Statistical Learning– MATH 6333
Set 1 (Introduction to Statistical Learning)

Tamer Oraby

UTRGV

tamer.oraby@utrgv.edu

* Last updated August 26, 2021

Statistical Learning

- SL provides a set of methods/tools for identifying patterns in large databases.
- Those patterns could be then used in prediction or decision making.
- Machine learning is a multi-disciplinary (statistics, mathematics, computer science, etc.) provides algorithms that are designed to learns from data.

Statistical Learning

- SL provides a set of methods/tools for identifying patterns in large databases.
- Those patterns could be then used in prediction or decision making.
- Machine learning is a multi-disciplinary (statistics, mathematics, computer science, etc.) provides algorithms that are designed to learns from data.

Statistical Learning

- SL provides a set of methods/tools for identifying patterns in large databases.
- Those patterns could be then used in prediction or decision making.
- Machine learning is a multi-disciplinary (statistics, mathematics, computer science, etc.) provides algorithms that are designed to learns from data.

Statistical Learning

The work in SL involves:

- Data curation, like for training data
- Data exploration
- Visualization and presentation
- Modeling relationships in the data; like, generative and discriminative
- Conclusion and decisions

Statistical Learning

The work in SL involves:

- Data curation, like for training data
- Data exploration
- Visualization and presentation
- Modeling relationships in the data; like, generative and discriminative
- Conclusion and decisions

Statistical Learning

The work in SL involves:

- Data curation, like for training data
- Data exploration
- Visualization and presentation
- Modeling relationships in the data; like, generative and discriminative
- Conclusion and decisions

Statistical Learning

The work in SL involves:

- Data curation, like for training data
- Data exploration
- Visualization and presentation
- Modeling relationships in the data; like, generative and discriminative
- Conclusion and decisions

Statistical Learning

The work in SL involves:

- Data curation, like for training data
- Data exploration
- Visualization and presentation
- Modeling relationships in the data; like, generative and discriminative
- Conclusion and decisions

Statistical Learning

A general modeling form for the relationship between

Y: output or response (quantitative/continuous or categorical)

$X = (X_1, X_2, \ldots, X_p)$: input or features/attributes/covariates /predictors

is

$$Y = f(X) + \epsilon$$

where ϵ is the random error term.

Statistical Learning

Observed (training) data: $(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i)$ for object i and $i = 1, 2, \ldots, n$.

Then,

$$Y_i = f(X_{i1}, X_{i2}, \ldots, X_{ip}) + \epsilon_i$$

for $i = 1, 2, \ldots, n$, and the errors ϵ_i are independent identically distributed random variables (iidrv) with mean 0 and are independent of the X's.

One goal in SL: To find the prediction $f(X_*)$ for some future or testing input $X_* = (X_{*1}, X_{*2}, \ldots, X_{*p})$.

Statistical Learning

Observed (training) data: $(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i)$ for object i and $i = 1, 2, \ldots, n$.

Then,

$$Y_i = f(X_{i1}, X_{i2}, \ldots, X_{ip}) + \epsilon_i$$

for $i = 1, 2, \ldots, n$, and the errors ϵ_i are independent identically distributed random variables (iidrv) with mean 0 and are independent of the X's.

One goal in SL: To find the prediction $f(X_*)$ for some future or testing input $X_* = (X_{*1}, X_{*2}, \ldots, X_{*p})$.

Statistical Learning

Observed (training) data: $(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i)$ for object i and $i = 1, 2, \ldots, n$.

Then,

$$Y_i = f(X_{i1}, X_{i2}, \ldots, X_{ip}) + \epsilon_i$$

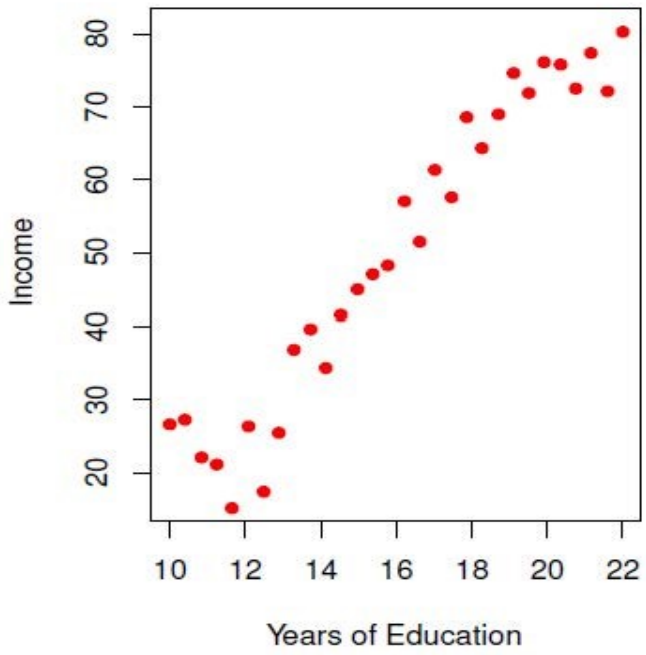
for $i = 1, 2, \ldots, n$, and the errors ϵ_i are independent identically distributed random variables (iidrv) with mean 0 and are independent of the X's.

One goal in SL: To find the prediction $f(X_*)$ for some future or testing input $X_* = (X_{*1}, X_{*2}, \ldots, X_{*p})$.

Statistical Learning

Example

Let X_1: years of education & Y: income.



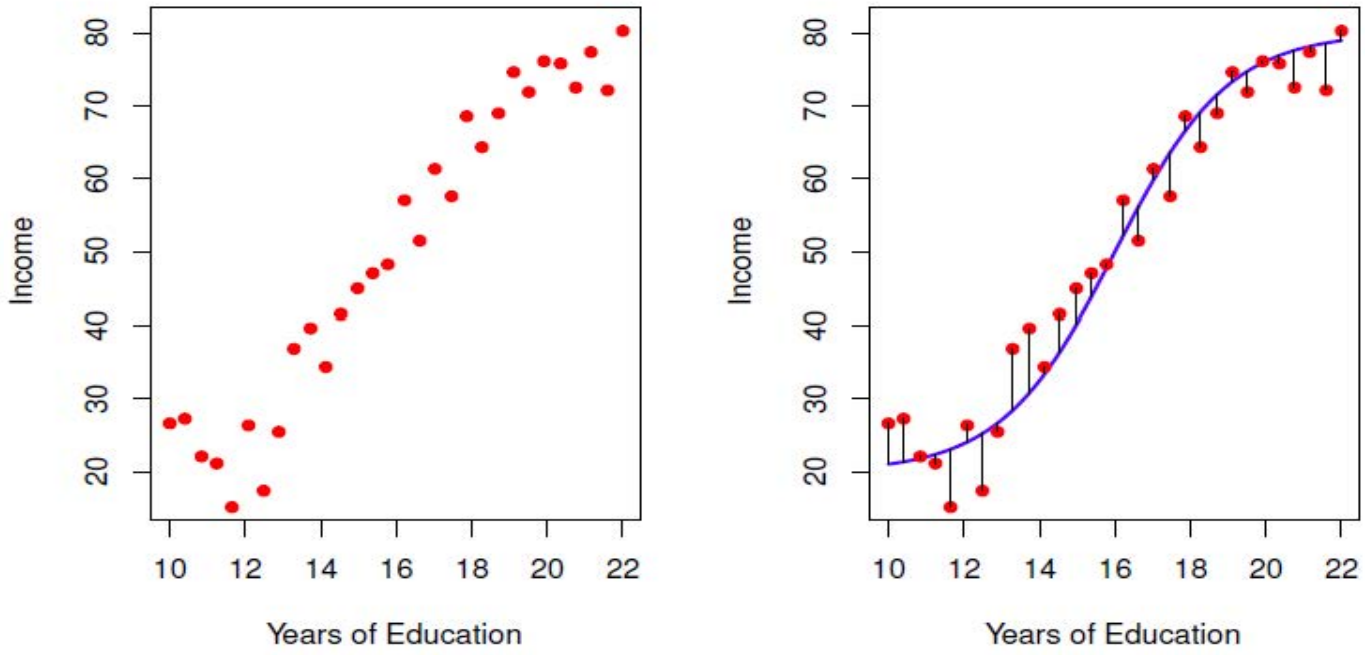
$$f(x) = \frac{1}{1 + \exp(-b_0 - b_1 x_1)}$$

