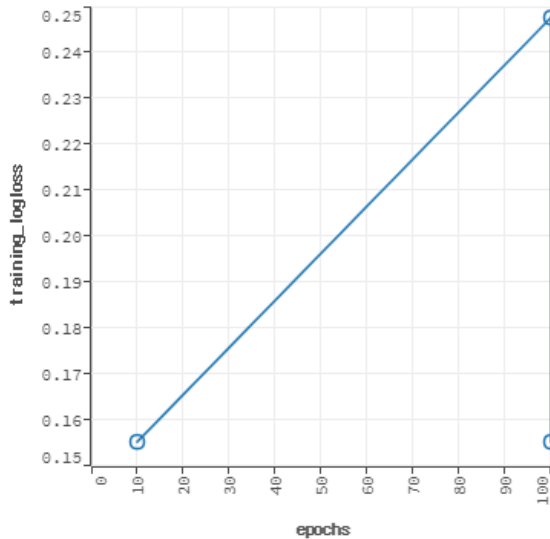
- Aggregator
- Deep Learning
- Distributed Random Forest
- Gradient Boosting Machine
- Generalized Linear Modelling
- Generalized Low Rank Modeling
- K-means
- Naive Bayes
- Principal Components Analysis
- Stacked Ensemble
- Word2Vec

Parameter	Description
ignore_const_cols	Ignore constant columns.
ntrees	Number of trees.
max_depth	Maximum tree depth.
min_rows	Fewest allowed (weighted) observations in a leaf.
nbins	For numerical columns (real/int), build a histogram of (at least) this many bins, then split at the best point
seed	Seed for pseudo random number generator (if applicable)
learn_rate	Learning rate (from 0.0 to 1.0)
sample_rate	Row sample rate per tree (from 0.0 to 1.0)
col_sample_rate	Column sample rate (from 0.0 to 1.0)

3. H2O 분석 - 딥러닝(1/2)

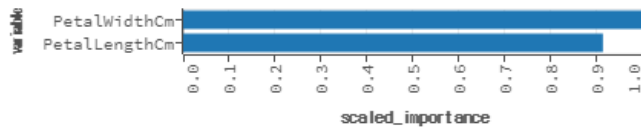
Sepal(꽃받침) 무시, 활성화함수 : Rectifier, 히든레이어 200,200 epochs 100
 분류결과는 versicolor 8개, virginica 4개, 총 12개 Error

SCORING HISTORY - LOGLOSS



변수 중요도
 꽃잎폭(1.0) > 꽃잎 길이(0.92)

VARIABLE IMPORTANCES



TRAINING METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

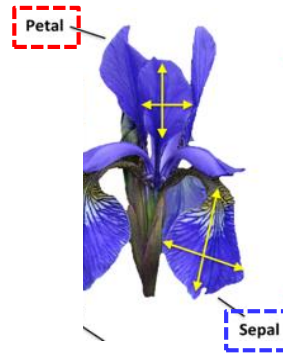
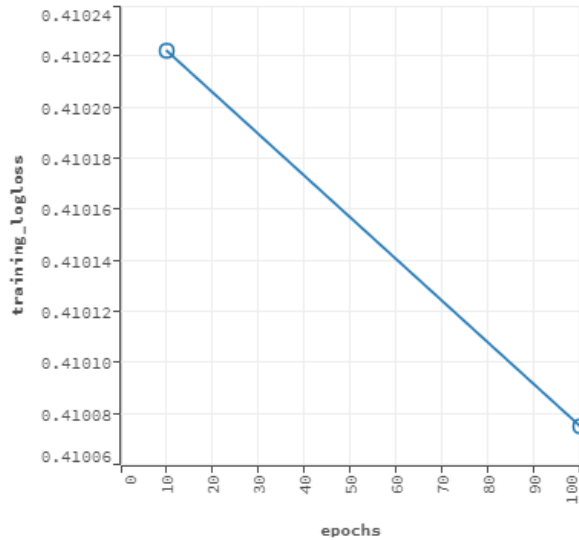
	Iris-setosa	Iris-versicolor	Iris-virginica	Error	Rate
Iris-setosa	50	0	0	0	0 / 50
Iris-versicolor	6	42	2	0.1600	8 / 50
Iris-virginica	0	4	46	0.0800	4 / 50
Total	56	46	48	0.0800	12 / 150



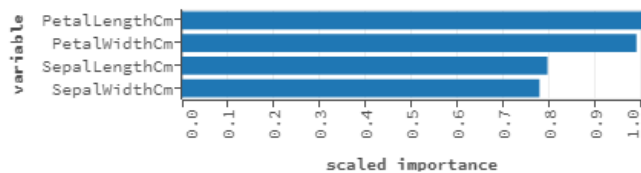
3. H2O 분석 - 딥러닝(2/2)

Sepal(꽃받침) 포함, 활성화함수 : Rectifier, 히든레이어 200,200, epochs 100
 virginica 13개 Error, 딥러닝은 변수조정이 관련이 있고 D/L 부적합으로 판단...

