

Data Preprocessing


การเตรียมข้อมูล

Data Preprocessing เป็นขั้นตอนในการเตรียมข้อมูลเพื่อที่จะนำไปประมวลผลขั้นตอนต่อ โดยขั้นตอนที่ต้องทำขึ้นอยู่กับข้อมูลที่เราได้มาต้องทำอะไรกับมันบ้าง เช่น หากมีข้อมูลหายเราสามารถลบข้อมูลแถวนั้นได้ (ไม่แนะนำ สามารถดูการจัดการแบบอื่นได้) หรือข้อมูลของคลาส **target** มี ชนิดเป็น **string** เราสามารถทำ**encoding** เป็นตัวเลขได้

ขั้นตอนการเตรียมข้อมูล



1) เริ่มแรกให้ทำการนำชุดข้อมูลที่กำหนดให้มาจากเว็บ UCI DATASET

← → ↺ <https://archive.ics.uci.edu/ml/datasets.html>




UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems


Browse Through: 452 Data Sets

Default Task	Name	Data Types	Default Task	Attribute
Classification (334) Regression (92) Clustering (80) Other (55)	 Abalone	Multivariate	Classification	Categorical Integer, R
Attribute Type Categorical (38) Numerical (292) Mixed (55)	 Adult	Multivariate	Classification	Categorical Integer

2) เลือกข้อมูล iris

 Iris	Multivariate	Classification	Real	150	4	1988
--	--------------	----------------	------	-----	---	------

3) ให้เลือกดาวน์โหลดข้อมูลตรง Data Folder




UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Iris Data Set

Download: [Data Folder](#) [Data Set Description](#)

Abstract: Famous database; from Fisher, 1936



Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	2166367

4) เลือก iris.data

	Parent Directory	-
	Index	03-Dec-1996 04:01 105
	bezdekIris.data	14-Dec-1999 12:12 4.4K
	iris.data	08-Mar-1993 16:27 4.4K
	iris.names	11-Jul-2000 21:30 2.9K

5) เมื่อเข้ามาจะมีข้อมูลของ iris น้อยๆ สามารถคัดลอกข้อมูลทั้งหมดแล้วไปวางลง Notepad พร้อมทั้งบันทึกให้เป็นชื่อ iris.txt หรือ สามารถคลิกขวา แล้วกด save as พร้อมเปลี่ยนชื่อเป็น iris.txt (กรณีนี้อาจมีปัญหาลoadข้อมูลแต่ละแถวจะเข้ามาบรรทัดเดียวกันสำหรับบางคนมั้ง 555)

```
← → ↻ 🔒 https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
5.2,3.5,1.5,0.2,Iris-setosa
```

5.1) กรณีข้อมูลมีปัญหาหลังคลิกขวาแล้ว save as หรือ บันทึกเป็น

iris.data - Notepad

```
File Edit Format View Help
5.1,3.5,1.4,0.2,Iris-setosa4.9,3.0,1.4,0.2,Iris-setosa4.7,3.2,1.3,0.2,Iris-seto
4.6,3.1,1.5,0.2,Iris-setosa5.0,3.6,1.4,0.2,Iris-setosa5.4,3.9,1.7,0.4,Iris-seto
4.6,3.4,1.4,0.3,Iris-setosa5.0,3.4,1.5,0.2,Iris-setosa4.4,2.9,1.4,0.2,Iris-seto
4.9,3.1,1.5,0.1,Iris-setosa5.4,3.7,1.5,0.2,Iris-setosa4.8,3.4,1.6,0.2,Iris-seto
4.8,3.0,1.4,0.1,Iris-setosa4.3,3.0,1.1,0.1,Iris-setosa5.8,4.0,1.2,0.2,Iris-seto
5.7,4.4,1.5,0.4,Iris-setosa5.4,3.9,1.3,0.4,Iris-setosa5.1,3.5,1.4,0.3,Iris-seto
5.7,3.8,1.7,0.3,Iris-setosa5.1,3.8,1.5,0.3,Iris-setosa5.4,3.4,1.7,0.2,Iris-seto
5.1,3.7,1.5,0.4,Iris-setosa4.6,3.6,1.0,0.2,Iris-setosa5.1,3.3,1.7,0.5,Iris-seto
4.8,3.4,1.9,0.2,Iris-setosa5.0,3.0,1.6,0.2,Iris-setosa5.0,3.4,1.6,0.4,Iris-seto
5.2,3.5,1.5,0.2,Iris-setosa5.2,3.4,1.4,0.2,Iris-setosa4.7,3.2,1.6,0.2,Iris-seto
4.8,3.1,1.6,0.2,Iris-setosa5.4,3.4,1.5,0.4,Iris-setosa5.2,4.1,1.5,0.1,Iris-seto
5.5,4.2,1.4,0.2,Iris-setosa4.9,3.1,1.5,0.1,Iris-setosa5.0,3.2,1.2,0.2,Iris-seto
```