for some constants b_0 and b_1.

Statistical Learning

Example

Let X_1: years of education & Y: income.



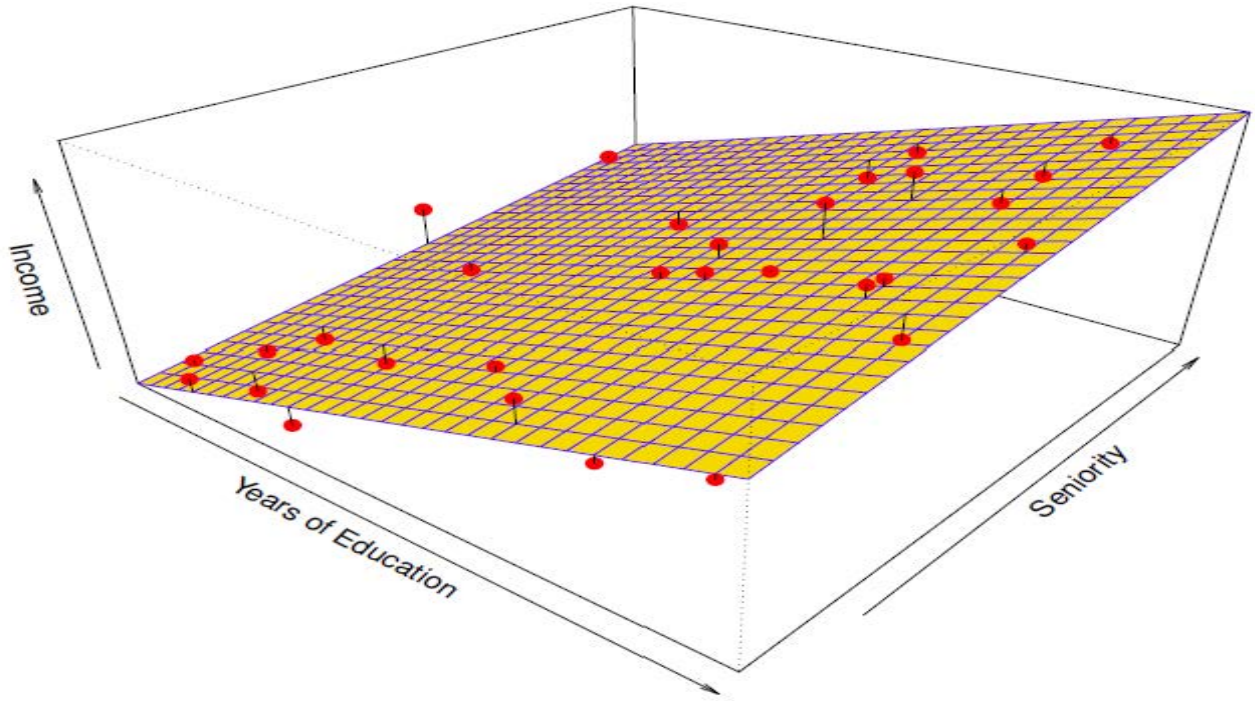
$$f(x) = \frac{1}{1 + \exp(-b_0 - b_1 x_1)}$$

for some constants b_0 and b_1.

Statistical Learning

Example

Let X_1: years of education & X_2: seniority & Y: income.



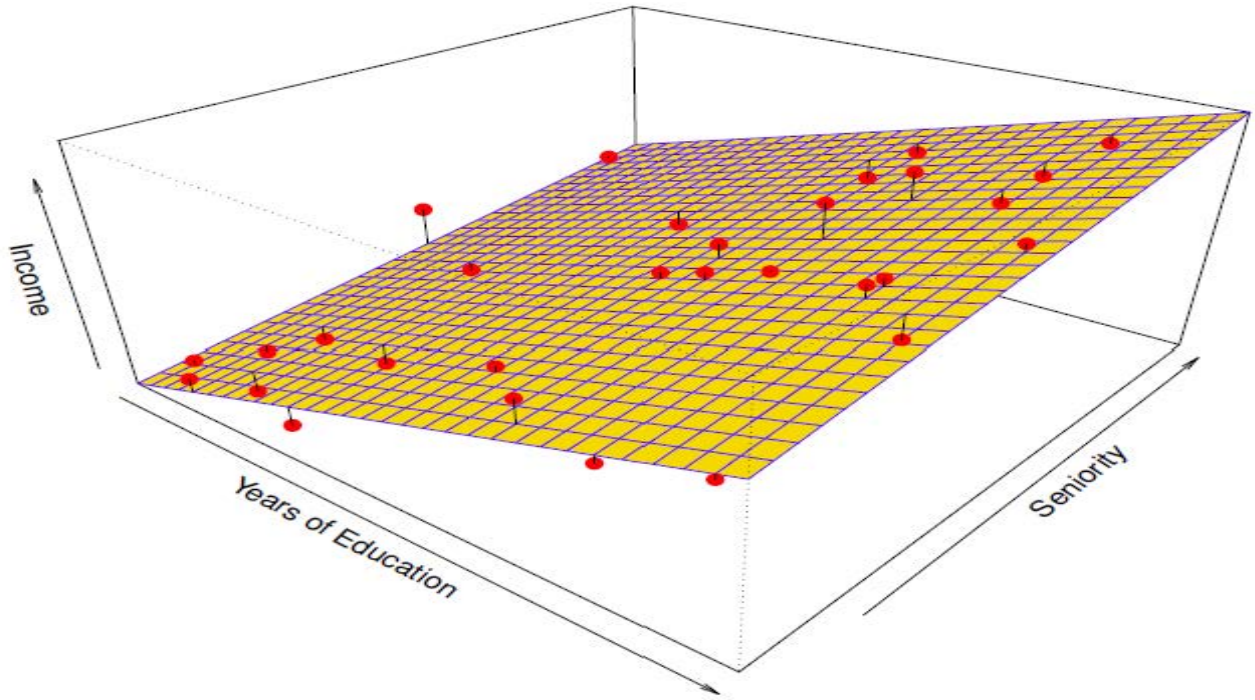
$$f(x) = b_0 + b_1 x_1 + b_2 x_2$$

for some constants b_0, b_1 and b_2.

