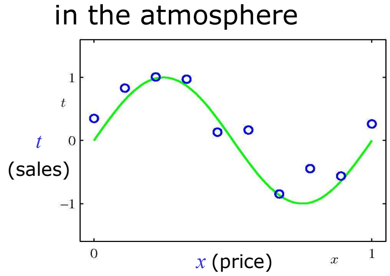
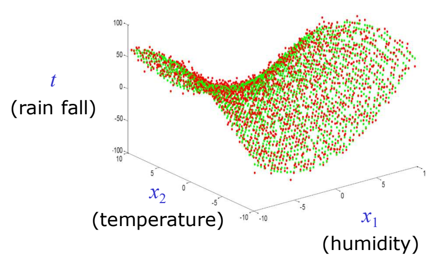
Regression (Prediction)

Prediction (Regression)

- Numeric prediction: Task of predicting continuous (or ordered) values for given input
- Example:
  - Predicting potential sales of a new product given its price
  - Predicting rain fall given the temperature and humidity in the atmosphere



$t = f(x)$



$t = f(\mathbf{x})$

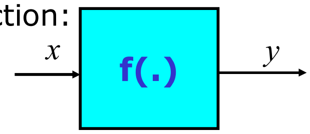
$\mathbf{x} = [x_1, x_2]^T$

- Regression and prediction are synonymous terms

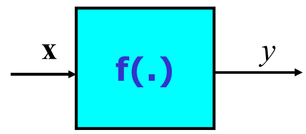
2

Prediction (Regression)

- Regression analysis is used to model the relationship between one or more independent (predictor) variable and a dependent (response) variable
  - Dependent variable is always continuous valued or ordered valued
  - Example: Dependent variable: Rain fall
    Independent variable(s): temperature, humidity
- The values of predictor variables are known
- The response variable is what we want to predict
- Regression analysis can be viewed as mapping function:



- Single independent variable (x)
- Single dependent variable (y)



- Multiple independent variable ($\mathbf{x} \in \mathbb{R}^d$)
- Single dependent variable (y)

3

Prediction (Regression)

- Regression is a two step process
  - Step1: Building a regression model
    - Learning from data (training phase)
    - Regression model is build by analysing or learning from a training data set made up of one or more independent variables and their dependent labels
    - Supervised learning: In supervised learning, each example is a *pair* consisting of an input example (independent variables) and a desired output value (dependent variable)
  - Step2: Using regression model for prediction
    - Testing phase
    - Predicting dependent variable
- Accuracy of a predictor:
  - How well a given predictor can predict for new values
- Target of learning techniques: Good generalization ability

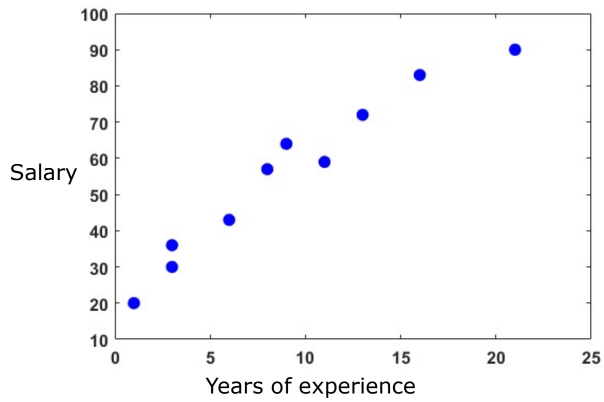
4

Illustration of Training Set: Salary Prediction

Years of experience (x)	Salary (in Rs 1000) (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Independent variable: Years of experience

Dependent variable: Salary

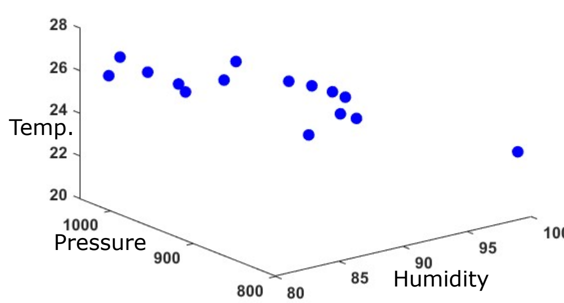


5

Illustration of Training Set: Temperature Prediction

Humidity (x_1)	Pressure (x_2)	Temp (y)
82.19	1036.35	25.47
83.15	1037.60	26.19
85.34	1037.89	25.17
87.69	1036.86	24.30
87.65	1027.83	24.07
95.95	1006.92	21.21
96.17	1006.57	23.49
98.59	1009.42	21.79
88.33	991.65	25.09
90.43	1009.66	25.39
94.54	1009.27	23.89
99.00	1009.80	22.51
98.00	1009.90	22.90
99.00	996.29	21.72
98.97	800.00	23.18

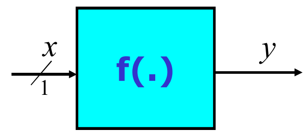
- Independent variable: Humidity, Pressure
- Dependent variable: Temperature (Temp)



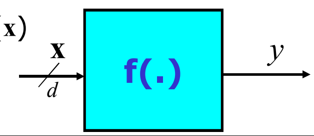
6

Linear Regression

- Linear approach to model the relationship between a scalar response, (y) (or dependent variable) and one or more predictor variables, (x *or* $\mathbf{x}$) (or independent variables)
- The response is going to be the linear function of input (one or more independent variables)
- Simple linear regression (straight-line regression):
  - Single independent variable (x)
  - Single dependent variable (y)
  - *Fitting a straight-line*



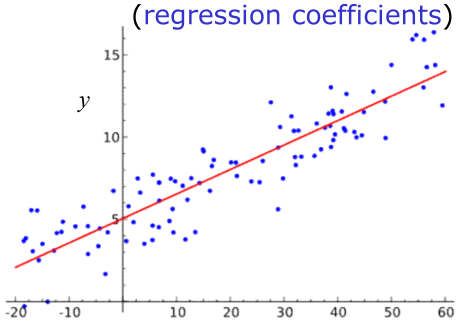
- Multiple linear regression:
  - One or more independent variable ($\mathbf{x}$)
  - Single dependent variable (y)
  - *Fitting a hyperplane*



7

Straight-Line (Simple Linear) Regression

- Given:- Training data: $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N},\ x_n \in \mathbb{R}^1$ and $y_n \in \mathbb{R}^1$
  - x_n: n^{th} input example (independent variable)
  - y_n: Dependent variable (output) corresponding to n^{th} independent variable
- Function governing the relationship between input and output: $y_n = f(x_n, w, w_0) = w\,x_n + w_0$
  - The coefficients w_0 and w are parameters of straight-line (regression coefficients) ***- Unknown***



- Function $f(x_n,w,w_0)$ is a linear function of x_n and it is a linear function of coefficients w and w_0
  - ***Linear model for regression***

8

Straight-Line (Simple Linear) Regression: Training Phase

- The values for the coefficients will be determined by fitting the linear function (straight-line) to the training data
- **Method of least squares**: Minimizes the squared error between the actual data (y_n) i.e. actual dependent variable and the estimate of line (predicted dependent variable ($\hat{y}_n$)) i.e. the function $f(x_n,w,w_0)$

$$\hat{y}_n = f(x_n, w, w_0) = w\, x_n + w_0$$

$$\underset{w,w_0}{\text{minimize}}\ \ E(w, w_0) = \frac{1}{2}\sum_{n=1}^{N}\left(\hat{y}_n - y_n\right)^2$$

- The derivatives of error function with respect to the coefficients will be linear in the elements of w and w_0
- Hence the minimization of the error function has unique solution and found in closed form

