DASC 4113 Machine Learning

Ukash Nakarmi

Lecture 1

Course Logistic and Syllabus

Meeting times: Tu/Th 11:00 AM – 12:15PM

Location: JBHT 0266

Delivery Mode: Face2Face

Useful web links for Course Logistics

Blackboard: https://learn.uark.edu

Course One Drive Link

DASC4113_FALL2023

https://uark-my.sharepoint.com/:f:/g/personal/unakarmi_uark_edu/Eoc1OLH5URlGmwtSBwoFUG0BEhKh7xM4CWBxywCmlWcYwg?e=hF4yHg

Office Hours

Time: Tuesdays, 2:PM – 3:00 PM

Location: JBHT 525

Or by Appointment Virtual/In-person

Correspondence and Email

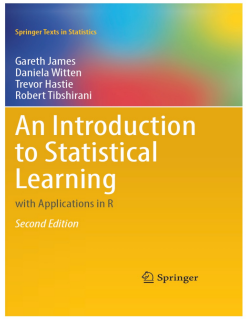
Email: unakarmi@uark.edu. (Please include "DASC 4113" in subject line)

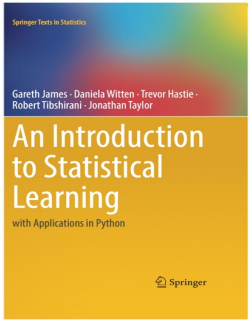
Tentative Course Topics:

Regression, Logistic regression and stochastic gradient descents	2 weeks
Ensemble methods	1 week
Support vector machines and kernel methods	2 weeks
Bayesian inference	1 weeks
Neural networks and deep learning	4 weeks
Reinforcement learning	2 weeks
Learning theory	1 week
Other Emerging Topics:	2 weeks

Books and Resources:

An Introduction to Statistical Learning, (Python Version) by G James, et. al. (2023).
Available online:
https://www.statlearning.com/
https://hastie.su.domains/ISLP/ISLP_website.pdf





- Understanding Machine Learning: From Theory to Algorithms, by Shai Shalev-Shwartz and Shai Ben-David (2014). Available online: https://www.cse.huji.ac.il/~shais/UnderstandingMachineLearning/
- Pattern Recognition and Machine Learning, by C. Bishop. Available Online: https://www.microsoft.com/en-us/research/people/cmbishop/prml-book
- Dive into Deep Learning, by Aston Zhang and Zachary C. Lipton and Mu Li and Alexander J. Smola (2020). Available online: https://d2l.ai/

Assignments and Grading:

Types of Assignments

i) Pre-Class Questionnaire **5 Points**

Done individually

Assignments and Grading:

Types of Assignments

ii) Data Survey and Visualization **10 Points**

Done individually

Assigned sometime in the second week of the semester

Assignments and Grading:

Types of Assignments

iii) Reading/Scribe/Code/Present **15 points**

Done in Group

Every week one of the group will be assigned Reading/Scribe/Code Assignments

Most Thursday Classes will be used for Student Presentations
Details to come later!

Assignments and Grading:

Types of Assignments

iv Peer Grading 5 points

Done in group.

Each group except the presenting group will Peer grade Presenting Group's Reading/Scribe/Code Assignment.

Assignments and Grading:

5 Types of Assignments

v) Quiz: **20 points**

Done individually

One random quiz and another quiz sometime in Early November.

Assignments and Grading:

Types of Assignments

vi) Mini Projects **2*10 = 20 points**

Done in Group

(Some time in Mid-Late November)

Assignments and Grading:

5 Types of Assignments

Major project: **25 points**

1) Literature Survey, Project Selection and Proposal - **5 points**
*(Will be assigned, **Early September**, Due: ~**1.5 weeks**)*

2) 1st Progress Report and Presentation - **10 points**, (*Most likely around early Mid October)*

4) Final Report and Poster – **10 points.** (*Most likely around early December before the Final Class week*

Grading Scale:

A scale like the following will be used to determine final course grades. However, **it is approximate and subject to change**.

i) A: over 90%
ii) B: 80-89%
iii) C: 70-79%
iv) D: 60-69%
v) F: below 60%

UNIVERSITY OF ARKANSAS

How to be successful in this class?

Success: Not just grades in the course!

a) Start Early on Assignments!
b) Do not be discouraged if you don't understand something on a first take!
c) Course expects significant chunk of self-study!

Some Recommendation:

a) Try to stick to one reliable resource.
b) Learn by doing! Projects, Projects, Projects!
c) Take Reading and other ungraded assignments sincerely.
d) Discussion, Discussion!
e) Don't Hesitate to come during Office Hours, even if it is for just some random chats, questions about your projects and other interests related to ML!