Statistical Learning

Example

Let X_1: years of education & X_2: seniority & Y: income.



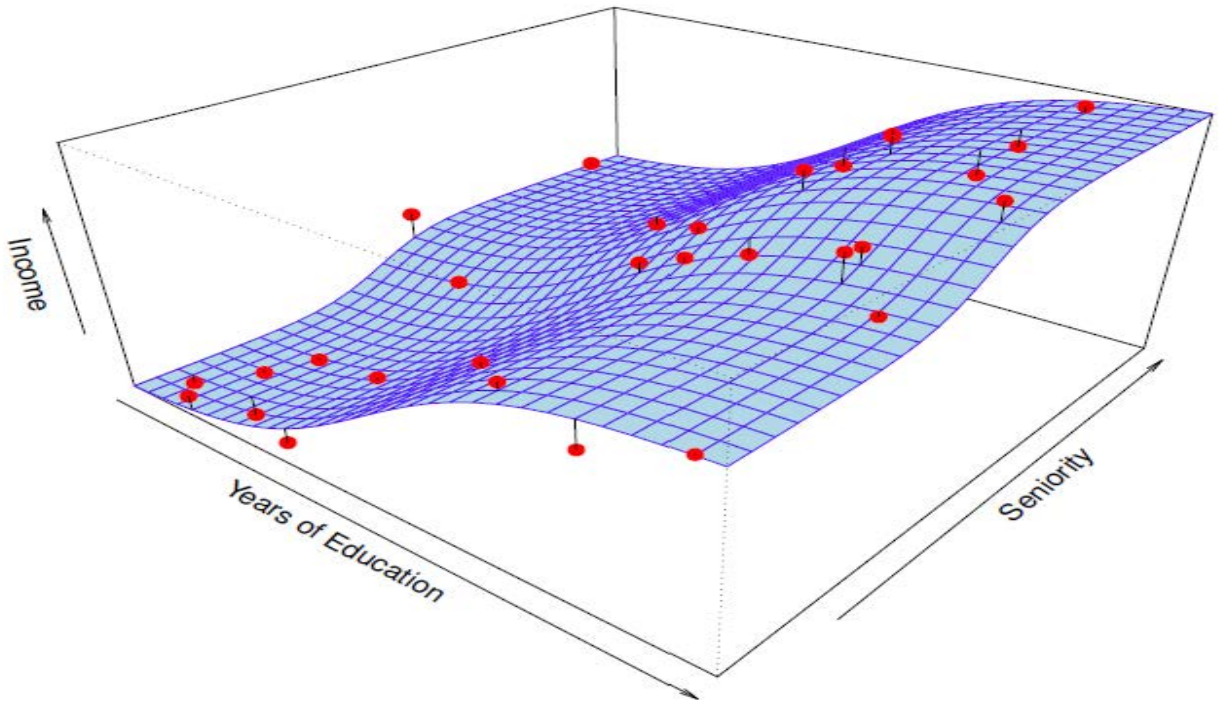
$$f(x) = b_0 + b_1 x_1 + b_2 x_2$$

for some constants b_0, b_1 and b_2.

Statistical Learning

Example

Let X_1: years of education & X_2: seniority & Y: income.



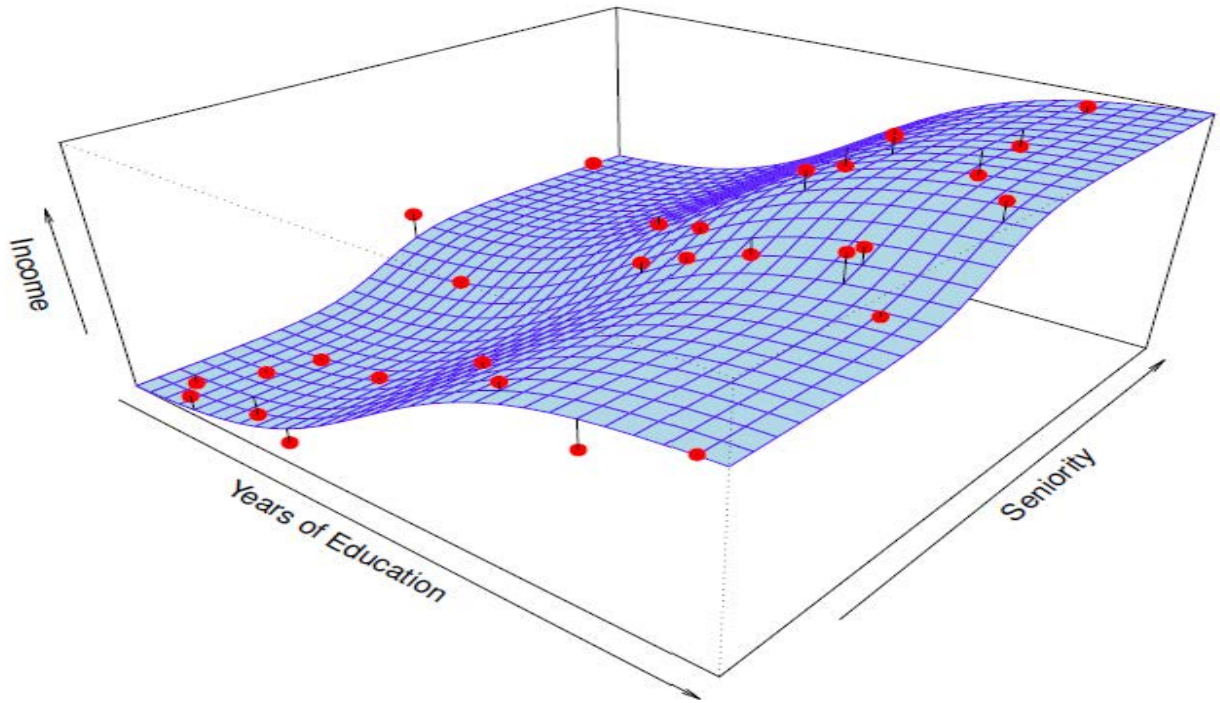
$$f(x) = \frac{c}{1 + \exp(-b_0 - b_1 x_1 - b_2 x_2)}$$

for some constants c, b_0, b_1 and b_2.

Statistical Learning

Example

Let X_1: years of education & X_2: seniority & Y: income.



$$f(x) = \frac{c}{1 + \exp(-b_0 - b_1 x_1 - b_2 x_2)}$$

for some constants c, b_0, b_1 and b_2.