9

Straight-Line (Simple Linear) Regression: Training Phase

- Cost function for optimization:

$$E(w, w_0) = \frac{1}{2}\sum_{n=1}^{N}\left(f(x_n, w, w_0) - y_n\right)^2$$

- Conditions for optimality: $\frac{\partial\, E(w, w_0)}{\partial w} = 0$ $\quad \frac{\partial\, E(w, w_0)}{\partial w_0} = 0$

$$\frac{\partial\, \frac{1}{2}\sum_{n=1}^{N}\left(w\, x_n + w_0 - y_n\right)^2}{\partial w} = 0 \qquad \frac{\partial\, \frac{1}{2}\sum_{n=1}^{N}\left(w\, x_n + w_0 - y_n\right)^2}{\partial w_0} = 0$$

- Solving this give optimal $\hat{w}$ and $\hat{w}_0$ as

$$\boxed{\hat{w} = \frac{\sum_{n=1}^{N}\left(x_n - \mu_x\right)\left(y_n - \mu_y\right)}{\sum_{n=1}^{N}\left(x_n - \mu_x\right)^2}} \qquad \boxed{\hat{w}_0 = \mu_y - w\mu_x}$$

- μ_x: sample mean of independent variable x
- μ_y: sample mean of independent variable y

10

Straight-Line (Simple Linear) Regression: Testing

- For any test example x, the predicted value is given by:

$$\hat{y} = f(x, w, w_0) = \hat{w}\, x + \hat{w}_0$$

- The prediction accuracy is measured in terms of squared error:

$$E = (\hat{y} - y)^2$$

- Let N_t be the total number of test samples
- The prediction accuracy of regression model is measured in terms of root mean squared error:

$$E_{\text{RMS}} = \sqrt{\frac{1}{N_t}\sum_{n=1}^{N_t}(\hat{y}_n - y_n)^2}$$

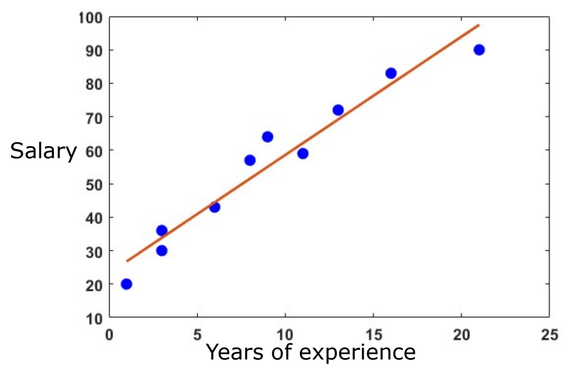
11

Illustration of Simple Linear Regression: Salary Prediction - Training

Years of experience (x)	Salary (in Rs 1000) (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

$$\hat{w} = \frac{\sum_{n=1}^{N}(x_n - \mu_x)(y_n - \mu_y)}{\sum_{n=1}^{N}(x_n - \mu_x)^2} \qquad \hat{w}_0 = \mu_y - w\mu_x$$

- μ_x: 9.1
- μ_y: 55.4
- $\hat{w}$: 3.54
- $\hat{w}_0$: 23.21

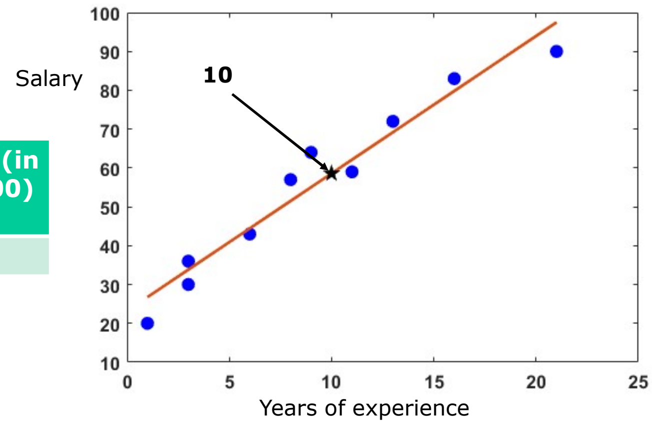


12

Illustration of Simple Linear Regression: Salary Prediction - Test

- $\hat{w}$: 3.54
- $\hat{w}_0$: 23.21

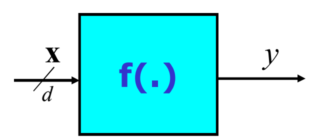
Years of experience (x)	Salary (in Rs 1000) (y)
10	-



- Predicted salary: 58.584
- Actual salary: 58.000
- Squared error: **0.34**

13

Multiple Linear Regression



- Multiple linear regression:
  - One or more independent variable ($\mathbf{x}$)
  - Single dependent variable (y)
- Given:- Training data: $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$, $\mathbf{x}_n \in \mathbb{R}^d$ and $y_n \in \mathbb{R}^1$
  - d: dimension of input example (number of independent variables)
- Function governing the relationship between input and output:

$$y_n = f(\mathbf{x}_n, \mathbf{w}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}_n + w_0 = \sum_{i=0}^{d} w_i x_i$$

  - The coefficients w_0, w_1, … , w_d are collectively denoted by the vector $\mathbf{w}$ ***- Unknown***
- Function $f(\mathbf{x}_n,\mathbf{w})$ is a linear function of $\mathbf{x}_n$ and it is a linear function of coefficients $\mathbf{w}$
  - ***Linear model for regression***

14

Linear Regression: Linear Function Approximation

- Linear function:
  - 2-dimensional space: The mapping function is a **line** specified by

$$f(\mathbf{x},\mathbf{w}) = w_1 x_1 + w_2 x_2 + w_0 = 0$$
$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{w_0}{w_2}$$

  - d-dimensional space: The mapping function is a **hyperplane** specified by

$$f(\mathbf{x},\mathbf{w}) = w_d x_d + \dots + w_2 x_2 + w_1 x_1 + w_0 = \sum_{i=0}^{d} w_i x_i = \mathbf{w}^\mathsf{T}\mathbf{x} = 0$$

where $\mathbf{w} = [w_0, w_1, \dots, w_d]^\mathsf{T}$ and $\mathbf{x} = [1, x_1, \dots, x_d]^\mathsf{T}$

15

Multiple Linear Regression

- The values for the coefficients will be determined by fitting the linear function to the training data
- **Method of least squares**: Minimizes the squared error between the actual data (y_n) i.e. actual dependent variable and predicted dependent variable ($\hat{y}_n$) i.e. the estimate linear function $f(\mathbf{x}_n, \mathbf{w})$, for any given value of $\mathbf{w}$

$$\hat{y}_n = f(\mathbf{x}_n, \mathbf{w}) = \mathbf{w}^\mathsf{T}\mathbf{x}_n + w_0 = \sum_{i=0}^{d} w_i x_i$$
$$\underset{\mathbf{w}}{\text{minimize}} \ E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \left(\hat{y}_n - y_n\right)^2$$

- The error function is a
  - quadratic function of the coefficients $\mathbf{w}$ and
  - The derivatives of error function with respect to the coefficients will be linear in the elements of $\mathbf{w}$
- Hence the minimization of the error function has unique solution and found in closed form

16

Multiple Linear Regression

- Cost function for optimization:
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(f(\mathbf{x}_n, \mathbf{w}) - y_n\right)^2$$
- Conditions for optimality: $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{0}$
- Application of optimality conditions gives optimal $\hat{\mathbf{w}}$:
$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{i=0}^{d} w_i x_{ni} - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$
$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^\top \mathbf{x}_n - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

