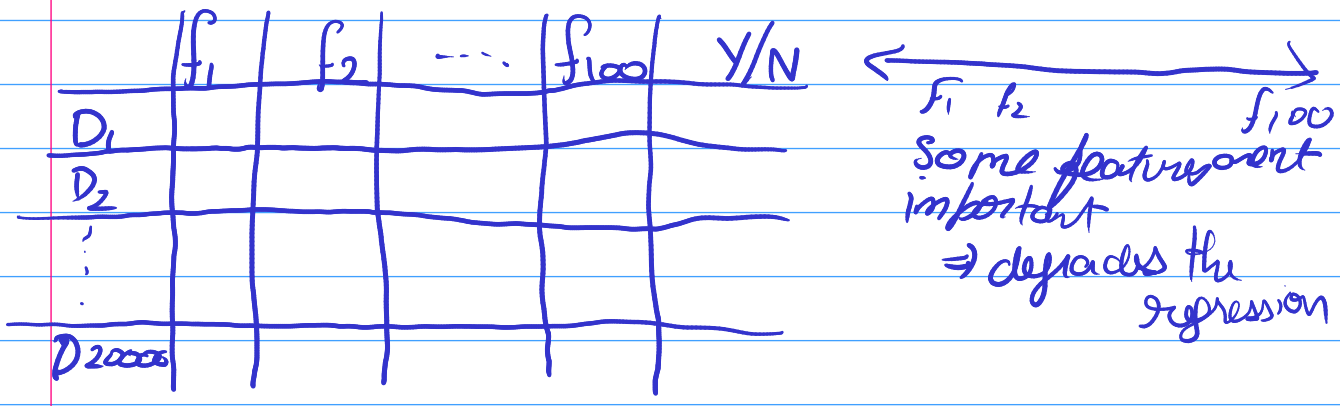
2 lakh → 2000 lakh



How do we find the subset that maximises

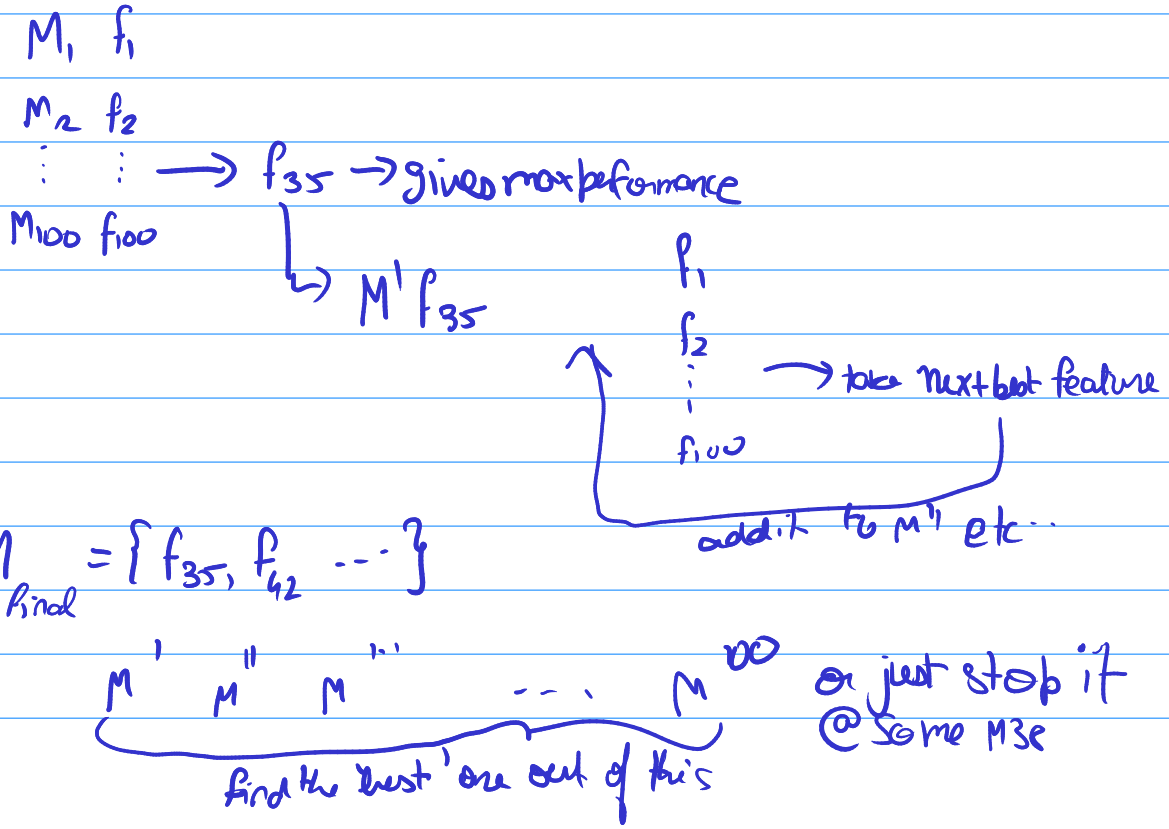
f_1 f_2 ... f_{100}

M

(1) take literally ALL SUBSETS TRAIN THEM ALL FIND BEST ACCURACY SUBSET REEEEEEEEE

(No fucking way)

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



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

greedy backward