Lecture 1

Motivation

1. We will look at 3 different data sets .

1. Study several information available from the data.

1. Analyze the information to understand several interesting problems we could answer from the data.

Some Data Sets

1. Wage

D

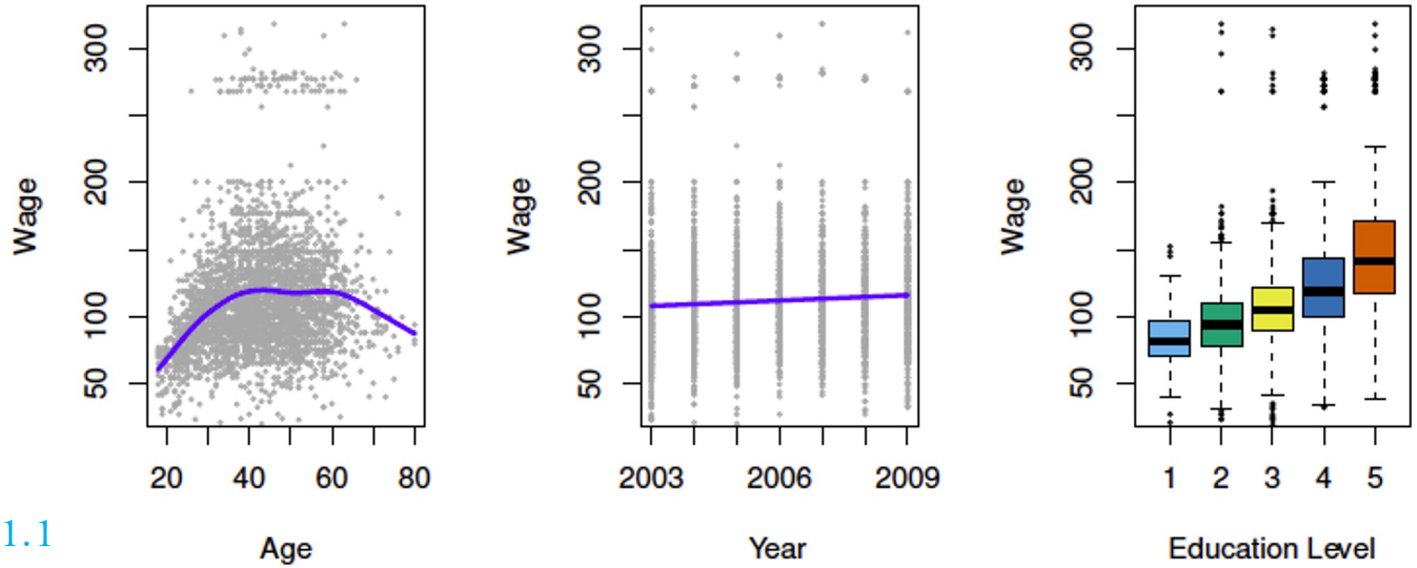


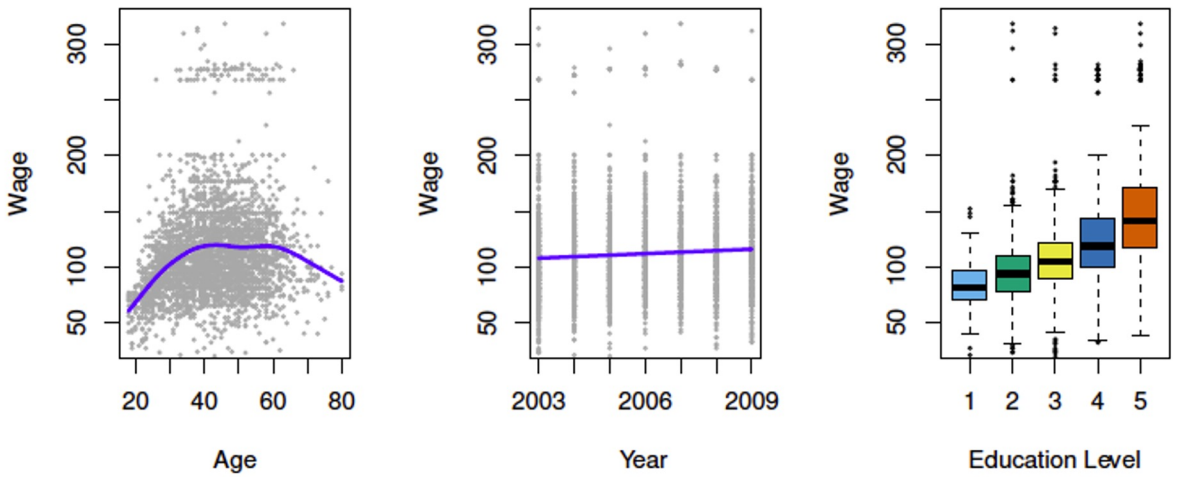
Fig 1.1

1. Income Survey Information for Men from the Central Atlantic Region of the US.

1. Shows relation between the Wage and several other parameters. (*Wage as function of features*)

1. Wage Data

Fig 1.1



Q. What can we say about the relation between:

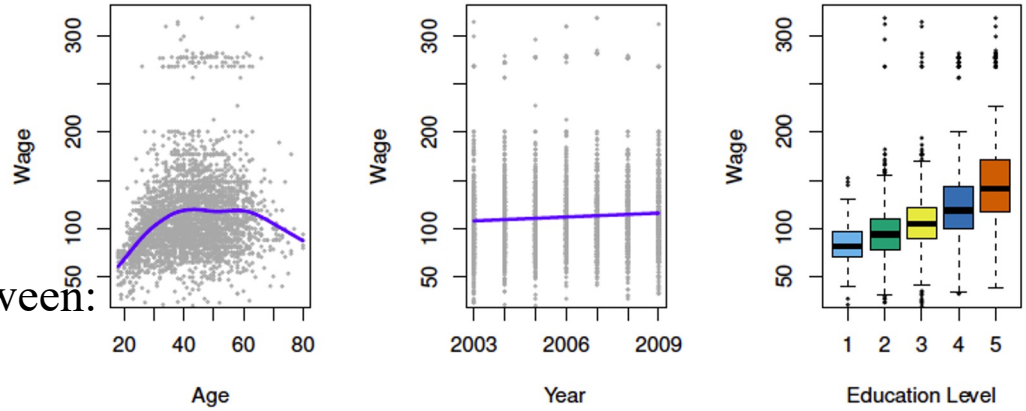
a) *Age and Wage?*

a) *Year and Wage?*

a) *Education Level and Wage?*

1. Wage Data

Fig 1.1



Q. What can we say about the relation between:

a) Age and Wage?

On average, wage increases age until 60 years and starts to decrease slowly.

a) Year and Wage?

There is a slow but steady increase in wage between 2003 to 2009.

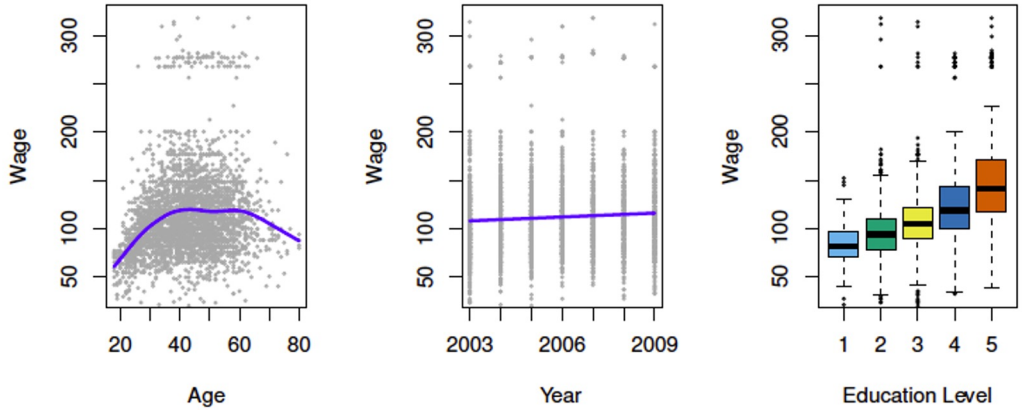
a) Education Level and Wage?

On average the wage increases with education level.

In general, we tried to understand the relation between the feature and the target variable. (Increase-Decrease, Linear-nonlinear …..)