SCORING HISTORY - LOGLOSS



VARIABLE IMPORTANCES



variable	relative_importance	scaled_importance	percentage
PetalLengthCm	1.0	1.0	0.2807
PetalWidthCm	0.9889	0.9889	0.2776
SepalLengthCm	0.7953	0.7953	0.2233
SepalWidthCm	0.7778	0.7778	0.2184

TRAINING METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: P

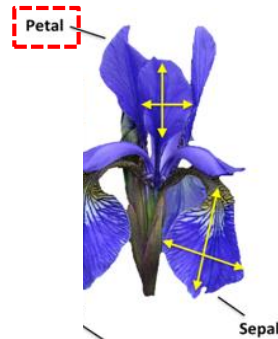
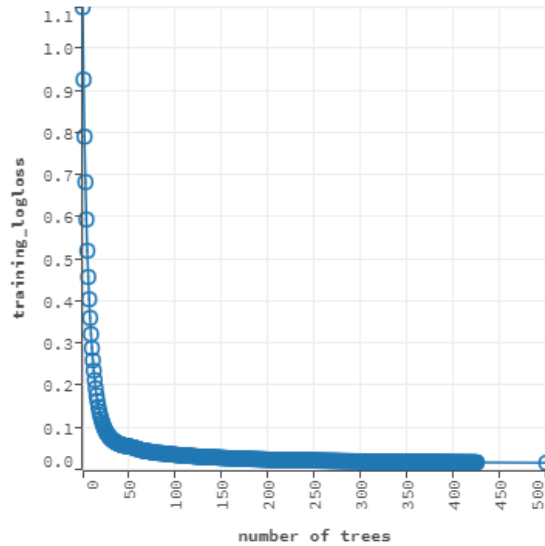
	Iris-setosa	Iris-versicolor	Iris-virginica	Error	Rate
Iris-setosa	50	0	0	0	0 / 50
Iris-versicolor	0	50	0	0	0 / 50
Iris-virginica	0	13	37	0.2600	13 / 50
Total	50	63	37	0.0867	13 / 150



3. H2O 분석 - GBM(1/2)

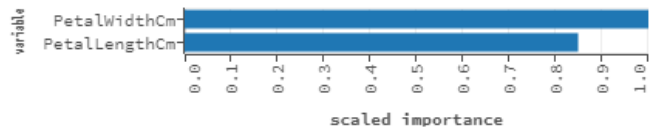
Sepal(꽃받침) 무시,
ntrees : 500으로 증가, 분류결과는 versicolor 1개로 Error 감소~~

SCORING HISTORY - LOGLOSS



변수 중요도
꽃잎폭(1.0) > 꽃잎 길이(0.83)

VARIABLE IMPORTANCES



TRAINING METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: P

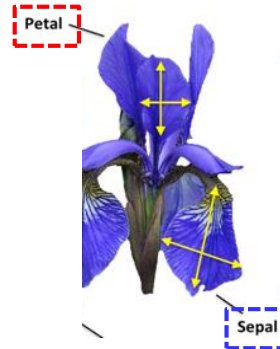
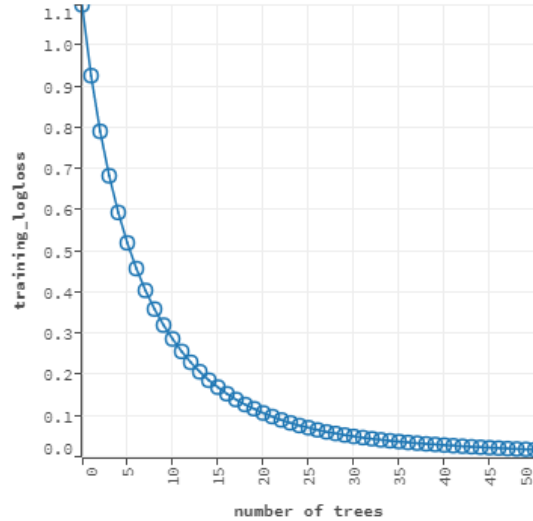
	Iris-setosa	Iris-versicolor	Iris-virginica	Error	Rate
Iris-setosa	50	0	0	0	0 / 50
Iris-versicolor	0	49	1	0.0200	1 / 50
Iris-virginica	0	0	50	0	0 / 50
Total	50	49	51	0.0067	1 / 150



3. H2O 분석 – GBM(2/2)

Sepal(꽃받침) 포함 시,
ntrees : 50, 분류결과는 Error 0개로 감소, Sepal 변수 포함이 결정적!