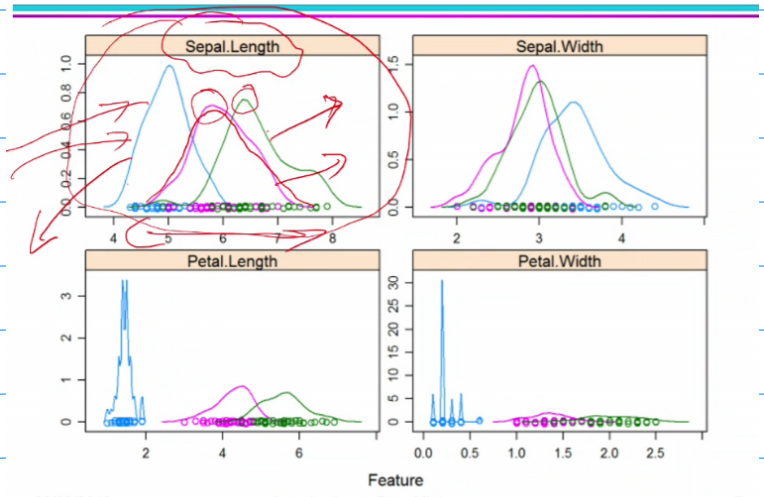
M_{100} { $M_{99} = M_{100} - f_1$ –
$M_{99} = M_{100} - f_5$ – take best one
M_{99} eg: $M_{99} = M_{100} - f_{67}$
$= M_{100} - f_{00}$ –

M_{99} { Build all → take best

M_{100} M_{99} — M_1 } 100 and take best

→ While there isn't preprocessing in the strict sense there no way out in this algorithm.
(We are making the model itself to preprocess data)