Statistical Learning

Two main reasons to perform statistical modeling and analyses:

1. Prediction of the output Y at a predictor value of X_* by

$$\hat{Y} = \hat{f}(X_*)$$

   while $\hat{f}$ could be like a crystal ball where there is no need to know how it works as long as it provides good prediction.

2. Inference for which inputs $X_1, X_2, \ldots, X_p$ are associated with the output Y and to which direction and at what degree.

It could be also a combination of both.

Statistical Learning

Two main reasons to perform statistical modeling and analyses:

1. Prediction of the output Y at a predictor value of X_* by

$$\hat{Y} = \hat{f}(X_*)$$

   while $\hat{f}$ could be like a crystal ball where there is no need to know how it works as long as it provides good prediction.

2. Inference for which inputs $X_1, X_2, \ldots, X_p$ are associated with the output Y and to which direction and at what degree.

It could be also a combination of both.

Statistical Learning

Two main reasons to perform statistical modeling and analyses:

1. Prediction of the output Y at a predictor value of X_* by

$$\hat{Y} = \hat{f}(X_*)$$

   while $\hat{f}$ could be like a crystal ball where there is no need to know how it works as long as it provides good prediction.
2. Inference for which inputs $X_1, X_2, \ldots, X_p$ are associated with the output Y and to which direction and at what degree.

It could be also a combination of both.

Statistical Learning

Two main reasons to perform statistical modeling and analyses:

1. Prediction of the output Y at a predictor value of X_* by

$$\hat{Y} = \hat{f}(X_*)$$

   while $\hat{f}$ could be like a crystal ball where there is no need to know how it works as long as it provides good prediction.
2. Inference for which inputs $X_1, X_2, \ldots, X_p$ are associated with the output Y and to which direction and at what degree.

It could be also a combination of both.

Statistical Learning

Two main reasons to perform statistical modeling and analyses:

1. Prediction of the output Y at a predictor value of X_* by

$$\hat{Y} = \hat{f}(X_*)$$

   while $\hat{f}$ could be like a crystal ball where there is no need to know how it works as long as it provides good prediction.
2. Inference for which inputs $X_1, X_2, \ldots, X_p$ are associated with the output Y and to which direction and at what degree.

It could be also a combination of both.

Statistical Learning

In prediction, the accuracy of prediction

$$E[(Y - \hat{Y})^2] = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{reducible error}} + \underbrace{Var(\epsilon)}_{\text{irreducible error}}$$

The focus is in estimating f so as to minimize the reducible error by selecting the appropriate Statistical Learning technique.

Meanwhile, the irreducible error depends on ϵ which is due to unmeasured features or covariates and/or random circumstances.

Statistical Learning

Statistical learning is split into:

1. Supervised learning (predictive) Methods performed over an output Y (response) and input X (features/attributes /covariates/predictors).
2. Unsupervised learning (descriptive) Methods performed over input X without the output Y.
3. Reinforcement learning.

Statistical Learning

Statistical learning is split into:

1. Supervised learning (predictive) Methods performed over an output Y (response) and input X (features/attributes /covariates/predictors).
2. Unsupervised learning (descriptive) Methods performed over input X without the output Y.
3. Reinforcement learning.

Statistical Learning

Statistical learning is split into:

1. Supervised learning (predictive) Methods performed over an output Y (response) and input X (features/attributes /covariates/predictors).
2. Unsupervised learning (descriptive) Methods performed over input X without the output Y.
3. Reinforcement learning.

Statistical Learning

Statistical learning is split into:

1. Supervised learning (predictive) Methods performed over an output Y (response) and input X (features/attributes /covariates/predictors).
2. Unsupervised learning (descriptive) Methods performed over input X without the output Y.
3. Reinforcement learning.

Statistical Learning

Methods that we will study in:

1. Supervised learning (predictive):
   - Regression: linear, multiple linear, logistic, Poisson, ridge, lasso
   - Classification: decision trees, nearest neighbour, discriminant analysis, logistic regression, naive Bayes, neural networks, support vector machines
2. Unsupervised learning (descriptive):
   - Association: finding correlations between variables
   - Cluster analysis: identify similarities and groupings between objects

Statistical Learning

Methods that we will study in:

1. Supervised learning (predictive):
   - Regression: linear, multiple linear, logistic, Poisson, ridge, lasso
   - Classification: decision trees, nearest neighbor, discriminant analysis, logistic regression, naïve Bayes, neural networks, support vector machines
2. Unsupervised learning (descriptive):
   - Association: finding correlations between variables
   - Cluster analysis: identify similarities and group similar objects

Statistical Learning

Methods that we will study in:

1. Supervised learning (predictive):
   - Regression: linear, multiple linear, logistic, Poisson, ridge, lasso
   - Classification: decision trees, nearest neighbor, discriminant analysis, logistic regression, naïve Bayes, neural networks, support vector machines
2. Unsupervised learning (descriptive):
   - Association: finding correlations between variables
   - Cluster analysis: identify similarities and group similar objects

Statistical Learning

Methods that we will study in:

1. Supervised learning (predictive):
   - Regression: linear, multiple linear, logistic, Poisson, ridge, lasso
   - Classification: decision trees, nearest neighbor, discriminant analysis, logistic regression, naïve Bayes, neural networks, support vector machines
2. Unsupervised learning (descriptive):
   - Association: finding correlations between variables
   - Cluster analysis: identify similarities and group similar objects

Statistical Learning

Methods that we will study in:

1. Supervised learning (predictive):
   - Regression: linear, multiple linear, logistic, Poisson, ridge, lasso
   - Classification: decision trees, nearest neighbor, discriminant analysis, logistic regression, naïve Bayes, neural networks, support vector machines
2. Unsupervised learning (descriptive):
   - Association: finding correlations between variables
   - Cluster analysis: identify similarities and group similar objects

Statistical Learning

Methods that we will study in:

1. Supervised learning (predictive):
   - Regression: linear, multiple linear, logistic, Poisson, ridge, lasso
   - Classification: decision trees, nearest neighbor, discriminant analysis, logistic regression, naïve Bayes, neural networks, support vector machines
2. Unsupervised learning (descriptive):
   - Association: finding correlations between variables
   - Cluster analysis: identify similarities and group similar objects

Statistical Learning

Methods that we will study in:

1. Supervised learning (predictive):
   - Regression: linear, multiple linear, logistic, Poisson, ridge, lasso
   - Classification: decision trees, nearest neighbor, discriminant analysis, logistic regression, naïve Bayes, neural networks, support vector machines
2. Unsupervised learning (descriptive):
   - Association: finding correlations between variables
   - Cluster analysis: identify similarities and group similar objects

Statistical Learning

Example (Classification of emails as spam or email/inbox)

- Training data: 4601 emails with spam or email as an output.
- Inputs are the percentages of 57 of words and punctuation marks that appear in emails and their percentages:

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

- Supervised learning (a classification problem)
- Goal: To set up rules of classification, e.g.

  if (%george < 0.6)&(%you > 1.5) then spam else email/inbox

Statistical Learning

Example (Classification of emails as spam or email/inbox)

- Training data: 4601 emails with spam or email as an output.
- Inputs are the percentages of 57 of words and punctuation marks that appear in emails and their percentages:

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

- Supervised learning (a classification problem)
- Goal: To set up rules of classification, e.g.

  if (%george < 0.6)&(%you > 1.5) then spam else email/inbox

Statistical Learning

Example (Classification of emails as spam or email/inbox)

- Training data: 4601 emails with spam or email as an output.
- Inputs are the percentages of 57 of words and punctuation marks that appear in emails and their percentages:

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

- Supervised learning (a classification problem)
- Goal: To set up rules of classification, e.g.

