

2 lakh \rightarrow 200 lakh

	f_1	f_2	...	f_{100}	Y/N
D_1					
D_2					
\vdots					
D_{20000}					

\leftarrow f_1 f_2 f_{100}
 Some features not important
 \Rightarrow degrades the regression

How do we find the subset that maximises

f_1 f_2 ... f_{100}

[M]

(1) take literally ALL SUBSETS TRAIN THEM ALL
 (No Fucking Way) FIND BEST ACCURACY SUBSET REEEEEEEEEEE

(2) greedy approach: Start with no features and keep adding the "best" one of remaining

M_1 f_1
 M_2 f_2
 \vdots \vdots
 M_{100} f_{100}

$\rightarrow f_{35} \rightarrow$ gives max performance

$\hookrightarrow M' f_{35}$

f_1
 f_2
 \vdots
 f_{100}

\rightarrow take next best feature

add it to M' etc...

$M_{\text{final}} = \{f_{35}, f_{42}, \dots\}$

M' M'' M''' ... M^{100}
 Find the best one out of this

or just stop it
 @ some MSE

problem: having selected f_{35} , f_{10} is best
(Basically problem of greedy approach)

greedy backward

$$M_{100} \begin{cases} M_{99} = M_{100} - f_1 - \\ M_{99} = M_{100} - f_5 - \\ M_{99} = M_{100} - f_{100} - \end{cases} \begin{matrix} \text{take best one} \\ \text{Eg: } M_{99} = M_{100} - f_{67} \end{matrix}$$

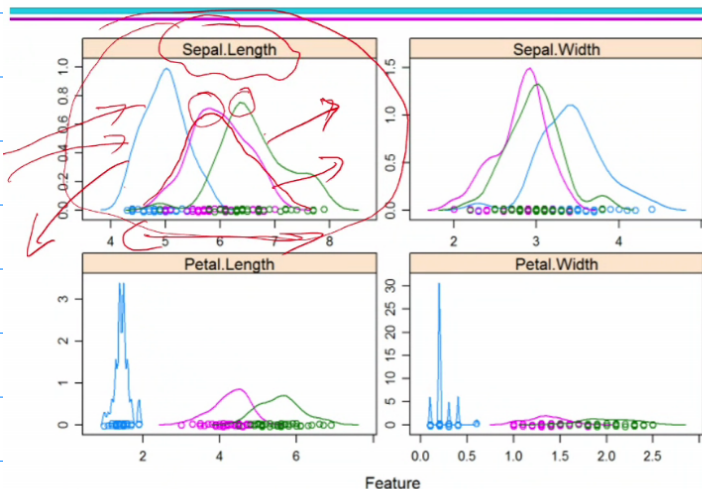
$M_{99} \left\{ \begin{matrix} \text{Build back} \\ \rightarrow \text{take best} \end{matrix} \right.$

$M_{100} \quad M_{99} \dots M_1 \} 100 \text{ and take best}$

→ While this isn't preprocessing in the strict sense thus no way out in this algorithm.

(We are involving the model itself to preprocess data)

Univariate Analysis using PDF



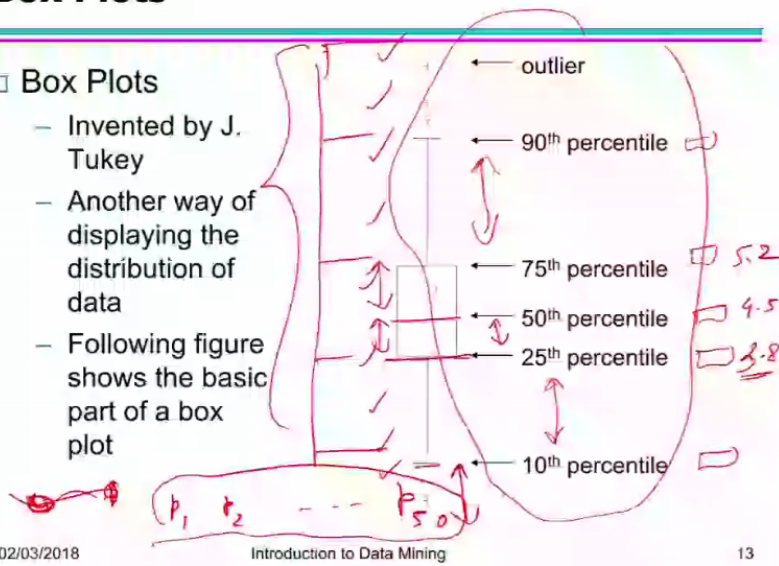
→ Let's say that we want to classify flowers.

