MM 225 – AI and Data Science

Day 23: Supervised Learning 1 : Regression Analysis

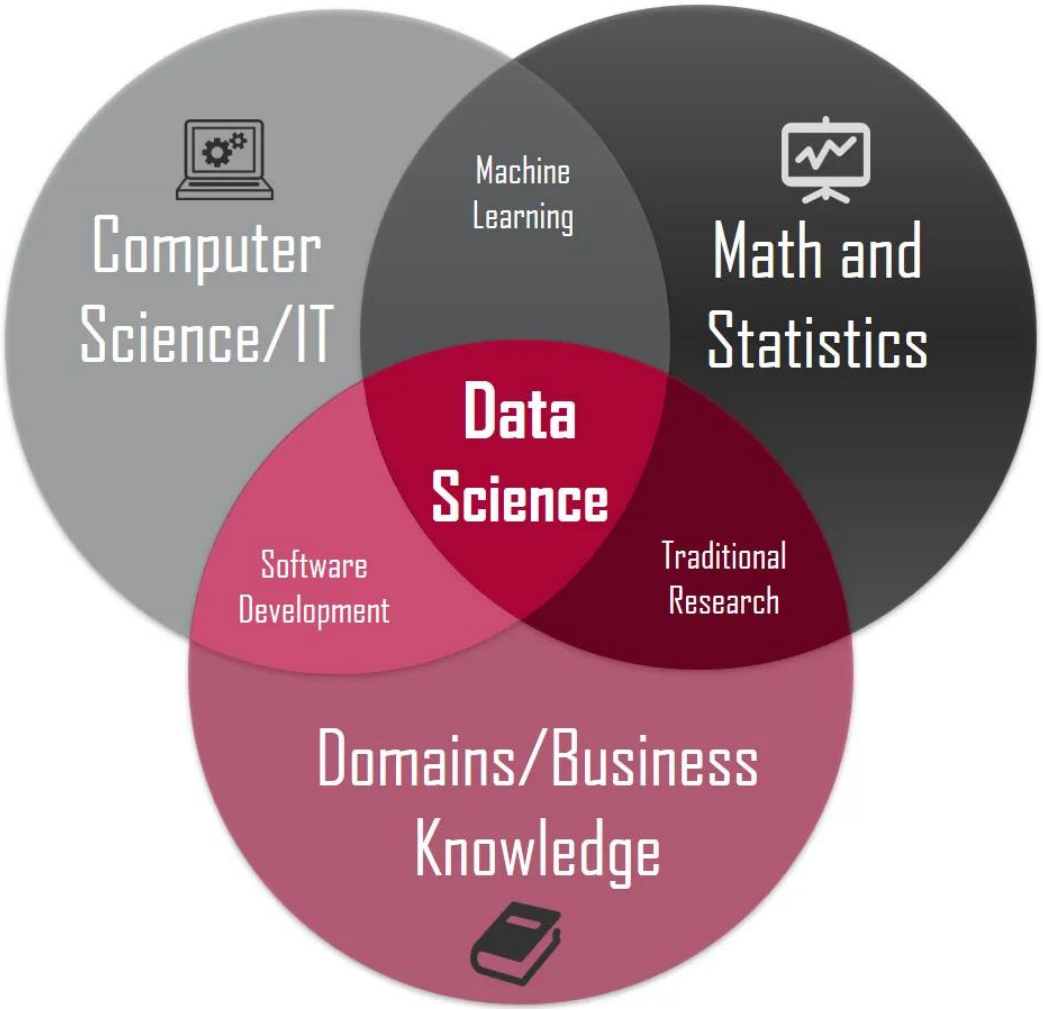
Instructors: Hina Gokhale, MP Gururajan, N. Vishwanathan

26 SEPTEMBER 2024

Text to follow:

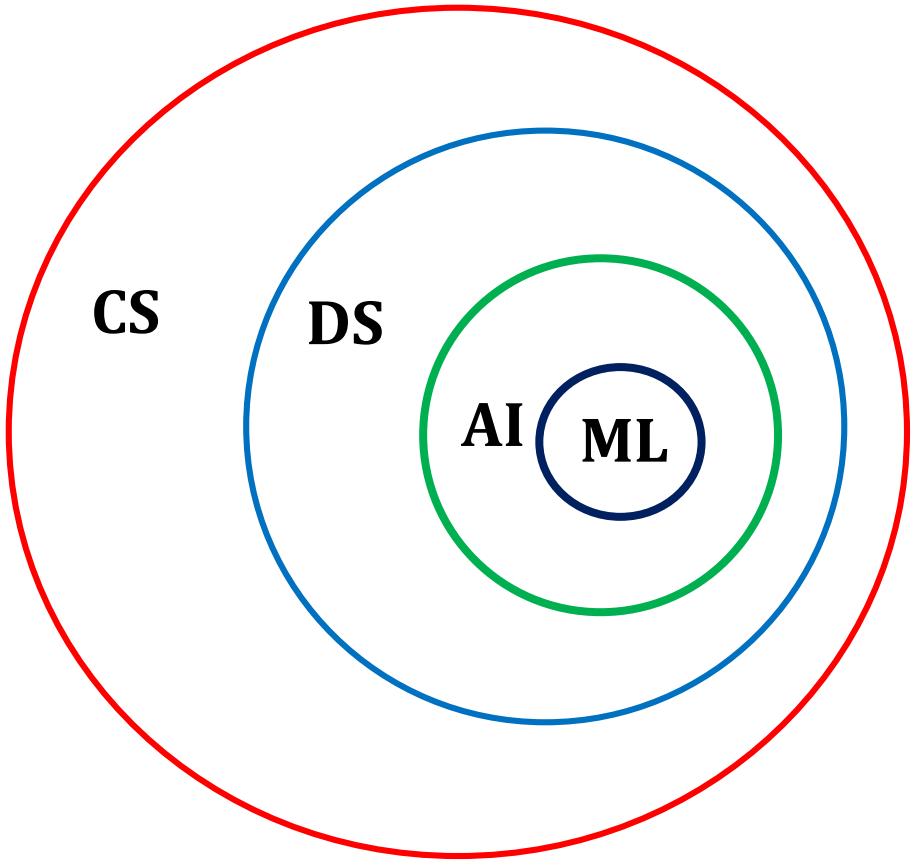
Principles and Techniques for Data Science, by Sam Lau, Joey Gonzalez and Deb Nolan, 2019 : https://www.textbook.ds100.org/intro

Machine Learning



Ref: Michael Barber, Towards Data Science, Jan 14 2018

Machine Learning



Oliver Theobald: "Machine Learning for Absolute Beginner"

Machine Learning

- Learning from data
- Three types of learning
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

- Supervised learning examples
  - Regression analysis
  - logistic regression
- Unsupervised learning
  - Exploratory Data Analysis (EDA) / Descriptive Statistics
  - Clustering – k mean clustering
- Reinforcement Learning

What is Learning in ML?

Ordinary programming : Decision Rules are defined and implemented through Algorithm