Univariate Analysis using PDF

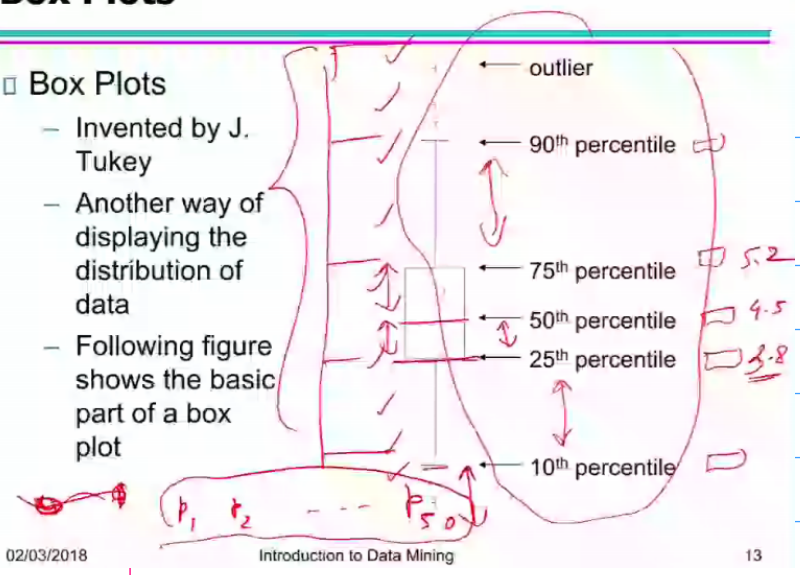


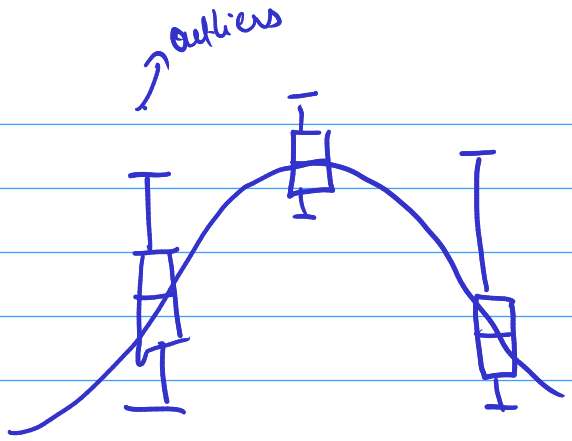
→ Let's say that we want to classify flowers.

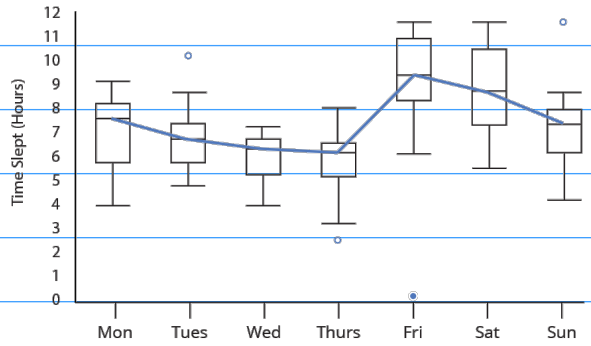
→ We can plot discrete distribution or continuous (By normalizing)

Box Plots

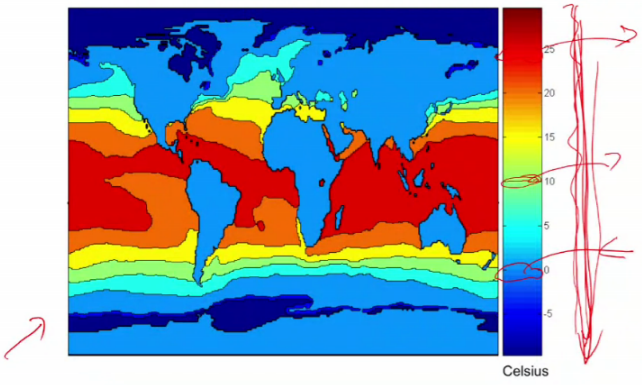
- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot





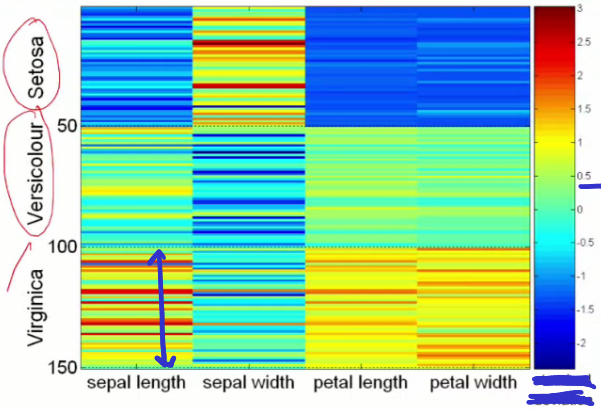


Contour Plot Example: SST Dec, 1998



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'