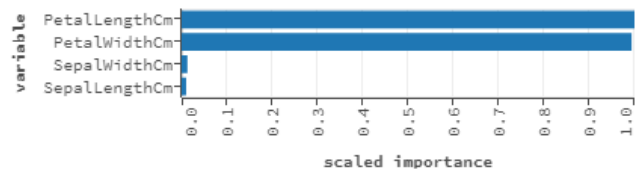
SCORING HISTORY - LOGLOSS



OUTPUT - VARIABLE IMPORTANCES

variable	relative_importance	scaled_importance	percentage
PetalLengthCm	227.9377	1.0	0.4960
PetalWidthCm	226.4932	0.9937	0.4928
SepalWidthCm	2.8897	0.0127	0.0063
SepalLengthCm	2.2480	0.0099	0.0049

VARIABLE IMPORTANCES



TRAINING METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	Iris-setosa	Iris-versicolor	Iris-virginica	Error	Rate
Iris-setosa	50	0	0	0	0 / 50
Iris-versicolor	0	50	0	0	0 / 50
Iris-virginica	0	0	50	0	0 / 50
Total	50	50	50	0	0 / 150



3. H2O 모델적용 - POJO 코드

생성된 Model을 POJO(Plain Old Java Object)코드 Export 기능
☞ 회사 내부 비즈니스 Process에 POJO코드 Porting 적용 가능성!

The screenshot shows the H2O Flow web interface in a browser. The address bar shows localhost:54321/flow/index.html. The main content area displays the 'PREVIEW POJO' section, which contains the following code:

```
/*
Licensed under the Apache License, Version 2.0
http://www.apache.org/licenses/LICENSE-2.0.html

AUTogenerated BY H2O at 2018-01-17T12:53:39.486+09:00
3.16.0.2

Standalone prediction code with sample test data for GBMModel named gbm_ccfcbe8c_02df_4a10_8316_71d80448d92a

How to download, compile and execute:
mkdir tmpdir
cd tmpdir
curl http://172.30.78.55:54321/3/h2o-genmodel.jar > h2o-genmodel.jar
curl http://172.30.78.55:54321/3/Models.java/gbm_ccfcbe8c_02df_4a10_8316_71d80448d92a > gbm_ccfcbe8c_02df_4a10_8316_71d80448d92a.java
javac -cp h2o-genmodel.jar -J-Xmx2g -J-XX:MaxPermSize=128m gbm_ccfcbe8c_02df_4a10_8316_71d80448d92a.java

(Note: Try java argument -XX:+PrintCompilation to show runtime JIT compiler behavior.)
*/
import java.util.Map;
import hex.genmodel.GenModel;
import hex.genmodel.annotations.ModelPojo;

@ModelPojo(name="gbm_ccfcbe8c_02df_4a10_8316_71d80448d92a", algorithm="gbm")
public class gbm_ccfcbe8c_02df_4a10_8316_71d80448d92a extends GenModel {
    public hex.ModelCategory getModelCategory() { return hex.ModelCategory.Multinomial; }

    public boolean isSupervised() { return true; }
    public int nfeatures() { return 4; }
    public int nclasses() { return 3; }

    // Names of columns used by model.
    public static final String[] NAMES = NamesHolder_gbm_ccfcbe8c_02df_4a10_8316_71d80448d92a.VALUES;
    // Number of output classes included in training data response column.
    public static final int NCLASSES = 3;

    // Column domains. The last array contains domain of response column.
    public static final String[][] DOMAINS = new String[][] {
        /* SepalLengthCm */ null,
        /* SepalWidthCm */ null,
        /* PetalLengthCm */ null,
    };
}
```

The right sidebar shows the 'HELP' section with a 'Quickstart Videos' button and a list of 'GENERAL' topics including 'Flow Web UI...', 'Importing Data', 'Building Models', 'Making Predictions', 'Using Flows', and 'Troubleshooting Flow'. The bottom status bar shows 'Connections: 0 H2O' and the system clock '오후 1:10 2018-01-17'.

목 차

1. Kaggle IRIS(붓꽃) Dataset 사례
2. H2o 분석
 - 딥러닝
 - GBM(Gradient Boosting Model)
3. Q&A



서강대학교
SOGANG UNIVERSITY

황 문기

산학협력교수, 혁신과 경쟁 연구센터, 서강 SSK 연구단
책임교수, 서강 머신러닝 파이낸스 클러스터

서울시 마포구 백범로 35, 서강대학교 남덕우경제관 803호 우) 04107
Mobile : 010-7685-3889 E-mail : mkhwang@sogang.ac.kr
moonkihw@naver.com