17

Multiple Linear Regression

- Cost function for optimization:
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(f(\mathbf{x}_n, \mathbf{w}) - y_n\right)^2$$
- Conditions for optimality: $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{0}$
- Application of optimality conditions gives optimal $\hat{\mathbf{w}}$:
$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^\top \mathbf{x}_n - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$
$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{y}$$
  - Assumption: $d < N$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & x_{2d} \\ - & - & - & - & - \\ 1 & x_{n1} & x_{n2} & \dots & x_{nd} \\ - & - & - & - & - \\ 1 & x_{N1} & x_{N2} & \dots & x_{Nd} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ - \\ y_n \\ - \\ y_N \end{bmatrix}$$

$\mathbf{X}$ is data matrix

18

Multiple Linear Regression: Testing

- Optimal coefficient vector $\mathbf{w}$ is given by

  $\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$

  $\hat{\mathbf{w}} = \mathbf{X}^{+} \mathbf{y}$

  where $\mathbf{X}^{+} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$ is the pseudo inverse of matrix $\mathbf{X}$
- For any test example $\mathbf{x}$, the predicted value is given by:

$$\hat{y} = f(\mathbf{x}, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^\top \mathbf{x} = \sum_{i=0}^{d} \hat{w}_i x_i$$

- The prediction accuracy is measured in terms of squared error: $E = \left(\hat{y} - y\right)^2$
- Let N_t be the total number of test samples
- The prediction accuracy of regression model is measured in terms of root mean squared error:

$$E_{\text{RMS}} = \sqrt{\frac{1}{N_t} \sum_{n=1}^{N_t} \left(\hat{y}_n - y_n\right)^2}$$

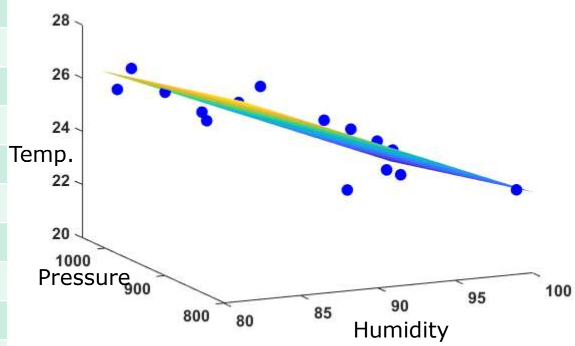
19

Illustration of Multiple Linear Regression: Temperature Prediction

Humidity (x_1)	Pressure (x_2)	Temp (y)
82.19	1036.35	25.47
83.15	1037.60	26.19
85.34	1037.89	25.17
87.69	1036.86	24.30
87.65	1027.83	24.07
95.95	1006.92	21.21
96.17	1006.57	23.49
98.59	1009.42	21.79
88.33	991.65	25.09
90.43	1009.66	25.39
94.54	1009.27	23.89
99.00	1009.80	22.51
98.00	1009.90	22.90
99.00	996.29	21.72
98.97	800.00	23.18

- Training:

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$



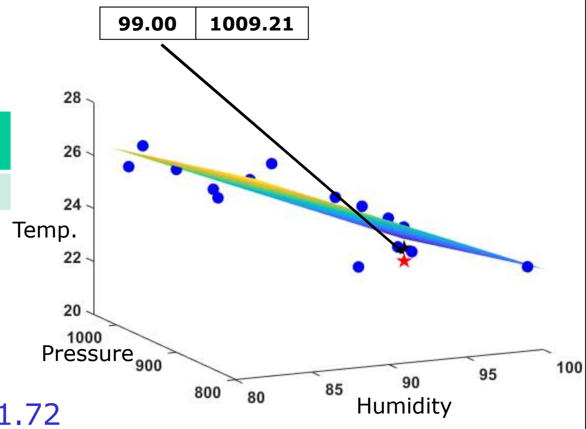
20

Illustration of Multiple Linear Regression: Temperature Prediction - Test

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Humidity (x_1)	Pressure (x_2)	Temp (y)
99.00	1009.21	-

$$y = f(\mathbf{x}, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^\top \mathbf{x}$$



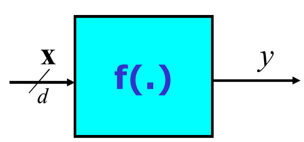
- Predicted rainfall: 21.72
- Actual rainfall: 21.24
- Squared error: **0.2347**

21

Application of Regression: A Method to Handle Missing Values

- Use most probable value to fill the missing value:
  - Use regression techniques to predict the missing value (regression imputation)
    - Let x_1, x_2, …, x_d be a set of d attributes
    - Regression (multivariate): The n^{th} value is predicted as

$$y_n = f(x_{n1}, x_{n2}, \ldots, x_{nd})$$



    - Simple or Multiple Linear regression: $y_n = w_1 x_{n1} + w_2 x_{n2} + \ldots + w_d x_{nd}$
    - Popular strategy
    - It uses the most information from the present data to predict the missing values
    - It preserves the relationship with other variables

Application of Regression: A Method to Handle Missing Values

- Training process:
  - Let y be the attribute, whose missing values to be predicted
  - Training examples: All $\mathbf{x}=[x_1, x_2, \ldots, x_d]^\mathsf{T}$, a set of d dependent attributes for which the independent variable y is available
  - The values for the coefficients will be determined by fitting the linear function to the training data

	Dates	Temperature	Humidity	Rain
2	08-07-2018	25.46875	82.1875	6.75
3	09-07-2018	26.19298	83.1491	1761.75
4	10-07-2018	25.17021	85.3404	652.5
5	11-07-2018	NaN	87.6866	963
6	12-07-2018	24.06923	87.6462	254.25
7	13-07-2018	21.20779	95.9481	339.75
8	15-07-2018	23.48571	96.1714	38.25
9	18-07-2018	NaN	98.5897	29.25
10	19-07-2018	25.09346	88.3271	4.5
11	20-07-2018	25.39423	90.4327	112.5
12	21-07-2018	NaN	94.5378	735.75
13	22-07-2018	22.5098	99	607.5
14	23-07-2018	22.904	98	717.75
15	24-07-2018	NaN	99	513
16	25-07-2018	23.18182	98.9697	195.75

- Dependent variable: Temperature
- Independent variables: Humidity and Rainfall

Application of Regression: A Method to Handle Missing Values

- Testing process (Prediction):
  - Optimal coefficient vector $\mathbf{w}$ is given by

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

  - For any test example $\mathbf{x}$, the predicted value is given by:

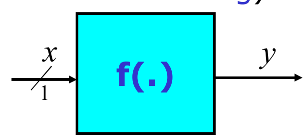
$$\hat{y} = f(\mathbf{x}, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^\mathsf{T}\mathbf{x} = \sum_{i=0}^{d} \hat{w}_i x_i$$

	Dates	Temperature	Humidity	Rain
2	08-07-2018	25.46875	82.1875	6.75
3	09-07-2018	26.19298	83.1491	1761.75
4	10-07-2018	25.17021	85.3404	652.5
5	11-07-2018	NaN	87.6866	963
6	12-07-2018	24.06923	87.6462	254.25
7	13-07-2018	21.20779	95.9481	339.75
8	15-07-2018	23.48571	96.1714	38.25
9	18-07-2018	NaN	98.5897	29.25
10	19-07-2018	25.09346	88.3271	4.5
11	20-07-2018	25.39423	90.4327	112.5
12	21-07-2018	NaN	94.5378	735.75
13	22-07-2018	22.5098	99	607.5
14	23-07-2018	22.904	98	717.75
15	24-07-2018	NaN	99	513
16	25-07-2018	23.18182	98.9697	195.75