1. Wage Data

Fig 1.1



Application of Data to Answer questions

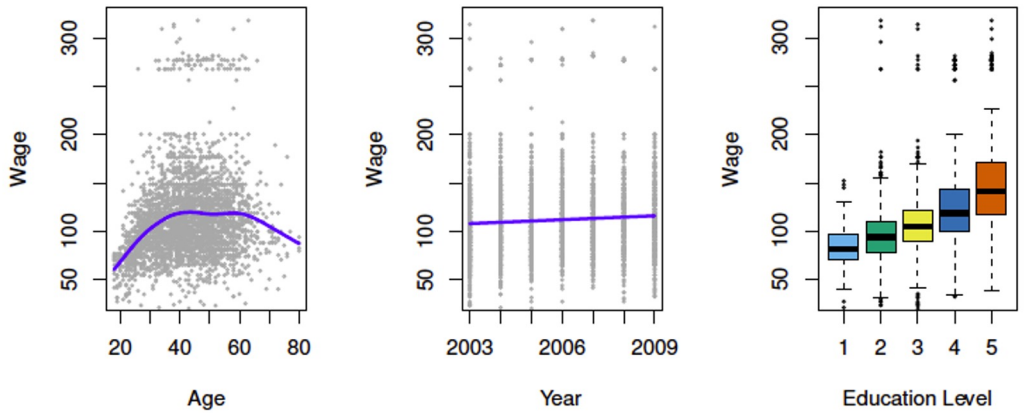
Q. Given a man's age can we predict his income?:

Q. If we make such prediction of a man's wage based on his age, how reliable will be our prediction?

Q. Given all these three data sets, how can we best predict a man's wage?

1. Wage Data

Fig 1.1



Q. Given a man's age can we predict his income?

Yes!

Q. If we make such prediction of a man's wage based on his age, how reliable will be our prediction?

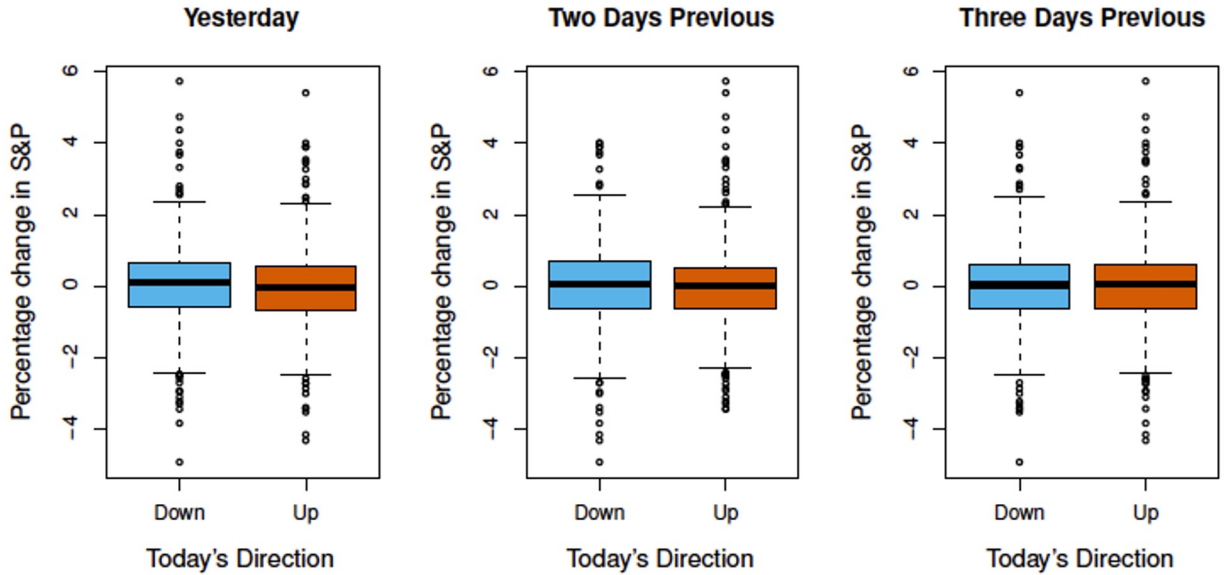
Probably not! Why not?

Q. Given all these three data sets, how can we best predict a man's age?

Combining the effect (contribution) all features, age, year and education level.

2. Stock Market Data (Smarket Data)

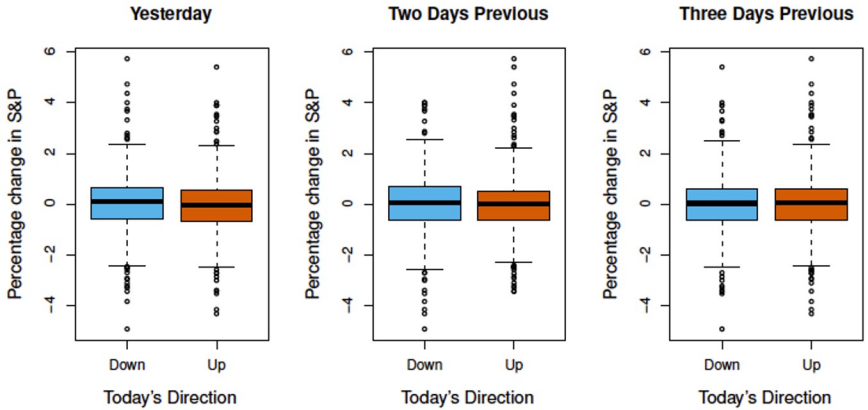
Fig 1.2



- Two Box plots of previous days percentage change in S&P index.
- One for 602 days for which the market decreased (DOWN) in the subsequent day. Another, for 648 days for which the market increased (UP) on the subsequent day.
- Note, the two plots look very identical.