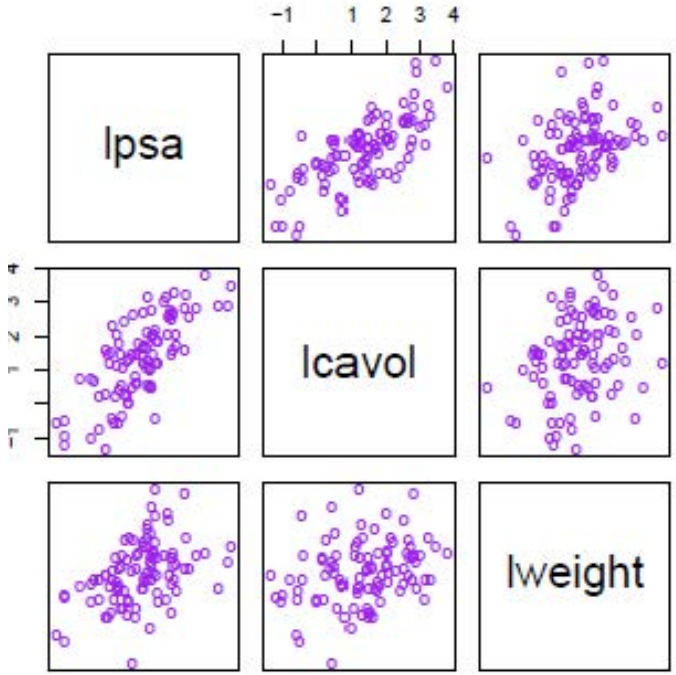
  if (%george < 0.6)&(%you > 1.5) then spam else email/inbox

Statistical Learning

Example (Predicting prostate cancer)

- Training data: 97 men
- Output: (logarithmic) levels of prostate specific antigen (lpsa)
- Inputs: a number of measurements, e.g. log of cancer volume (lcavol) and log prostate weight (lweight)
- Supervised learning (a regression problem)
  - Goal: Set up regression model to be used for prediction, e.g.

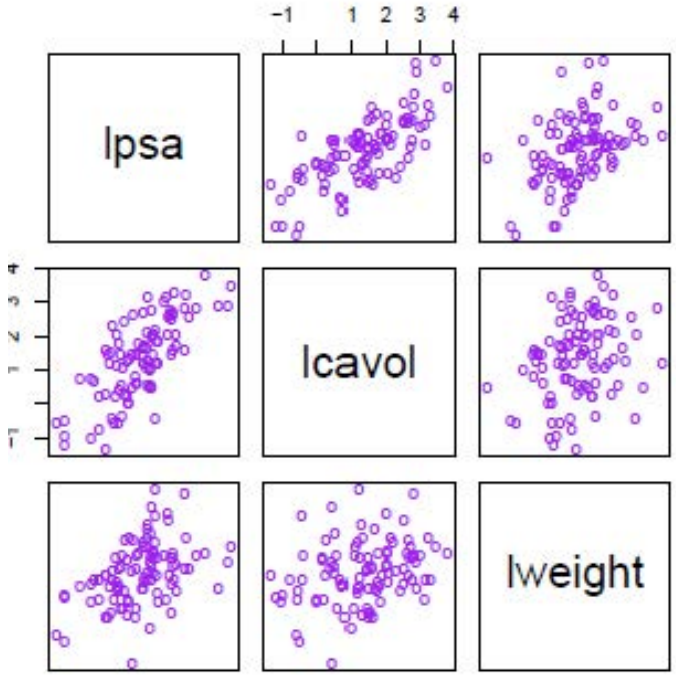
$$\text{lpsa} = 2.46 + 0.68\ \text{lcavol} + 0.26\ \text{lweight} + \ldots$$



Statistical Learning

Example (Predicting prostate cancer)

- Training data: 97 men
- Output: (logarithmic) levels of prostate specific antigen (lpsa)
- Inputs: a number of measurements, e.g. log of cancer volume (lcavol) and log prostate weight (lweight)
- Supervised learning (a regression problem)



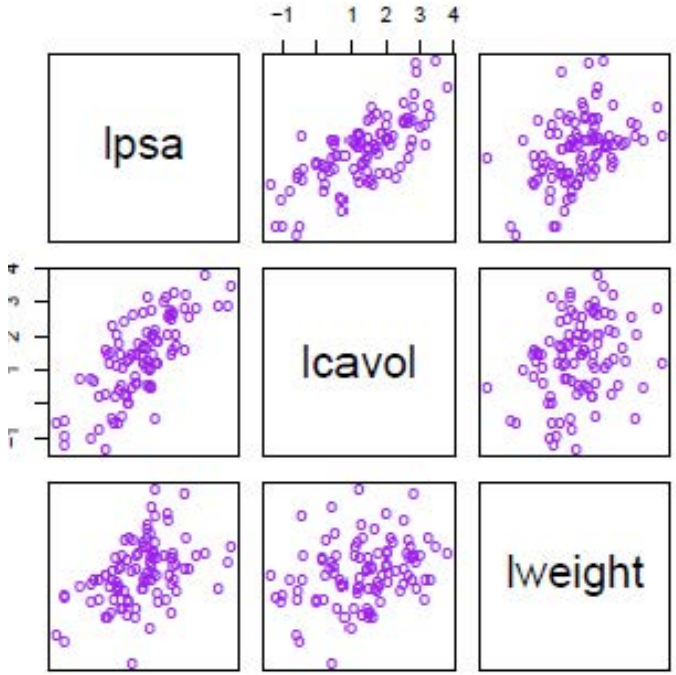
- Goal: Set up regression model to be used for prediction, e.g.

$$\text{lpsa} = 2.46 + 0.68\ \text{lcavol} + 0.26\ \text{lweight} + \ldots$$

Statistical Learning

Example (Predicting prostate cancer)

- Training data: 97 men
- Output: (logarithmic) levels of prostate specific antigen (lpsa)
- Inputs: a number of measurements, e.g. log of cancer volume (lcavol) and log prostate weight (lweight)
- Supervised learning (a regression problem)



  - Goal: Set up regression model to be used for prediction, e.g.

$$\text{lpsa} = 2.46 + 0.68\ \text{lcavol} + 0.26\ \text{lweight} + \ldots$$

Statistical Learning

Example (Classification of iris flowers)

Setosa

Versicolor

Virginica

Goal: To classify an iris flower based on the length and width of both its sepal and petal (inputs). A supervised learning problem with flower type as an output.