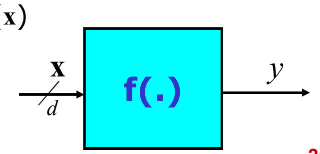
	Dates	Temperature	Humidity	Rain
2	08-07-2018	25.46875	82.1875	6.75
3	09-07-2018	26.19298	83.1491	1761.75
4	10-07-2018	25.17021	85.3404	652.5
5	11-07-2018	**24.2**	87.6866	963
6	12-07-2018	24.06923	87.6462	254.25
7	13-07-2018	21.20779	95.9481	339.75
8	15-07-2018	23.48571	96.1714	38.25
9	18-07-2018	**21.5**	98.5897	29.25
10	19-07-2018	25.09346	88.3271	4.5
11	20-07-2018	25.39423	90.4327	112.5
12	21-07-2018	**23.7**	94.5378	735.75
13	22-07-2018	22.5098	99	607.5
14	23-07-2018	22.904	98	717.75
15	24-07-2018	**21.6**	99	513
16	25-07-2018	23.18182	98.9697	195.75

Nonlinear Regression

- Nonlinear approach to model the relationship between a scalar response, (y) (or dependent variable) and one or more predictor variables, (x *or* $\mathbf{x}$) (or independent variables)
- The response is going to be the nonlinear function of input (one or more independent variables)
- Simple nonlinear regression (Polynomial curve fitting):
  - Single independent variable (x)
  - Single dependent variable (y)
  - *Fitting a curve*

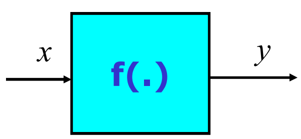


- Nonlinear regression (Polynomial regression):
  - One or more independent variable ($\mathbf{x}$)
  - Single dependent variable (y)
  - *Fitting a surface*



25

Polynomial Curve Fitting

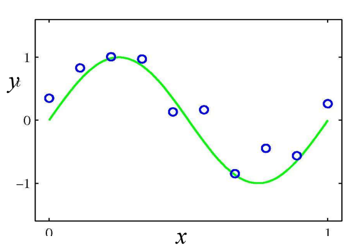


- Given:-Training data:

  $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N},\ x_n \in \mathbb{R}^1$ and $y_n \in \mathbb{R}^1$

- Function governing the relationship between input and output given by a polynomial function of degree p:

$$y_n = f(x_n, \mathbf{w}) = w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_p x_n^p = \sum_{j=0}^{p} w_j x_n^j$$



$y = f(x, \mathbf{w})$

- The coefficients $\mathbf{w}=[w_0, w_1, \ldots, w_p]$ are parameters of polynomial curve (regression coefficients)
  - *Unknown*
- Polynomial function $f(x_n,\mathbf{w})$ is a nonlinear function of x_n and it is a linear function of coefficients $\mathbf{w}$
  - ***Linear model for regression***

26

Polynomial Curve Fitting: Training Phase

- The values for the coefficients will be determined by fitting the polynomial curve to the training data
- **Method of least squares**: Minimizes the squared error between the actual data (y_n) i.e. actual dependent variable and the estimate of line (predicted dependent variable ($\hat{y}_n$) i.e. the function $f(x_n,\mathbf{w})$

$$\hat{y}_n = f(x_n, \mathbf{w}) = w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_p x_n^p$$

$$\underset{\mathbf{w}}{\text{minimize}} \;\; E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \left(\hat{y}_n - y_n\right)^2$$

- The error function is a quadratic function of the coefficients $\mathbf{w}$ and
- Derivatives of error function with respect to the coefficients will be linear in the elements of $\mathbf{w}$
- Hence the minimization of the error function has unique solution and found in closed form

27

Polynomial Curve Fitting: Training Phase

$$\hat{y}_n = f(x_n, \mathbf{w}) = w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_p x_n^p = \sum_{j=0}^{p} w_j x_n^j$$

- Lets consider:

$$\begin{array}{ccccc} x_n & x_n^2 & x_n^3 & & x_n^p \\ \downarrow & \downarrow & \downarrow & \cdots & \downarrow \\ z_{n1} & z_{n2} & z_{n3} & & z_{np} \end{array}$$

p is degree of polynomial

$$\hat{y}_n = f(\mathbf{z}_n, \mathbf{w}) = w_0 + w_1 z_{n1} + w_2 z_{n2} + \ldots + w_p z_{np}$$

$$\hat{y}_n = f(\mathbf{z}_n, \mathbf{w}) = \sum_{j=0}^{p} w_j z_{nj} = \mathbf{w}^\mathsf{T}\mathbf{z}_n$$

where $\mathbf{w} = [w_0, w_1, \ldots, w_p]^\mathsf{T}$ and $\mathbf{z}_n = [1, z_{n1}, \ldots, z_{np}]^\mathsf{T}$

28

Polynomial Curve Fitting: Training Phase

- Cost function for optimization:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(f(\mathbf{z}_n, \mathbf{w}) - y_n\right)^2$$

- Conditions for optimality: $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = 0$
- Application of optimality conditions gives optimal $\hat{\mathbf{w}}$:

$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{j=0}^{p} w_j z_{nj} - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^\top \mathbf{z}_n - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

29

Polynomial Curve Fitting: Training Phase

- Cost function for optimization:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(f(\mathbf{z}_n, \mathbf{w}) - y_n\right)^2$$

- Conditions for optimality: $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = 0$
- Application of optimality conditions gives optimal $\hat{\mathbf{w}}$:

$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^\top \mathbf{z}_n - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

$$\hat{\mathbf{w}} = \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1}\mathbf{Z}^\top \mathbf{y}$$

– Assumption: $p < N$

$\mathbf{Z}$ is Vandermonde matrix

$$\mathbf{Z} = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1p} \\ 1 & z_{21} & z_{22} & \cdots & z_{2p} \\ - & - & - & - & - \\ 1 & z_{n1} & z_{n2} & \cdots & z_{np} \\ - & - & - & - & - \\ 1 & z_{N1} & z_{N2} & \cdots & z_{Np} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ - \\ y_n \\ - \\ y_N \end{bmatrix}$$

where, $z_{nj} = x_n^j$

30

Polynomial Curve Fitting: Testing

- Optimal coefficient vector $\mathbf{w}$ is given by
  $$\hat{\mathbf{w}} = \left(\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{y}$$
  $$\hat{\mathbf{w}} = \mathbf{Z}^{+}\mathbf{y}$$
  where $\mathbf{Z}^{+} = \left(\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathsf{T}}$ is the pseudo inverse of matrix $\mathbf{Z}$
- For any test example x, the predicted value is given by:
  $$\hat{y} = f(x, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\mathsf{T}}\mathbf{z} = \sum_{j=0}^{p} \hat{w}_i x^j$$
- The prediction accuracy is measured in terms of squared error: $E = \left(\hat{y} - y\right)^2$
- Let N_t be the total number of test samples
- The prediction accuracy of regression model is measured in terms of root mean squared error:
  $$E_{\mathrm{RMS}} = \sqrt{\frac{1}{N_t}\sum_{n=1}^{N_t}\left(\hat{y}_n - y_n\right)^2}$$

31

Determining p, Degree of Polynomial

- This is determined experimentally
- Starting with p=1, test set is used to estimate the accuracy, in terms of error, of the regression model
- This process is repeated each time by incrementing p
- The regression model with p that gives the minimum error on test set may be selected