2. Stock Market Data (Smarket Data)

Fig 1.2

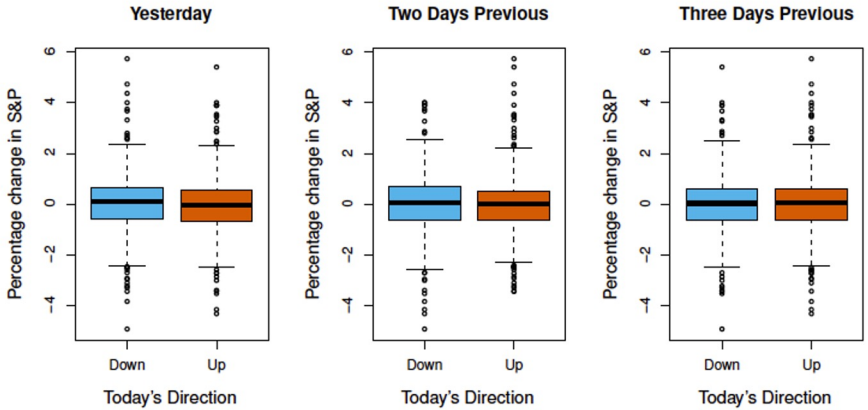


Q. Looking at the two Box Plots (Up and Down), can we say the market would go UP or down today?

Q. If Yes, how did you infer that? In not, why we cannot ?

2. Stock Market Data (Smarket Data)

Fig 1.2



1. Looking at the two Box Plots (Up and Down), can we say the market would go UP or down today?

No, the two box plots look very identical. Meaning for both box plots their mean value for percentage age is almost same (0).

2. If Yes, how did you infer that? In not, why we cannot ?

No, we cannot. This makes sense. It is obvious the market value depends not only on simple percentage change relation.

2. Stock Market Data (Smarket Data)

Fig 1.2

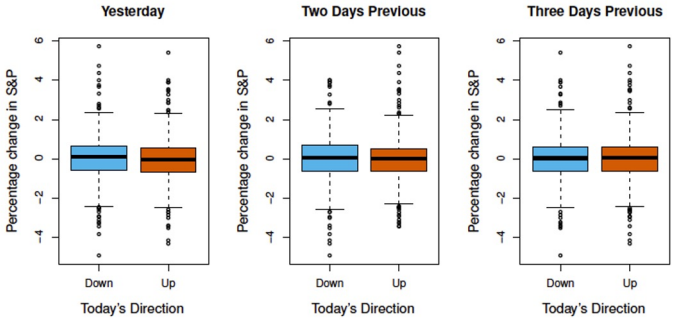
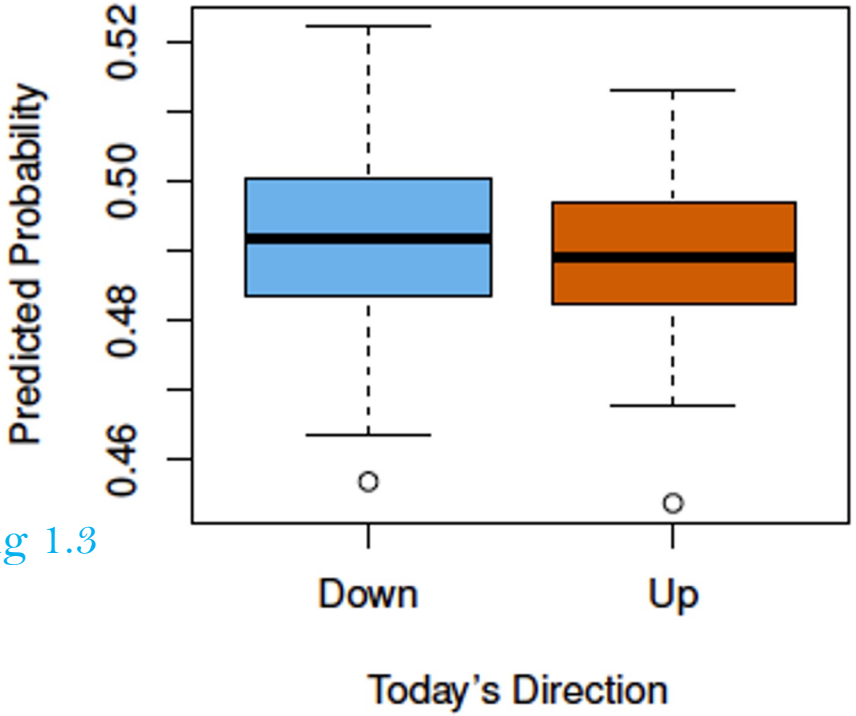
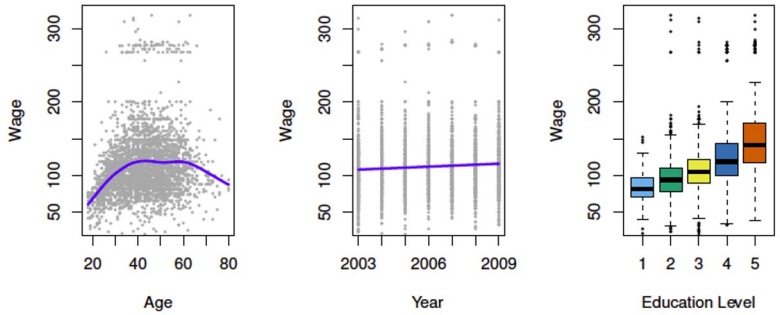