Statistical Learning

Example (Classification of iris flowers)

Setosa

Versicolor

Virginica

Goal: To classify an iris flower based on the length and width of both its sepal and petal (inputs). A supervised learning problem with flower type as an output.

Statistical Learning

Example (Classification of iris flowers)

Setosa

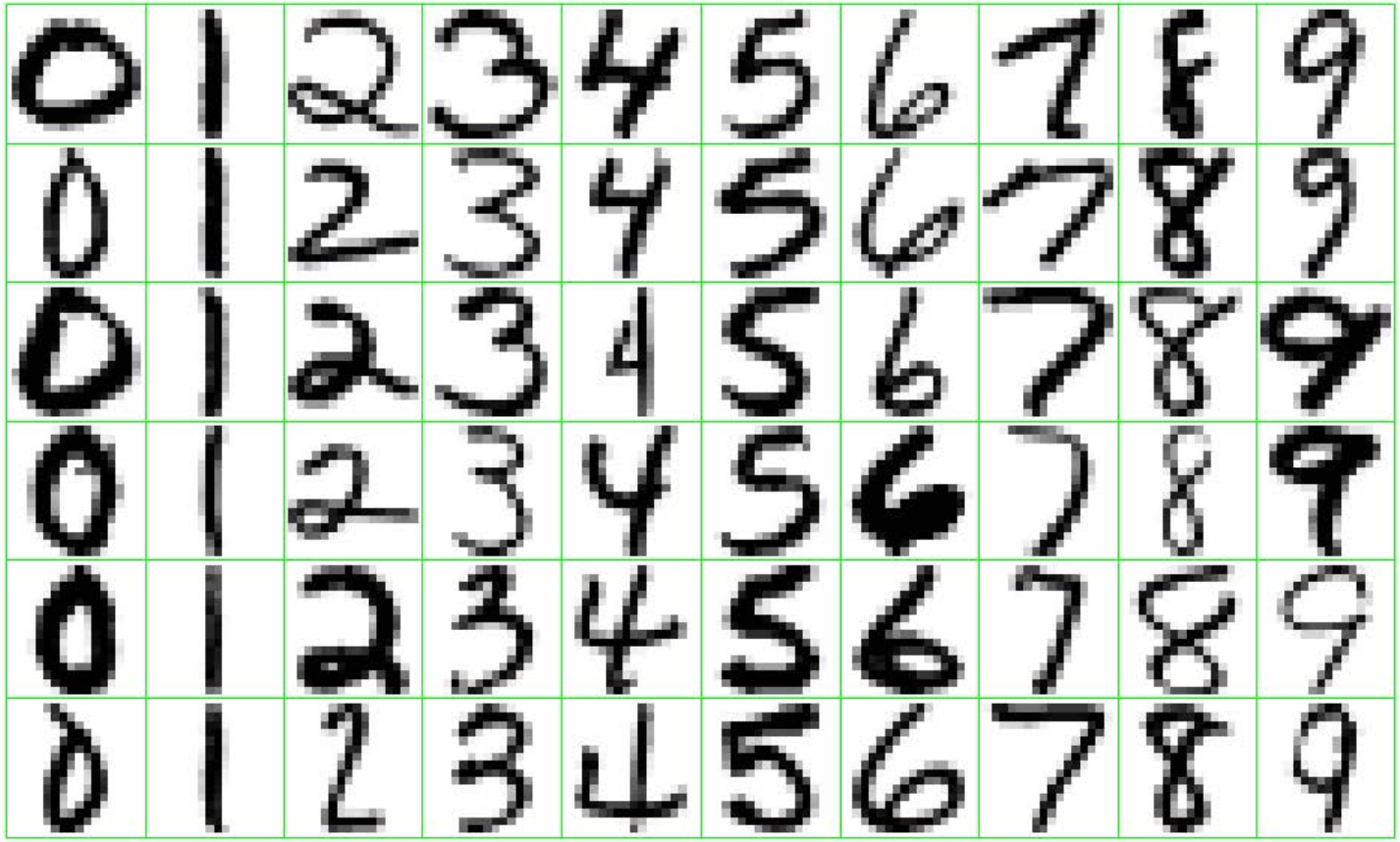
Versicolor

Virginica

Goal: To classify an iris flower based on the length and width of both its sepal and petal (inputs). A supervised learning problem with flower type as an output.

Statistical Learning

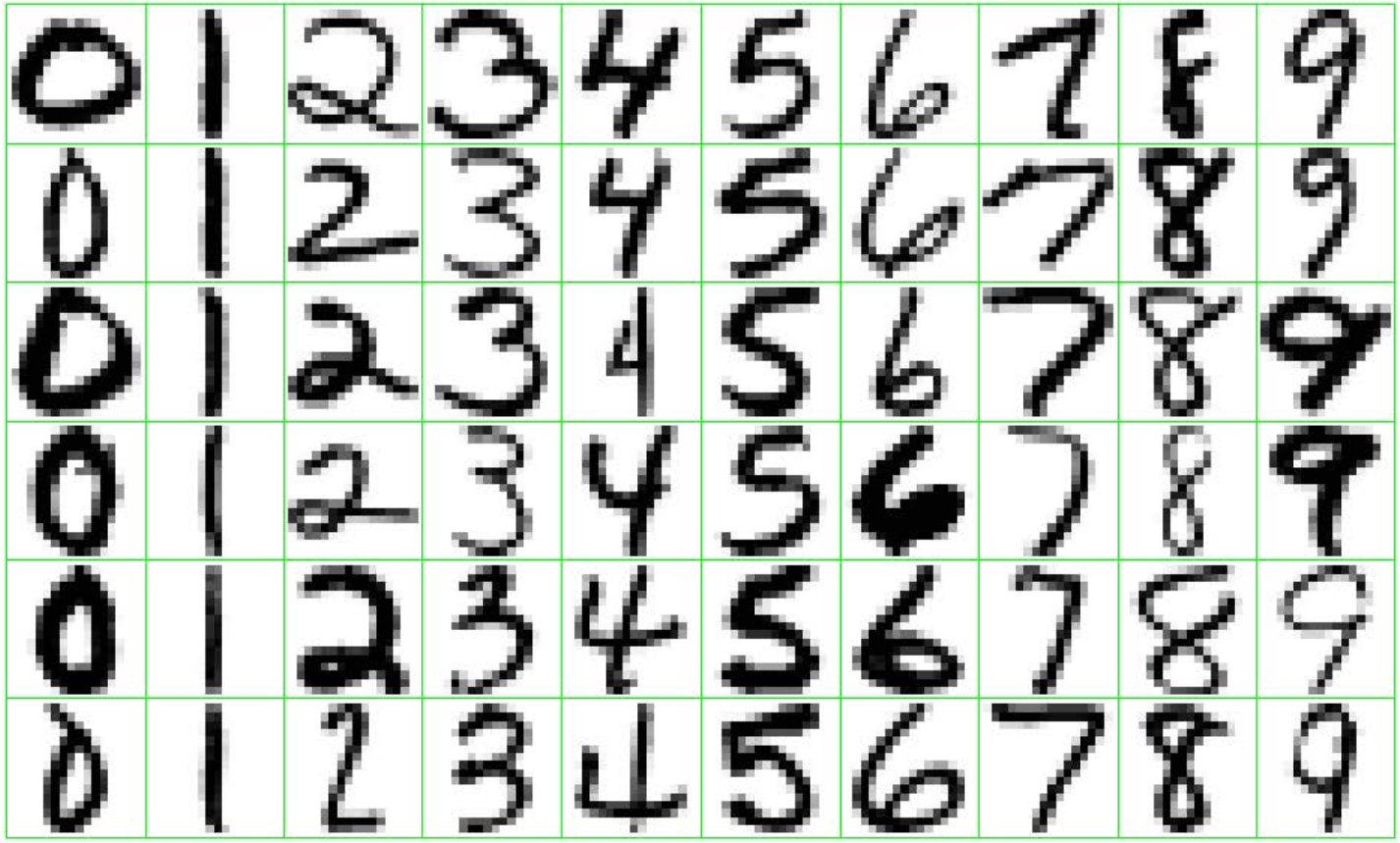
Example (Handwriting and image recognition)