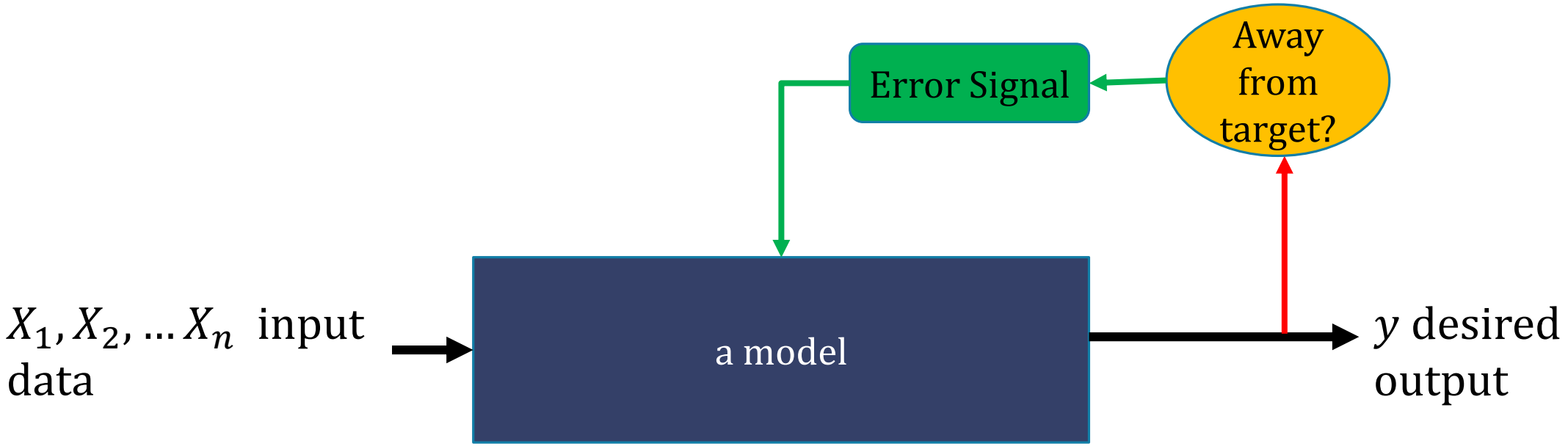
Typical ML Algorithm

- learns from data by analyzing patterns
- decisions are based on patterns
- errors are corrected in the process
- new data may lead to different decision
- uses statistics and probability

Supervised Learning

Data: [y; $X_1, X_2, \ldots X_n$]