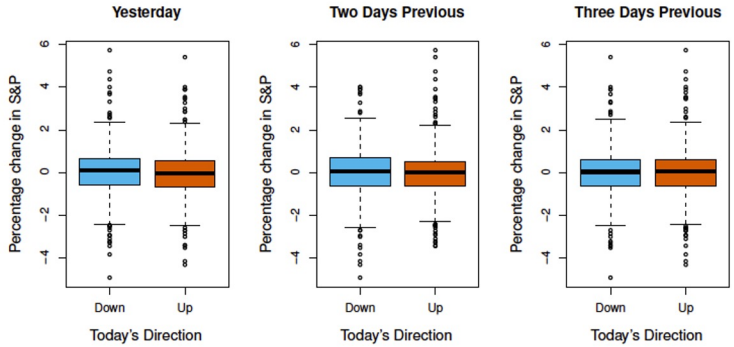
Fig 1.3



- Smarket data fitted using Quadratic Discriminant Analysis
  (We will talk about this in chapter 4).
- We can see the the probability the Market Goes DOWN is greater than the probability the market goes UP.
- We can make a prediction; the market goes UP.

Compare Wage Data and Smarket Data



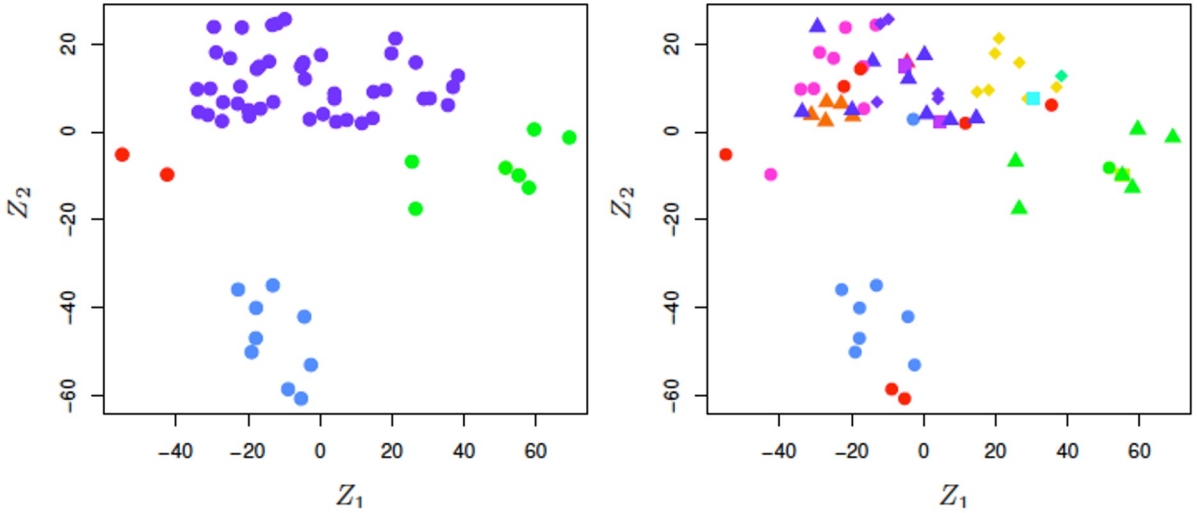


1. Given some features what is the the Wage of the Man?
   - Continuous or quantitative prediction
   - Regression Problem