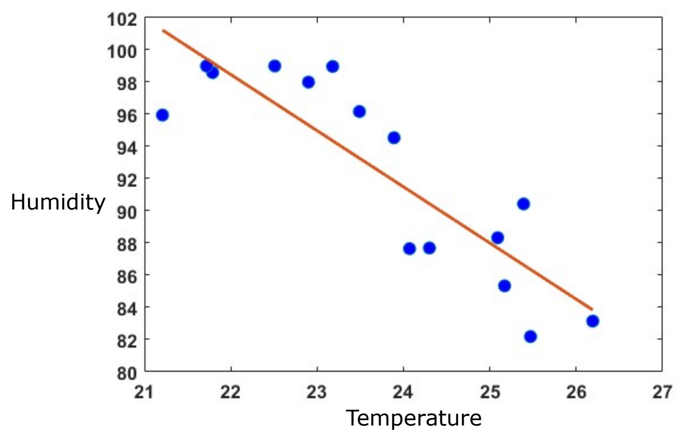
32

Illustration of Polynomial Curve Fitting: Humidity Prediction - Training

Temp (x)	Humidity (y)
25.47	82.19
26.19	83.15
25.17	85.34
24.30	87.69
24.07	87.65
21.21	95.95
23.49	96.17
21.79	98.59
25.09	88.33
25.39	90.43
23.89	94.54
22.51	99.00
22.90	98.00
21.72	99.00
23.18	98.97

- Degree of polynomial p : 1

$$\hat{\mathbf{w}} = \left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\mathbf{y}$$

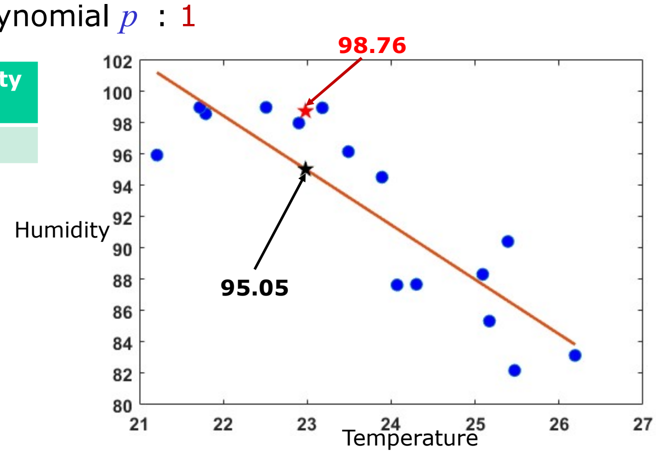


33

Illustration of Polynomial Curve Fitting: Humidity Prediction - Test

- Degree of polynomial p : 1

Temp (x)	Humidity (y)
22.98	--



- Predicted humidity: 95.05
- Actual humidity: 98.76
- Squared error: **13.77**

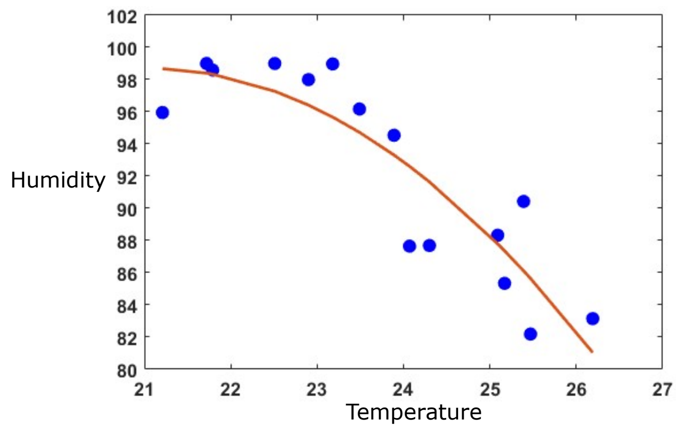
34

Illustration of Polynomial Curve Fitting: Humidity Prediction - Training

Temp (x)	Humidity (y)
25.47	82.19
26.19	83.15
25.17	85.34
24.30	87.69
24.07	87.65
21.21	95.95
23.49	96.17
21.79	98.59
25.09	88.33
25.39	90.43
23.89	94.54
22.51	99.00
22.90	98.00
21.72	99.00
23.18	98.97

- Degree of polynomial p : 2

$$\hat{\mathbf{w}} = \left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\mathbf{y}$$

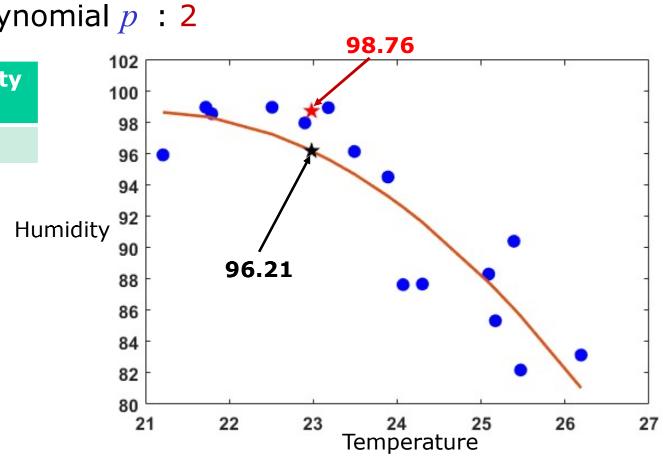


35

Illustration of Polynomial Curve Fitting: Humidity Prediction - Test

- Degree of polynomial p : 2

Temp (x)	Humidity (y)
22.98	--



- Predicted humidity: 96.21
- Actual humidity: 98.76
- Squared error: **06.49**

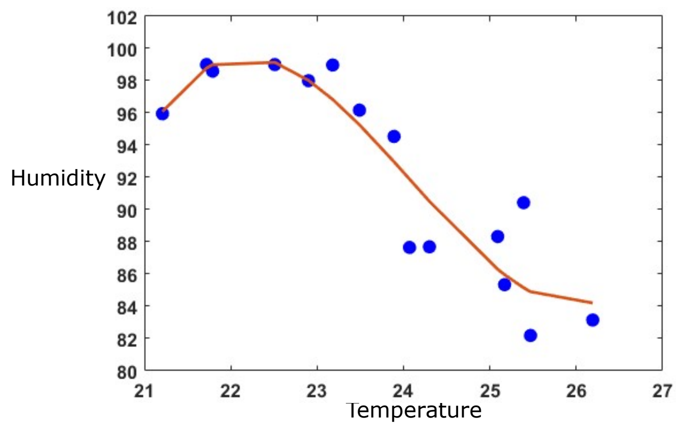
36

Illustration of Polynomial Curve Fitting: Humidity Prediction - Training

Temp (x)	Humidity (y)
25.47	82.19
26.19	83.15
25.17	85.34
24.30	87.69
24.07	87.65
21.21	95.95
23.49	96.17
21.79	98.59
25.09	88.33
25.39	90.43
23.89	94.54
22.51	99.00
22.90	98.00
21.72	99.00
23.18	98.97

- Degree of polynomial p : 3

$$\hat{\mathbf{w}} = \left(\mathbf{Z}^{\top}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\mathbf{y}$$