Model imagined is : $y = f(X_1, X_2, \ldots X_n) + \varepsilon$

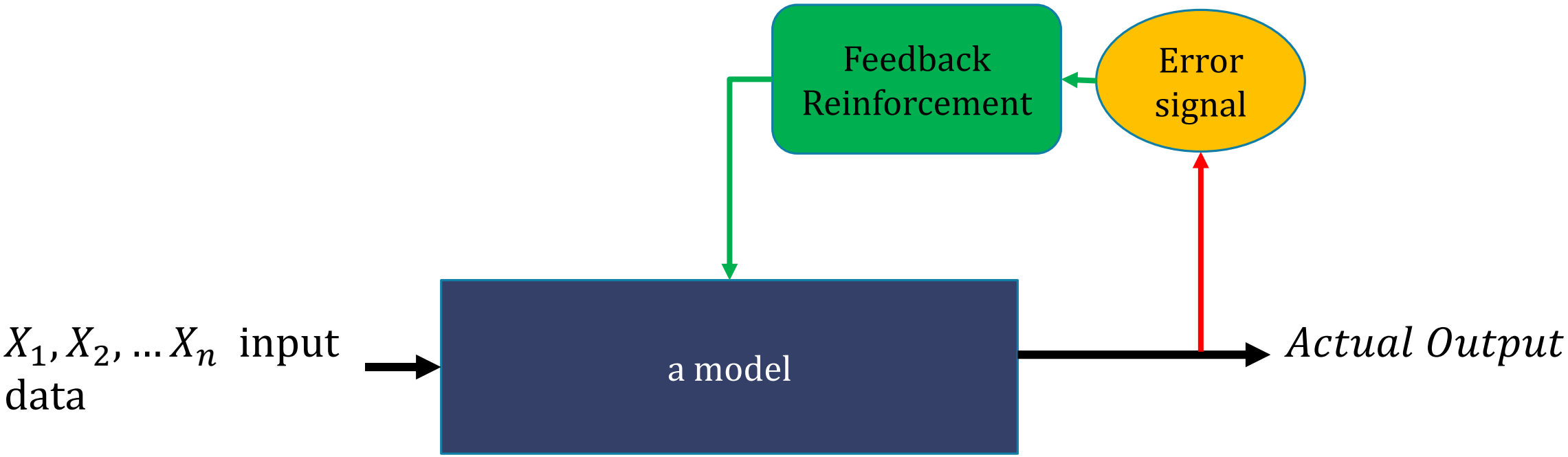


Unsupervised Learning

Data: $[X_1, X_2, \ldots X_n]$

Reinforced Learning

Data: $[X_1, X_2, \ldots X_n]$

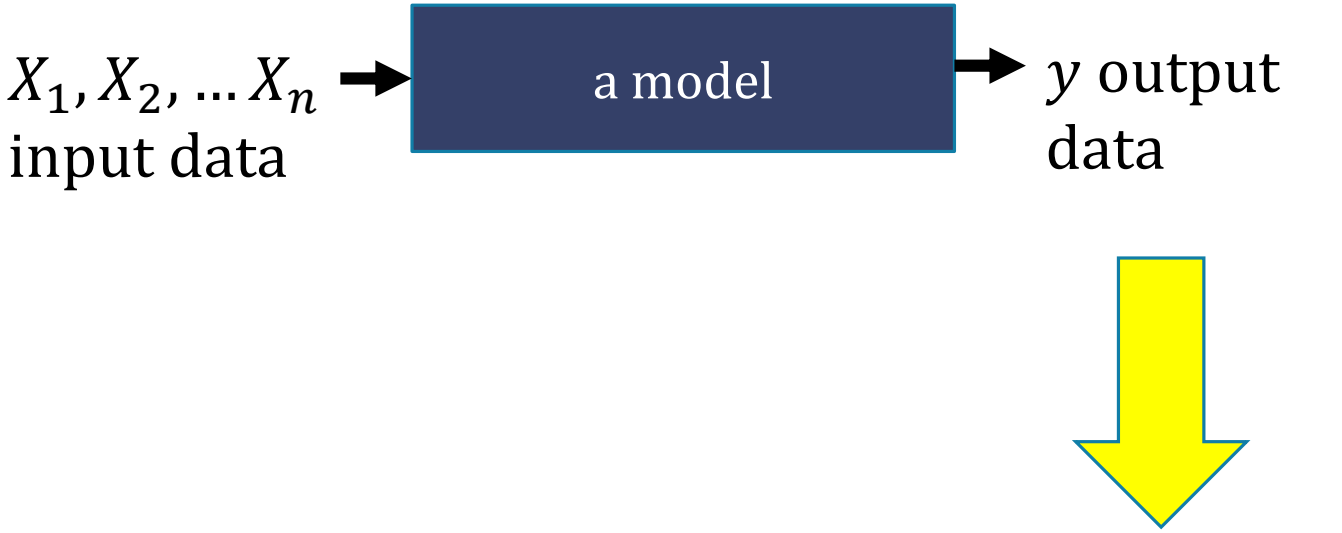


Supervised Learning – Data

Date	Prod(Tap)	Coke Rate	PCI	Sinter %	HB Press	HBT	Top Press	Top Temp 1	Top Temp 2	Tey Vel	Slag Volume	Oxygen Enr	Steam	RAFT
01-Jan-21	850	464	83	63	1.56	1051	0.47	148	188	241	345	962	1.19	1995
02-Jan-21	1004	443	105	63	1.68	1055	0.54	150	197	250	365	1314	1.17	2049
03-Jan-21	1026	441	114	64	1.71	1058	0.57	135	182	255	375	1465	0.85	2101
04-Jan-21	1065	440	114	64	1.70	1058	0.56	131	179	256	345	1780	0.86	2130
05-Jan-21	1056	438	113	63	1.70	1058	0.55	138	183	253	340	1700	0.78	2127
06-Jan-21	1071	436	119	64	1.70	1058	0.55	145	190	255	345	1863	0.82	2124
07-Jan-21	1043	442	103	65	1.65	1058	0.56	137	173	258	360	1516	0.91	2081
08-Jan-21	1083	436	120	66	1.70	1058	0.58	126	173	256	350	1995	0.74	2149
09-Jan-21	1042	444	119	66	1.66	1059	0.55	131	172	254	345	1805	0.83	2122
10-Jan-21	1074	434	119	65	1.72	1059	0.56	142	184	254	395	1858	0.82	2118
11-Jan-21	1086	431	117	65	1.73	1060	0.56	125	163	253	345	1797	0.83	2113