→ We can plot discrete distribution
or continuous
(By normalizing)

Box Plots

Box Plots

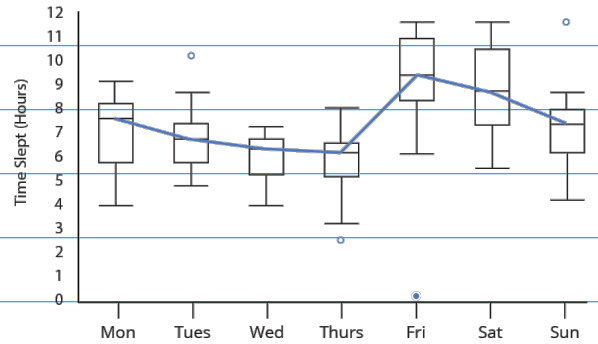
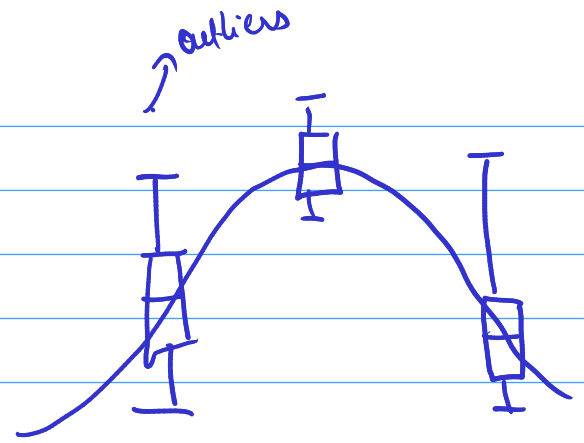
- Invented by J. Tukey
- Another way of displaying the distribution of data
- Following figure shows the basic part of a box plot



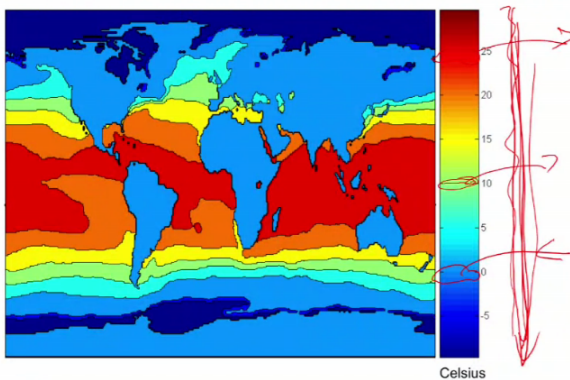
02/03/2018

Introduction to Data Mining

13

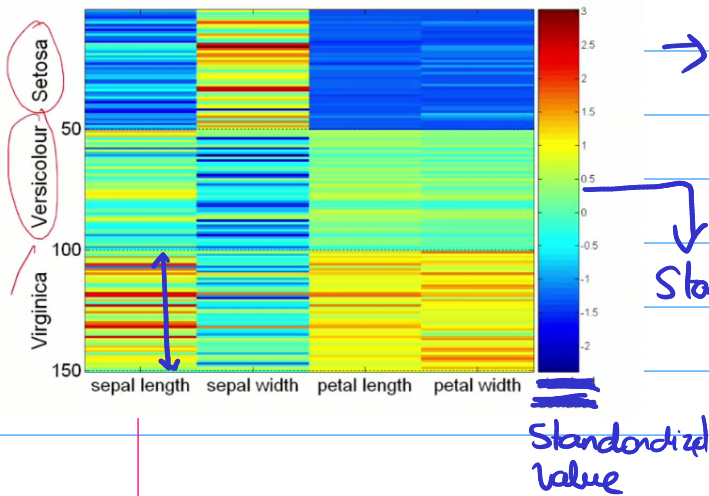


Contour Plot Example: SST Dec, 1998



some kind of average temperature

Or we can even take color corresponding to actual C here (it's smooth)



→ take samples and standardize

Standardized values

we can say that virginica petal length on average has a standard deviation of '1'