แก้ข้อมูลสลับเลย... ขอแนะนำให้คัดลอกข้อความต้นทางมาดีกว่า

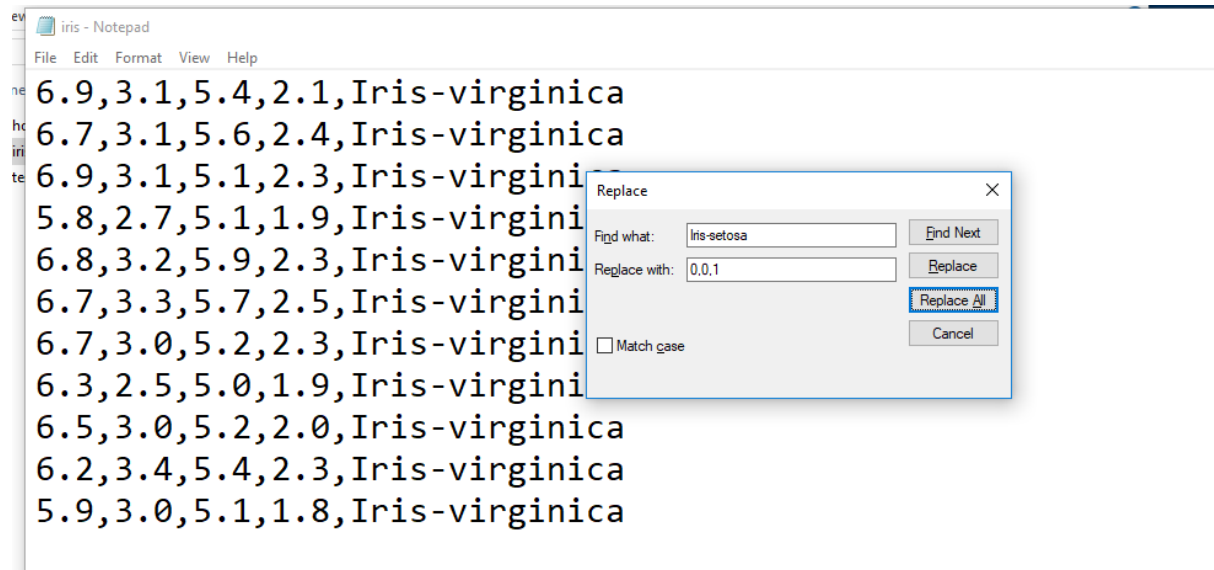
6) หลังจากทำการเซฟข้อมูลเสร็จแล้วเราเห็นได้ว่ามีข้อมูล target เราเป็น string เราจึงทำ Encoding ให้มันเป็นตัวเลข
ซะ โดยใช้วิธี One Hot (google ช่วยท่านได้)

ให้ Iris-setosa -> 0,0,1

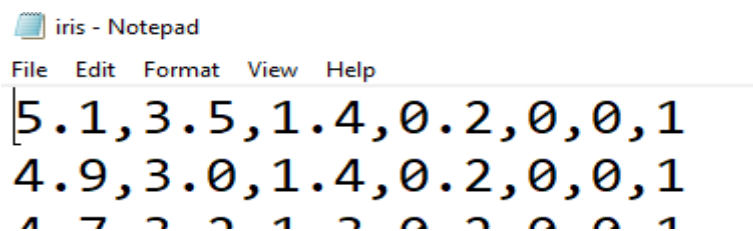
Iris-versicolor -> 0,1,0

Iris-virginica -> 1,0,0

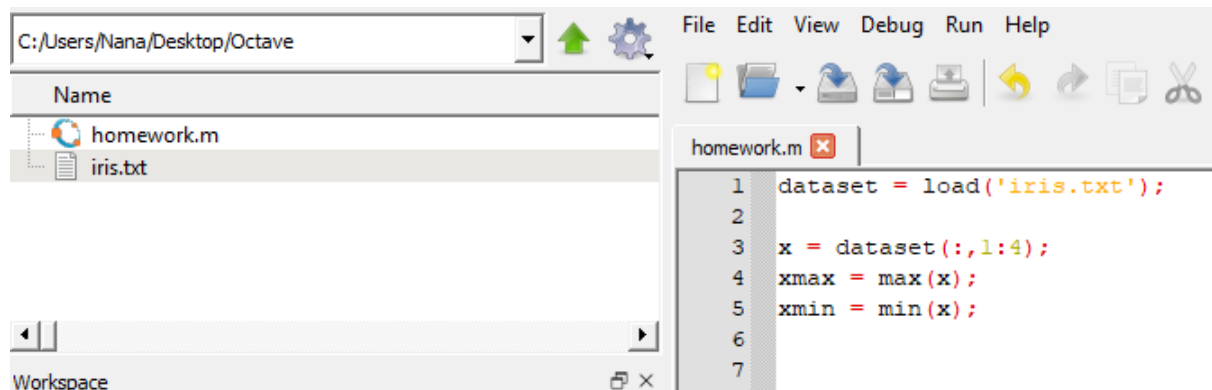
ทำได้ที่ Notepad เลย กด Ctrl + H และทำการเปลี่ยนข้อมูล