Supervised Learning – Data

SNo	X_1	X_2	y
0	-0.86914	0.38931	0
1	-0.99347	-0.61059	0
2	-0.83406	0.239236	0
3	-0.13647	0.632003	1
4	0.403887	0.310784	1
5	-0.56931	-0.24668	0
6	-0.10998	0.930917	1
7	0.288994	-0.53269	1
8	0.319782	0.664582	1
9	0.558686	-0.62118	1
10	0.886302	-0.7767	0
11	0.288676	-1	0
12	0.497748	0.344112	1
13	0.12674	0.966287	1
14	-0.72826	0.331071	0
15	-0.31453	0.716929	1



Unsupervised Learning

TC_NO	HEAT_NO	SIZE	GS_CODE	DP_CODE	TEMP	YS	UTS	EL	ROFA
13903	G1172	160 Dia	4.00	5.00	25	1154.00	1446.00	12.00	20.00
13903	G1172	160 Dia	4.00	5.00	25	1197.00	1418.00	15.00	29.00
13903	G1172	160 Dia	4.00	5.00	25	1232.00	1481.00	14.00	24.00
13903	G1172	160 Dia	4.00	5.00	25	1239.00	1463.00	14.00	26.00
13903	G1172	160 Dia	4.00	5.00	650	972.00	1082.00	20.00	33.00
13903	G1172	160 Dia	4.00	5.00	650	993.00	1116.00	16.00	35.00
13903	G1172	160 Dia	4.00	5.00	650	1027.00	1136.00	24.00	41.00
13903	G1172	160 Dia	4.00	5.00	650	1040.00	1174.00	17.00	47.00
13903	G1172	160 Dia	4.00	5.00	650	1079.00	1167.00	16.00	34.00
14106	G1214	160 Dia	1.00	5.00	25	1160.00	1380.00	24.00	42.00
14106	G1214	160 Dia	1.00	5.00	25	1210.00	1415.00	21.00	44.00
14106	G1214	160 Dia	1.00	5.00	650	970.00	1090.00	14.00	23.00
14107	G1212	160 Dia	1.00	2.00	25	1180.00	1380.00	26.00	52.00
14107	G1212	160 Dia	1.00	2.00	25	1200.00	1400.00	20.00	50.00

Unsupervised Learning

- Data: $\{X_1, X_2, \ldots X_n\}$
- No explicit relationship
- Explore possibilities
  - distribution – *example EDA*
  - outliers – *example EDA*
  - clusters - *example k-mean cluster*
  - …
- Reduction of Dimensionality

Simple Linear Model

- Simplest form of supervised learning
- Consider y a response variable and x_1, x_2, …., x_r as independent variables.
- *Assumption: there is a causal relationship between y and x_1, x_2, …., x_r*
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \epsilon,$
- where β_0, β_1, … β_r are some constants and ϵ represents random error in this relationship
- If r = 1 , then $Y = \beta_0 + \beta_1 x_1 + \epsilon$ is called simple linear regression model
- In general with r independent variables it is called multiple regression model
- $\beta_0, \beta_1, \beta_2 \ldots \beta_r$ are called regression coefficients….and need to be estimated from the data

Variety of Regression models

- Regression can cover a wide variety of relationships
  - Polynomial relationship
  - Power function
  - Exponential