Goal: To recognize pictures of the numerals on US postal envelopes. Still a supervised learning problem with outputs given by the actual numerals.

Statistical Learning

Example (Handwriting and image recognition)

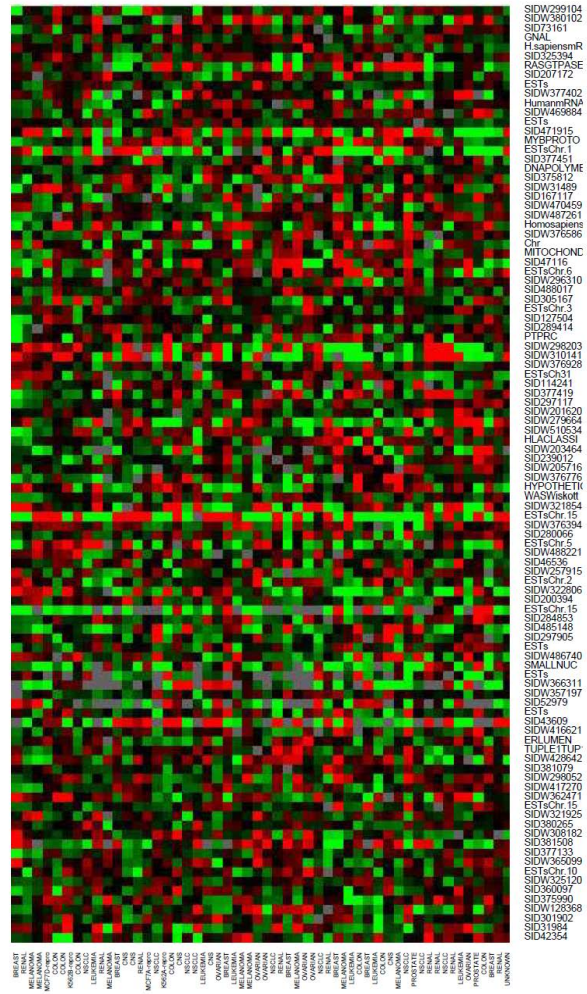


Goal: To recognize pictures of the numerals on US postal envelopes. Still a supervised learning problem with outputs given by the actual numerals.

Statistical Learning

Example (Gene Expression in DNA micro-array Data - NCI60 data set)

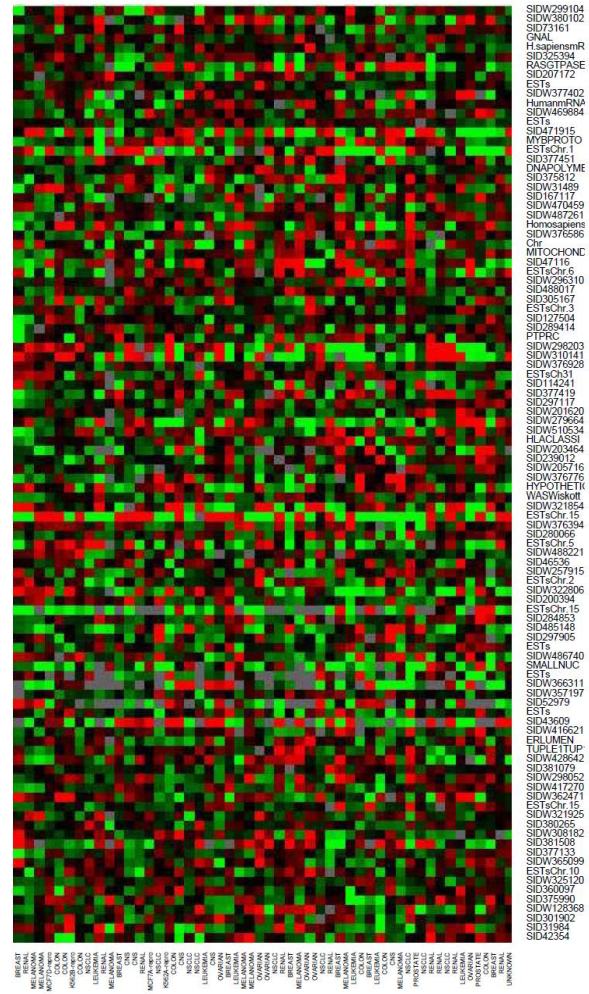
- A heat map of 100 rows of randomly selected from the 6,380 genes and 64 columns (sampled cancer patients with 14 types).
- Outputs: Levels of expression from green (negative or under-expressed gene) and red (positive or over-expressed gene). Grey means missing.
- Goal: To predict the levels of gene expression using the 64 patients' cancer cell lines with or without the 14 types of cancer (inputs). It is a supervised learning.