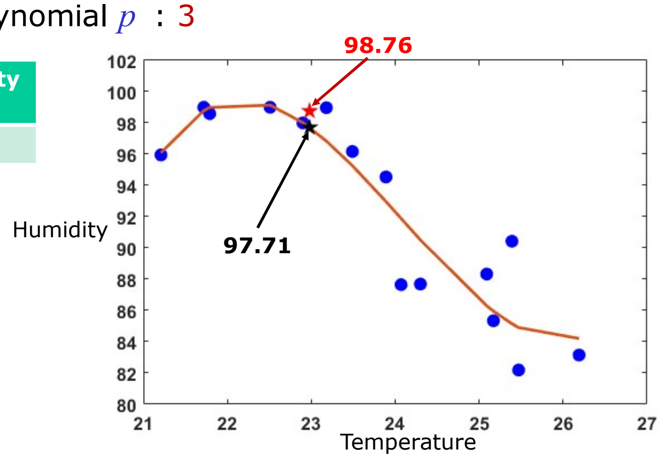


37

Illustration of Polynomial Curve Fitting: Humidity Prediction - Test

- Degree of polynomial p : 3

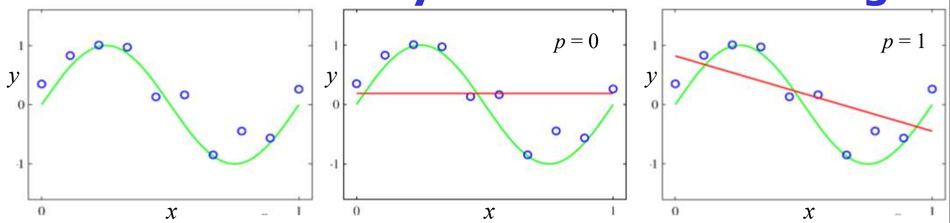
Temp (x)	Humidity (y)
22.98	--



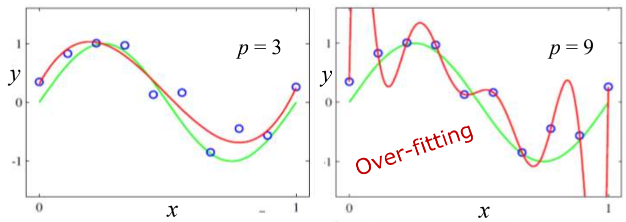
- Predicted humidity: 97.71
- Actual humidity: 98.76
- Squared error: **01.11**

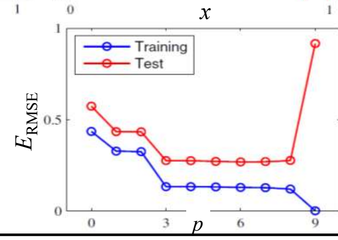
38

Illustration: Polynomial Curve Fitting



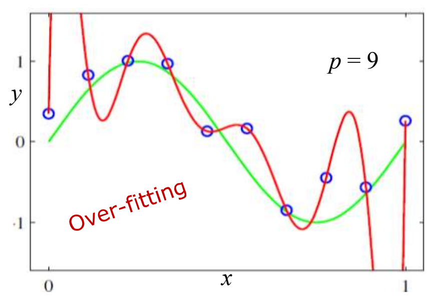
- Condition: Less number of training examples (N=10)
- Effect of increasing the degree of polynomial (p)

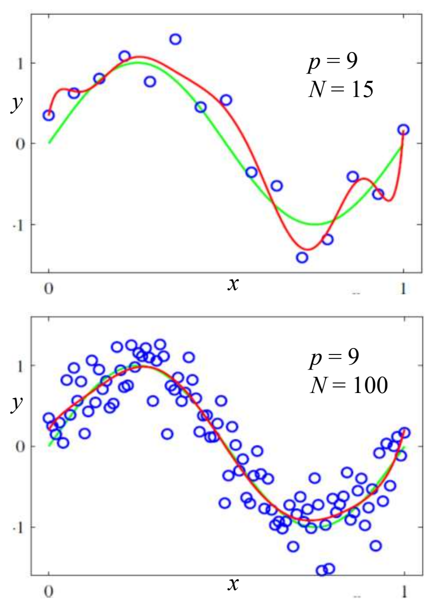




C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

Illustration: Polynomial Curve Fitting





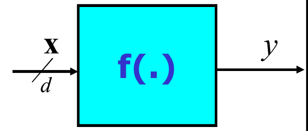
- Increasing the size of the data set reduces the over-fitting problem

C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

40

Nonlinear Regression: Polynomial Regression

- Polynomial regression:
  - One or more independent variable ($\mathbf{x}$)
  - Single dependent variable (y)



- Given:- Training data: $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}, \ \mathbf{x}_n \in \mathbb{R}^d$ and $y_n \in \mathbb{R}^1$
- Function governing the relationship between input and output given by a polynomial function of degree p:

$$y_n = f(\mathbf{x}_n, \mathbf{w}) = f(\boldsymbol{\varphi}(\mathbf{x}_n), \mathbf{w}) = \sum_{j=0}^{D-1} w_j \varphi_j(\mathbf{x}_n)$$

  - D is the number of monomials of polynomial up to degree p
  - $\varphi_j(\mathbf{x}_n)$ is the jth monomial of degree p for $\mathbf{x}_n$
- For 2-dimensional input, $\mathbf{x}_n = [x_{n1}, x_{n2}]^{\mathsf{T}}$ and degree, $p=2$

$$\boldsymbol{\varphi}(\mathbf{x}_n) = [\varphi_0(\mathbf{x}_n),\ \varphi_1(\mathbf{x}_n),\ \varphi_2(\mathbf{x}_n),\ \varphi_3(\mathbf{x}_n),\ \varphi_4(\mathbf{x}_n),\ \varphi_5(\mathbf{x}_n)]^{\mathsf{T}}$$

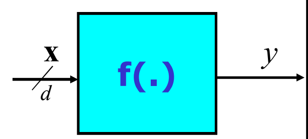
$$\boldsymbol{\varphi}(\mathbf{x}_n) = [\ 1,\ \ \sqrt{2}x_{n1},\ \ \sqrt{2}x_{n2},\ \ x_{n1}^2,\ \ x_{n2}^2,\ \ \sqrt{2}x_{n1}x_{n2}]^{\mathsf{T}}$$

$D = 6$

41

Nonlinear Regression: Polynomial Regression

- Polynomial regression:
  - One or more independent variable ($\mathbf{x}$)
  - Single dependent variable (y)



- Given:- Training data: $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}, \ \mathbf{x}_n \in \mathbb{R}^d$ and $y_n \in \mathbb{R}^1$
- Function governing the relationship between input and output given by a polynomial function of degree p:

$$y_n = f(\mathbf{x}_n, \mathbf{w}) = f(\boldsymbol{\varphi}(\mathbf{x}_n), \mathbf{w}) = \sum_{j=0}^{D-1} w_j \varphi_j(\mathbf{x}_n)$$

  - D is the number of monomials of polynomial up to degree p
  - $\varphi_j(\mathbf{x}_n)$ is the jth monomial of degree p for $\mathbf{x}_n$

The number of monomials D for the polynomial of degree p and the dimension of d is given by

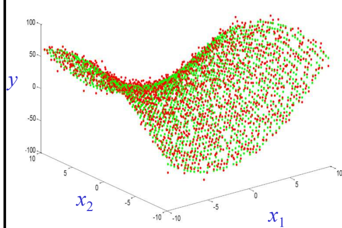
$$D = \frac{(d+p)!}{d!\,p!}$$

42

Nonlinear Regression: Polynomial Regression

- Given:- Training data: $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$, $\mathbf{x}_n \in \mathbb{R}^d$ and $y_n \in \mathbb{R}^1$
- Function governing the relationship between input and output given by a polynomial function of degree p:

$$y_n = f(\mathbf{x}_n, \mathbf{w}) = f(\boldsymbol{\varphi}(\mathbf{x}_n), \mathbf{w}) = \sum_{j=0}^{D-1} w_j \varphi_j(\mathbf{x}_n)$$