6.1) ข้อมูลใหม่ที่ได้มี 150 แถว 7 คอลัมน์



7) เมื่อเตรียมข้อมูลเสร็จแล้วให้เราเข้า octave (หรือ matlab) ให้ไฟล์ iris กับไฟล์โค้ดเราอยู่ด้วยกัน



8) ยังคงอยู่กระบวนการจัดการข้อมูลที่ได้มา ตอนนี้จะพูดถึงกระบวนการ **feature scaling** หรือการปรับระดับค่าของตัว **feature** นั้นเอง (feature ในนี้คือค่า input ที่ไม่ใช่ตัว **target**) เราจะเรียกการทำนี้ว่า **normalization** หรือ **normalize** ข้อมูล โดยปรับสเกลของมันให้อยู่ระหว่าง 0 กับ 1 ซึ่งจะส่งผลให้การประมวลผลเร็วขึ้น

ขั้นตอนการทำ

- 1) โหลดข้อมูลไฟล์ iris เข้ามา
- 2) ใช้สูตรในการทำ **normalize** ใส่ตัวแปร **Xnorm** โดยการนำค่า **dataset** จากคอลัมน์ที่ 1 ถึง 4
- 3) **T** เป็นตัวแปรสำหรับเก็บค่า **Target** (คอลัมน์ 5 ถึง 7)

```
1 dataset = load('iris.txt');
2
3 x = dataset(:,1:4);
4 xmax = max(x);
5 xmin = min(x);
6 Xnorm = (x-xmin) ./ (xmax-xmin);
7
8 T = dataset(:,5:end);
9
```

หมายเหตุ จะใช้ตัวแปร **Xnorm** แทน **X** ในหนังสือหน้า 117 (หนังสือโครงข่ายประสาทเทียม(Artificial Neural Network) เขียนโดย ผศ.ดร. สิริภัทร เชี่ยวชาญวัฒนา)

8.1) เปรียบเทียบข้อมูลที่ได้จากข้อมูลเดิมและข้อมูลที่ทำกร **normalize** แล้ว

1	2	3	4
5.1000	3.5000	1.4000	0.2000
4.9000	3	1.4000	0.2000
4.7000	3.2000	1.3000	0.2000
4.6000	3.1000	1.5000	0.2000
5	3.6000	1.4000	0.2000
5.4000	3.9000	1.7000	0.4000

ภาพ ก

1	2	3	4
0.2222	0.6250	0.0678	0.0417
0.1667	0.4167	0.0678	0.0417
0.1111	0.5000	0.0508	0.0417
0.0833	0.4583	0.0847	0.0417
0.1944	0.6667	0.0678	0.0417
0.3056	0.7917	0.1186	0.1250

ภาพ ข

จะเห็นว่าภาพ ข เป็นข้อมูลจากภาพ ก ที่ถูกปรับให้อยู่ระหว่าง 0 – 1

9) ทำการแบ่งข้อมูลให้เป็น **train** และ **test** โดยการแบ่งส่วนที่แนะนำให้เป็น 70:30 นั่นคือ **train** จะมีข้อมูล

150*70/100 ได้ 105 แต่หนังสือเอา 100 เวกก็จะเอา 100 ส่วน **test** จะมีข้อมูล 150*30/100 ได้ 45 หนังสือเอา 50

```
sz = size(dataset,1);
I = randperm(sz);
xTrain = Xnorm (I(1:100),:);
xTest = Xnorm (I(101:end),:);
tTrain = T(I(1:100),:);
tTest = T(I(101:end),:);
```

Tip การแบ่งข้อมูลมีการแบ่งแบบ 50:50 , 70:30 , 90:10 โดยการแบ่งข้อมูล 70:30 เป็นที่นิยมใช้ 50:50 ไม่แนะนำ เพราะการ **Train** ควรมีข้อมูลมากกว่า