Uses of Regression Analysis

1. Prediction and Forecasting
   - Overlap with machine learning application
   - Need to justify the predictive power

2. Setting up causal relationship
   - Examples shown in the previous slides
   - Need to justify the existence of the relationship

Regression as conditional Expectation

- In regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \epsilon$, ϵ is called error term
- *The random error ϵ is assumed to have expected value 0, then*
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \epsilon$ can also be written as
- $E(y|x_1, x_2, \ldots, x_r) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$
- We want to estimate the regression coefficients $\beta_0, \beta_1, \beta_2 \ldots \beta_r$ for given the observed values of $x_1, x_2, \ldots, x_r$ with following assumptions

1. ϵ is random, $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$ are independent
2. $x_1, x_2, \ldots, x_r$ are independent

Case of simple linear regression

- Want to estimate β_0, β_1 in the regression relation

$$y = \beta_0 + \beta_1 x + \epsilon$$

- There are several methods to find estimator of the regression coefficients β_0, β_1

1. Least Squares Estimators
2. Maximum Likelihood Estimators when $\epsilon \sim some\ distribution$
   i. Normal: Linear Regression
   ii. Bernoulli trial: Logistic Regression
   iii. Poisson : Poisson Regression

} **General Linear Model**

Least Squares Estimators of β_0, β_1

- Let $(Y_i, x_i) : i = 1, 2, \dots, n$ be the data.
- These can be expressed as $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ : \ i = 1, 2, \dots, n$
- Want to find β_0, β_1 by minimizing the squared error between values of Y_i and its estimator $\beta_0 + \beta_1 x_i$
- Let us denote estimated value of β_0, β_1 as A and B respectively
- Want find β_0 and β_1 that would minimize sum of squares of deviations from the observations from the regression line:
- $SS = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$

$$\frac{\partial SS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial SS}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Simplify these two equation leads to

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Least Squares Normal Equations

Further simplification we will get

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum(x_i - \bar{x})y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Now, let r = correlation coefficient between (x, y)

$$r = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\left((x_i - \bar{x})^2\right)\left((y_i - \bar{y})^2\right)}} = \frac{Sxy}{\sqrt{SxxSyy}}$$

And thus,

$$\hat{y} = \bar{y} + r\sqrt{\frac{Syy}{Sxx}}\sum(x_i - \bar{x})$$

Thus r measures the strength of linear relationship

However, it **does not necessarily** confirm any linear association!