1. Can we predict if the Market Goes UP or DOWN?
   - Categorical or qualitative prediction
   - Classification Problem

UNIVERSITY OF ARKANSAS

3. Gene Expression Data

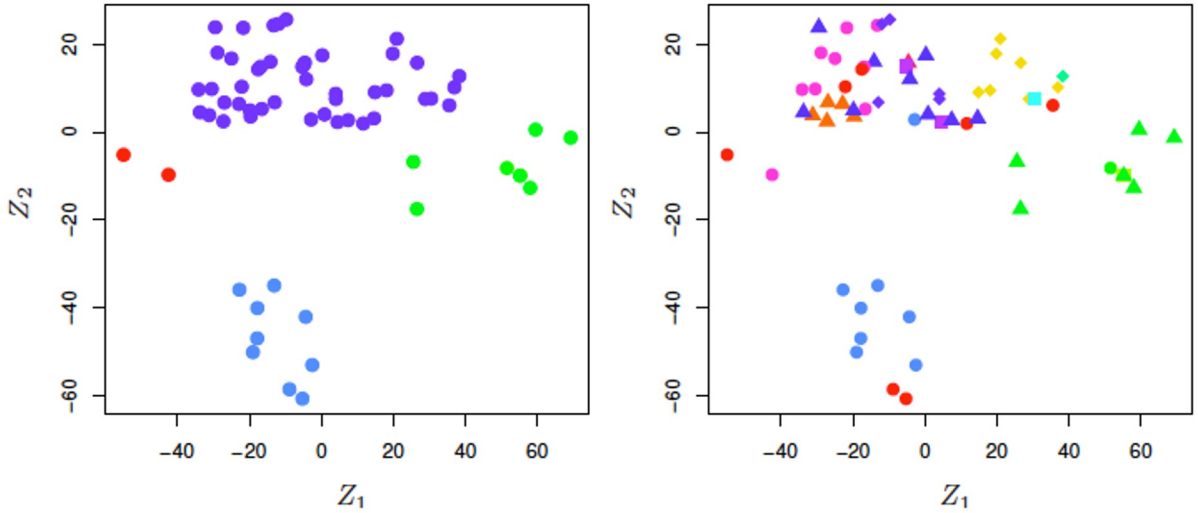


Representation of NCI60 gene expression data set in a 2-D space, Z1 and Z2. The data set consisted of 6300 gene expression measurement of 64 cancer cell lines.

The gene expression is a high dimensional data (6300 features). It is hard to visualize this data to understand relation between gene expression and caner cell lines. (*Is there any groups/clustering among cell lines based on their gene expression?*)

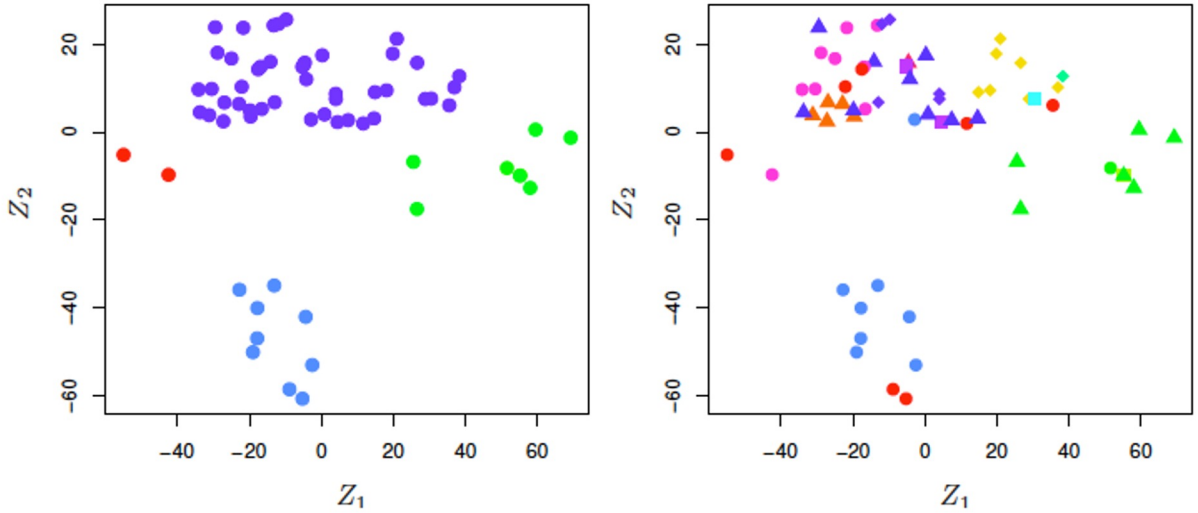
Clustering Problem !