$y = f(\mathbf{x}_n, \mathbf{w})$

$\mathbf{x} = [x_1, x_2]^{\mathsf{T}}$

Fitting a surface

- The coefficients $\mathbf{w}=[w_0, w_1,\ldots, w_{D-1}]$ are parameters of surface (polynomial function) (regression coefficients) ***- Unknown***
- Polynomial function $f(\mathbf{x}_n,\mathbf{w})$ is a nonlinear function of $\mathbf{x}_n$ and it is a linear function of coefficients $\mathbf{w}$
  - ***Linear model for regression***

43

Nonlinear Regression: Polynomial Regression

- The values for the coefficients will be determined by fitting the polynomial to the training data
- **Method of least squares**: Minimizes the squared error between the actual data (y_n) i.e. actual dependent variable and the estimate of line (predicted dependent variable ($\hat{y}_n$) i.e. the function $f(x_n,\mathbf{w})$

$$\hat{y}_n = f(\mathbf{x}_n, \mathbf{w}) = f(\boldsymbol{\varphi}(\mathbf{x}_n), \mathbf{w}) = \sum_{j=0}^{D-1} w_j \varphi_j(\mathbf{x}_n)$$

$$\underset{\mathbf{w}}{\text{minimize}} \;\; E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \left(\hat{y}_n - y_n\right)^2$$

- The error function is a quadratic function of the coefficients $\mathbf{w}$
- Derivatives of error function with respect to the coefficients will be linear in the elements of $\mathbf{w}$
- Hence the minimization of the error function has unique solution and found in closed form

44

Polynomial Regression : Training Phase

$$\hat{y}_n = f(\mathbf{x}_n, \mathbf{w})$$

$$\hat{y}_n = f(\boldsymbol{\varphi}(\mathbf{x}_n), \mathbf{w})$$

$$\hat{y}_n = \sum_{j=0}^{D-1} w_j \varphi_j(\mathbf{x}_n)$$

$$\hat{y}_n = \mathbf{w}^{\mathsf{T}} \boldsymbol{\varphi}(\mathbf{x}_n)$$

where $\mathbf{w} = [w_0, w_1, ..., w_{D-1}]^{\mathsf{T}}$ and

$$\boldsymbol{\varphi}(\mathbf{x}_n) = [\varphi_0(\mathbf{x}_n), \varphi_1(\mathbf{x}_n), \varphi_2(\mathbf{x}_n), ..., \varphi_{D-1}(\mathbf{x}_n)]^{\mathsf{T}}$$

45

Polynomial Regression : Training Phase

- Cost function for optimization:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(f(\boldsymbol{\varphi}(\mathbf{x}_n), \mathbf{w}) - y_n \right)^2$$

- Conditions for optimality: $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{0}$

- Application of optimality conditions gives optimal $\hat{\mathbf{w}}$:

$$\frac{\partial \frac{1}{2} \sum_{n=1}^{N} \left(\sum_{j=0}^{D-1} w_j \varphi_j(\mathbf{x}_n) - y_n \right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

$$\frac{\partial \frac{1}{2} \sum_{n=1}^{N} \left(\mathbf{w}^{\mathsf{T}} \boldsymbol{\varphi}(\mathbf{x}_n) - y_n \right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

46

Polynomial Regression : Training Phase

- Cost function for optimization:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(f(\boldsymbol{\varphi}(\mathbf{x}_n), \mathbf{w}) - y_n\right)^2$$

- Conditions for optimality: $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = 0$
- Application of optimality conditions gives optimal $\hat{\mathbf{w}}$:

$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^{\top}\boldsymbol{\varphi}(\mathbf{x}_n) - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

$$\hat{\mathbf{w}} = \left(\mathbf{\Phi}^{\top}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\top}\mathbf{y}$$

  - Assumption: $D < N$

$$\mathbf{\Phi} = \begin{bmatrix} \varphi_0(\mathbf{x}_1) & \varphi_1(\mathbf{x}_1) & \dots & \varphi_{D-1}(\mathbf{x}_1) \\ \varphi_0(\mathbf{x}_2) & \varphi_1(\mathbf{x}_2) & \dots & \varphi_{D-1}(\mathbf{x}_2) \\ - & - & - & - \\ \varphi_0(\mathbf{x}_n) & \varphi_1(\mathbf{x}_n) & \dots & \varphi_{D-1}(\mathbf{x}_n) \\ - & - & - & - \\ \varphi_0(\mathbf{x}_N) & \varphi_1(\mathbf{x}_N) & \dots & \varphi_{D-1}(\mathbf{x}_N) \end{bmatrix}$$

47

Polynomial Regression: Testing

- Optimal coefficient vector $\mathbf{w}$ is given by

  $\hat{\mathbf{w}} = \left(\mathbf{\Phi}^{\top}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\top}\mathbf{y}$

  $\hat{\mathbf{w}} = \mathbf{\Phi}^{+}\mathbf{y}$

  where $\mathbf{\Phi}^{+} = \left(\mathbf{\Phi}^{\top}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\top}$ is the pseudo inverse of matrix $\mathbf{\Phi}$
- For any test example $\mathbf{x}$, the predicted value is given by:

$$\hat{y} = f(\mathbf{x}, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\top}\boldsymbol{\varphi}(\mathbf{x}) = \sum_{j=0}^{D-1} w_j \varphi_j(\mathbf{x})$$

- The prediction accuracy is measured in terms of squared error: $E = (\hat{y} - y)^2$
- Let N_t be the total number of test samples
- The prediction accuracy of regression model is measured in terms of root mean squared error:

$$E_{\text{RMS}} = \sqrt{\frac{1}{N_t}\sum_{n=1}^{N_t}(\hat{y}_n - y_n)^2}$$

48

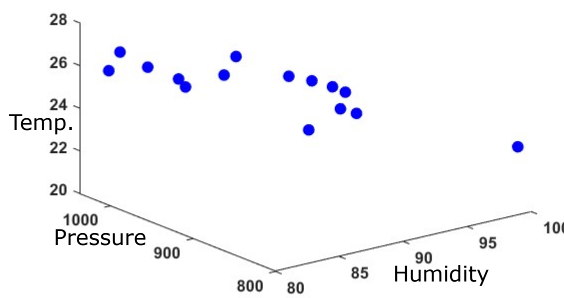
Determining p, Degree of Polynomial

- This is determined experimentally
- Starting with p=1, test set is used to estimate the accuracy, in terms of error, of the regression model
- This process is repeated each time by incrementing p
- The regression model with p that gives the minimum error on test set may be selected

49

Illustration of Polynomial Regression: Temperature Prediction

Humidity (x_1)	Pressure (x_2)	Temp (y)
82.19	1036.35	25.47
83.15	1037.60	26.19
85.34	1037.89	25.17
87.69	1036.86	24.30
87.65	1027.83	24.07
95.95	1006.92	21.21
96.17	1006.57	23.49
98.59	1009.42	21.79
88.33	991.65	25.09
90.43	1009.66	25.39
94.54	1009.27	23.89
99.00	1009.80	22.51
98.00	1009.90	22.90
99.00	996.29	21.72
98.97	800.00	23.18