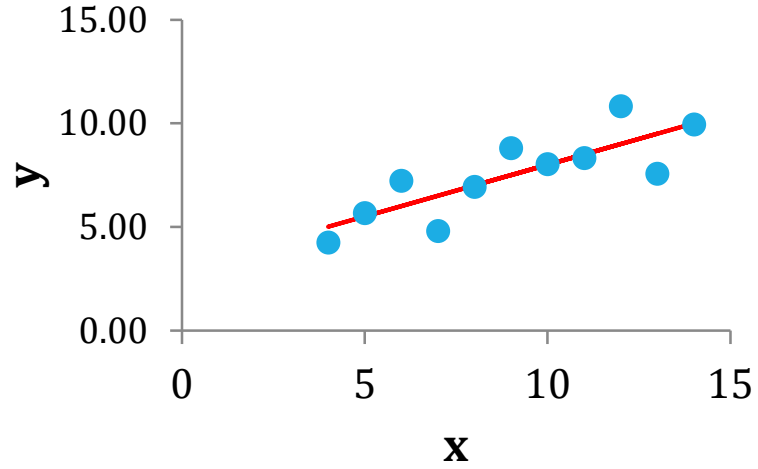
Anscombe's quartet

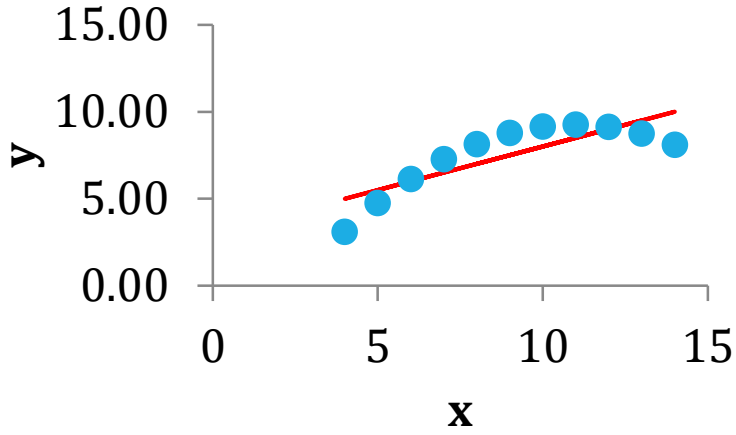
Dataset I		Dataset II		Dataset III		Dataset IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

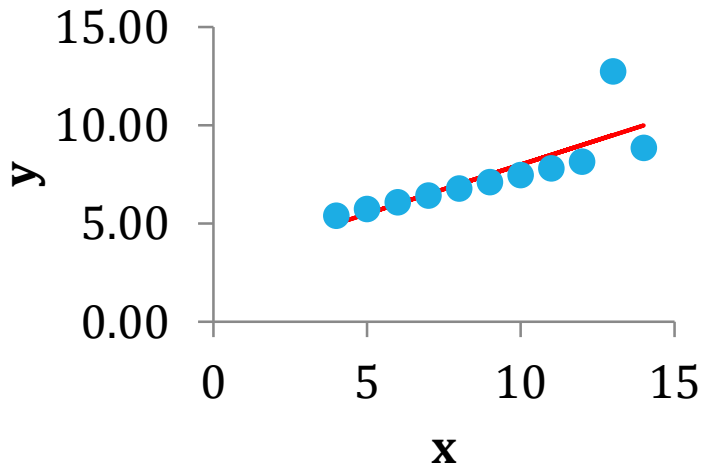
Anscombe's Quartet

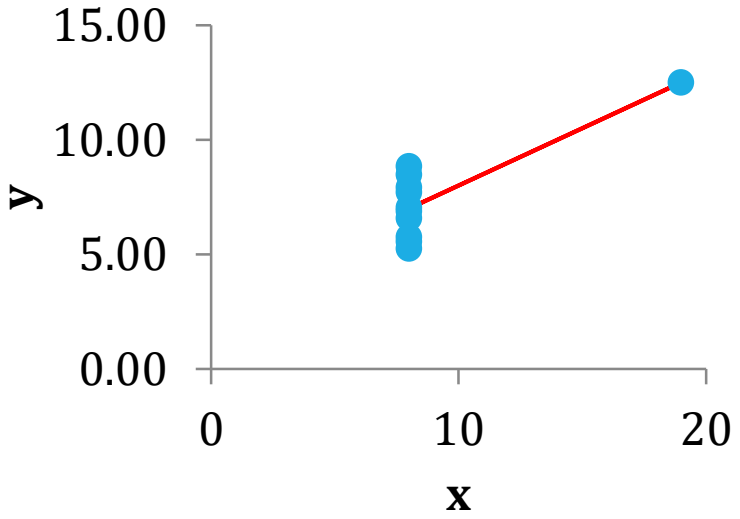
1. All data set have same correlation coefficient = 0.816
2. All data set have the same regression line: y = 3.0 + 0.5x
3. All the data set have the same coefficient of determination = 0.67
4. But they do not display the same relationship

Anscombe's quartet









Summary

General introduction to Machine Learning

- Supervised Learning
- Unsupervised Learning

Basic case of Supervised Learning – Linear Regression Model

Simple regression model and least squares estimates of coefficients

Anscombe's quartet to explain the difference between liner model and linear relationship

Thank you….