Statistical Learning

Example (Gene Expression in DNA micro-array Data - NCI60 data set)

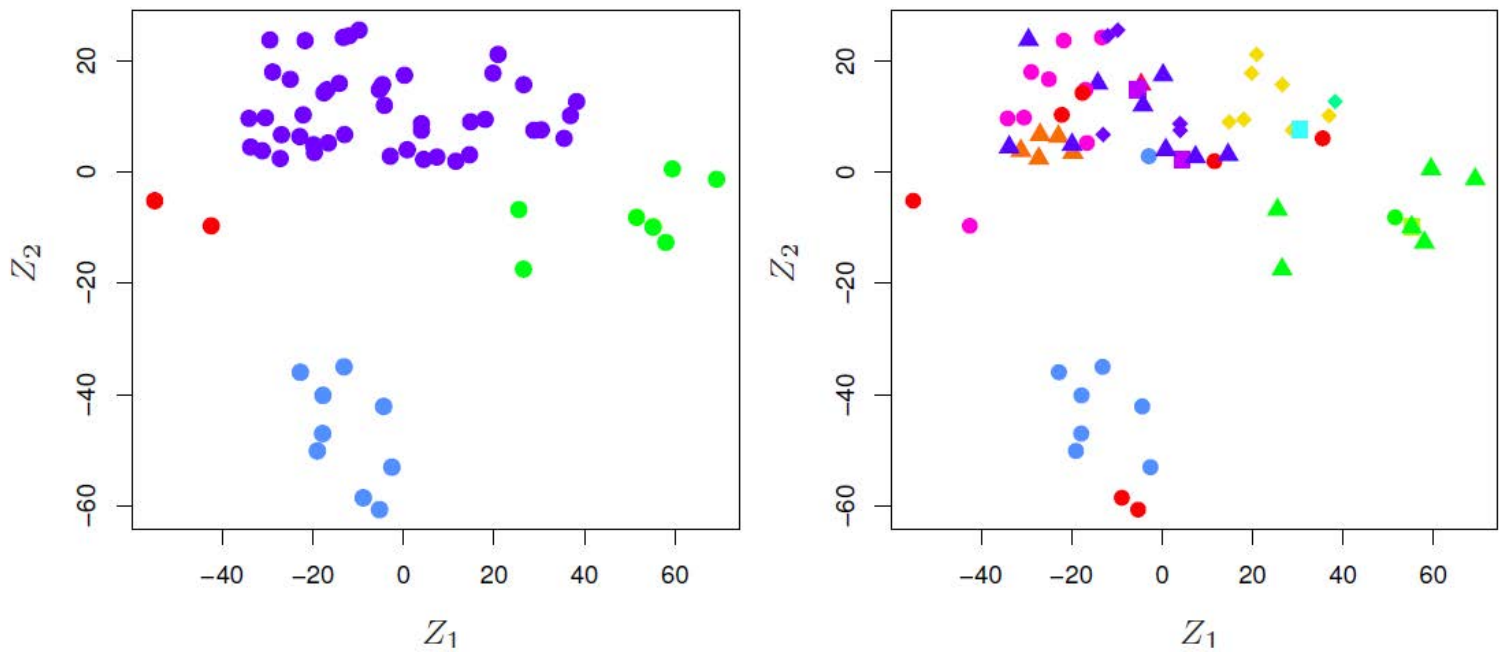
- A heat map of 100 rows of randomly selected from the 6,380 genes and 64 columns (sampled cancer patients with 14 types).
- Outputs: Levels of expression from green (negative or under-expressed gene) and red (positive or over-expressed gene). Grey means missing.
- Goal: To predict the levels of gene expression using the 64 patients' cancer cell lines with or without the 14 types of cancer (inputs). It is a supervised learning.



Statistical Learning

Example (Gene Expression Data - NCI60 data set)

Goal: To just identify groups (clusters) of the 64 patients' cancer cell lines without and with the 14 types of cancer. It is an "unsupervised" learning problem (clustering problem), if we ignore the levels of gene expression as outputs.

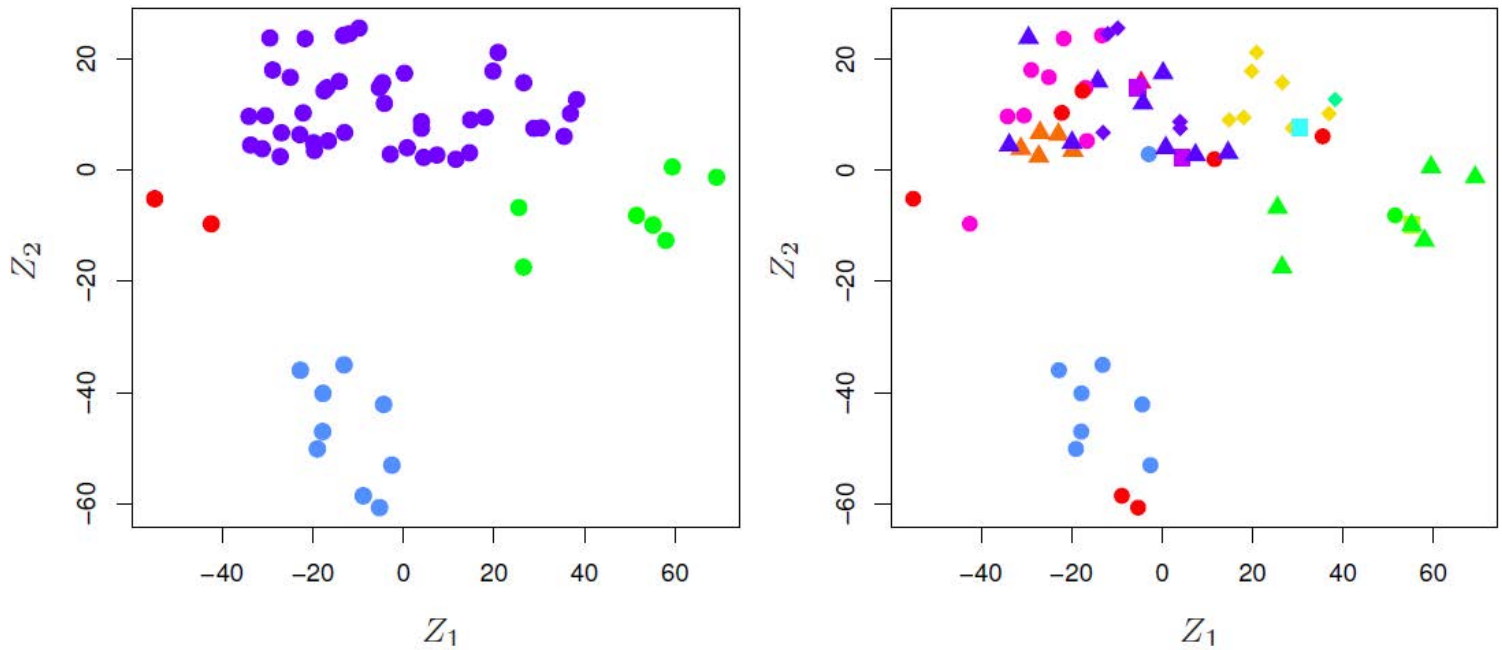


Z_1 and Z_2 are the first two principal components that summarize the 6,830 gene expression measurements.

Statistical Learning

Example (Gene Expression Data - NCI60 data set)

Goal: To just identify groups (clusters) of the 64 patients' cancer cell lines without and with the 14 types of cancer. It is an "unsupervised" learning problem (clustering problem), if we ignore the levels of gene expression as outputs.



Z_1 and Z_2 are the first two principal components that summarize the 6,830 gene expression measurements.

End of Set 1