3. Gene Expression Data



Left: 6300 gene expression of 64 cell lines. Each point corresponds to one cell line. We see there are roughly 4 clusters. Right: Same data but each type of 14 cancer is plotted using different color and symbol.

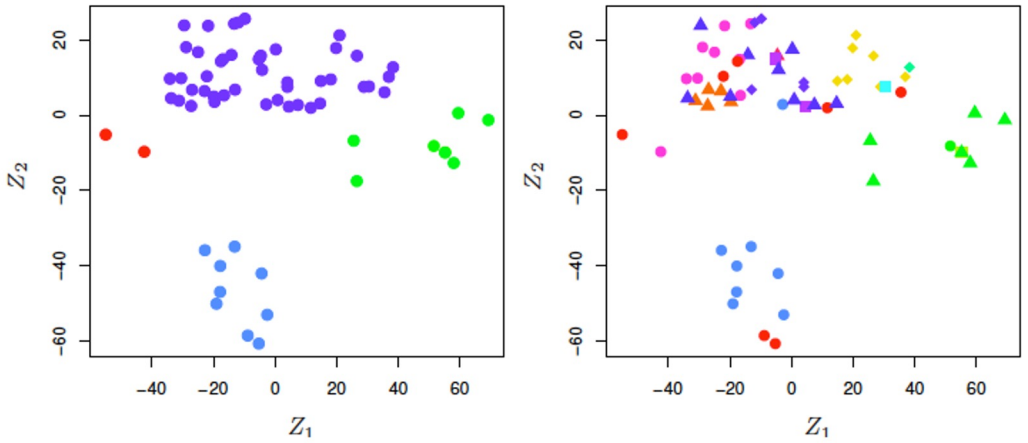
3. Gene Expression Data



Q. How is this data different from Wage and Smarket Data?

Q. How is the task we are interested in this data different from Wage and Smarket Data?

3. Gene Expression Data



Q. How is this data different from Wage and Smarket Data?

Wage and Smarket Data have both input (features) and Output (targets). *(Supervised Learning)* *This data only has observed input variables.* *(Unsupervised Learning)*

Q. How is the task we are interested in this data different from Wage and Smarket Data?

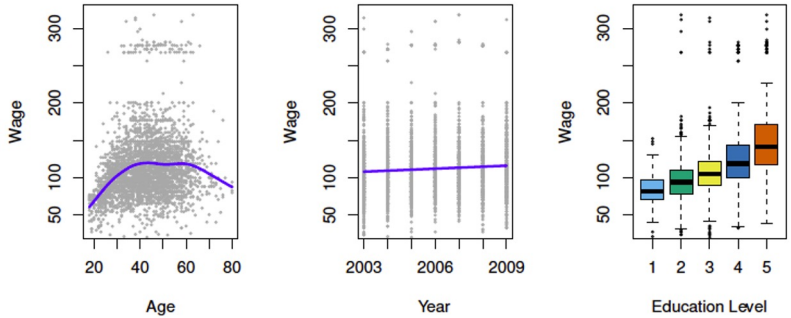
Not exactly a prediction problem. We want to understand the relation between the features or observed variables. **Clustering Problem.**

However, once we form the cluster, we might pose a new prediction problem, given a gene expression, can we predict which

Take Away:

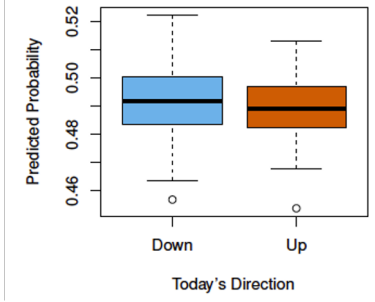
Regression

Continuous, Quantitative, Supervised



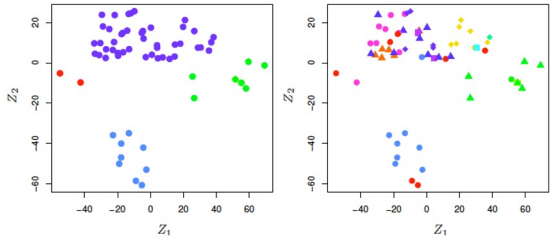
Classification

Categorical, Qualitative, Supervised



Clustering

No input-output pair, Unsupervised



Reading Assignment for Next Class

Introduction to Statistical Learning

Chapter 2, Page 15- 42

This reading assignment is ungraded.

Assignment 1

Pre-Class Questionnaire **5 Points**

Will be Assigned Tonight!