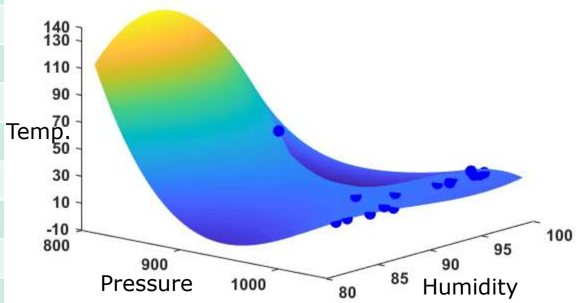
50

Illustration of Polynomial Regression: Temperature Prediction

Humidity (x_1)	Pressure (x_2)	Temp (y)
82.19	1036.35	25.47
83.15	1037.60	26.19
85.34	1037.89	25.17
87.69	1036.86	24.30
87.65	1027.83	24.07
95.95	1006.92	21.21
96.17	1006.57	23.49
98.59	1009.42	21.79
88.33	991.65	25.09
90.43	1009.66	25.39
94.54	1009.27	23.89
99.00	1009.80	22.51
98.00	1009.90	22.90
99.00	996.29	21.72
98.97	800.00	23.18

- **Training**:
- Polynomial Degree p = 3

$$\hat{\mathbf{w}} = \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \mathbf{y}$$



51

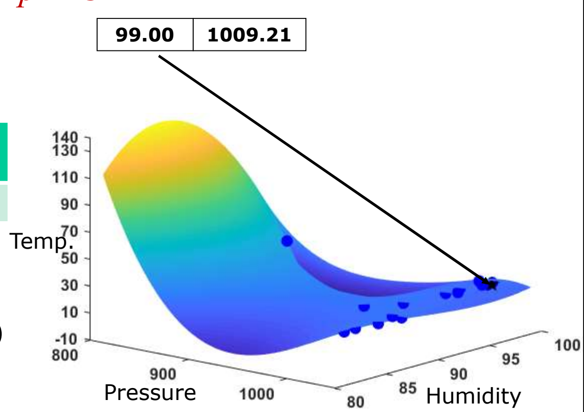
Illustration of Polynomial Regression: Temperature Prediction - Test

- Degree of polynomial p = 3

$$\hat{\mathbf{w}} = \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \mathbf{y}$$

Humidity (x_1)	Pressure (x_2)	Temp (y)
99.00	1009.21	-

$$\hat{y} = f(\mathbf{x}, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^\top \boldsymbol{\varphi}(\mathbf{x}) = \sum_{j=0}^{D-1} w_j \varphi_j(\mathbf{x})$$

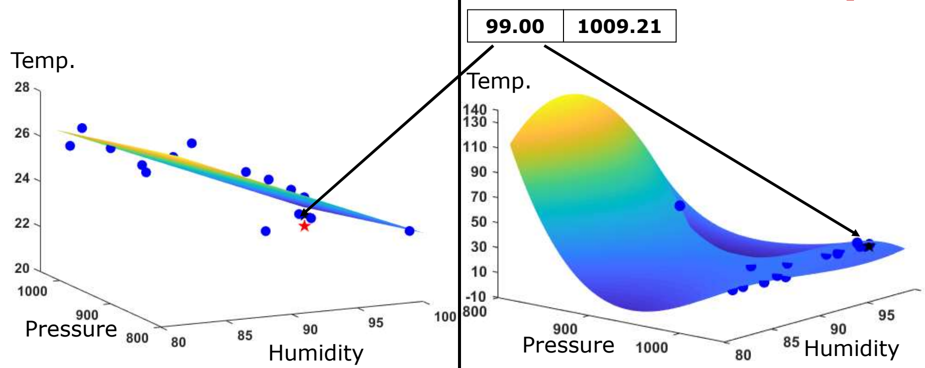


- Predicted Temperature: 21.05
- Actual Temperature: 21.24
- Squared error: **0.035**

52

Multiple Linear Regression vs Polynomial Regression Temperature Prediction

- Multiple Linear Regression
- Polynomial Regression
  - Degree of polynomial p = 3



- Predicted rainfall: 21.72
- Actual rainfall: 21.24
- Squared error: **0.2347**

- Predicted Temperature: 21.05
- Actual Temperature: 21.24
- Squared error: **0.035**

53

Autoregression (AR)

Autoregression (AR)

- Regression on the values of same attribute
- Autoregression is a time series model that
  - uses observations from previous time steps as input to a linear regression equation to predict the value at the next time step

Time Series Data

- Time series is a sequential set of data points, measured typically over successive times
- Time series data are simply a collection of observations gathered over time
- Time series data is given as:

$$\mathbf{X} = (x_1, x_2, \ldots, x_t, \ldots, x_T)$$

  - x_t is the observation at time t
  - T be the number of observations
- Example:
  - Weekly sales – time interval is week
  - Daily temperature in Kamand – time interval is day
- Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data
- Scope: We consider single variable x_t

Time Series Data and Dependence

- Time series data is given as:

$$\mathbf{X} = (x_1, x_2, \ldots, x_t, \ldots, x_T)$$

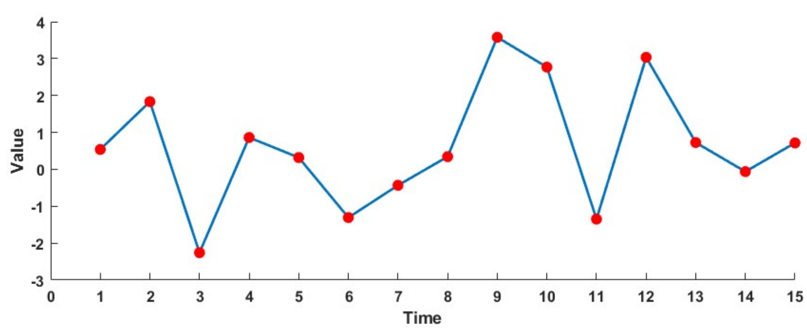
  - x_t is the observation at time t
  - T be the number of observations
- In time series data, value of each element at time t (x_t) is dependent on the values elements at previous p time steps (x_{t-1} , x_{t-2}, …, x_{t-p}) – p time lag

57

Time Series Data and Dependence

- Example: Data series in i.i.d
  - x_t is a random number drawn from $\mathcal{N}(0,1)$
- Each element at time t (x_t) is not dependent on the values elements at previous p time steps (x_{t-1} , x_{t-2}, …, x_{t-p}) – p time lag

0.54	1.83	-2.26	0.86	0.32	-1.31	-0.43	0.34	3.58	2.77	-1.35	3.03	0.73	-0.06	0.71

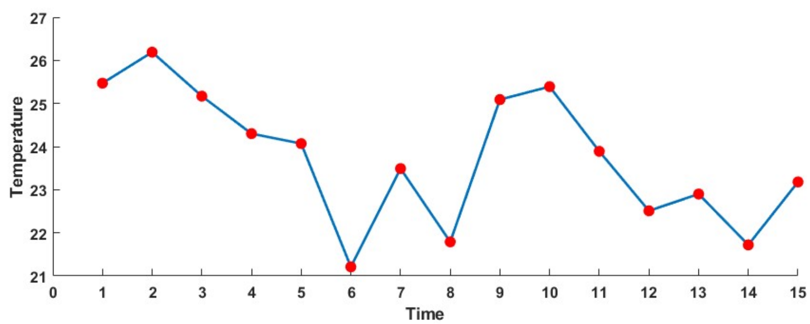


58

Time Series Data and Dependence

- Example: Daily temperature at Kamand
- Each element at time t (x_t) is dependent on the values elements at previous p time steps (x_{t-1} , x_{t-2}, …, x_{t-p}) – p time lag

25.47	26.19	25.17	24.3	24.07	21.21	23.49	21.79	25.09	25.39	23.89	22.51	22.9	21.72	23.18



59

Checking Dependency

- It's not always easy to just look at a time-series plot and say whether or not the series is independent
- x_t in a series is independent means that knowing previous values doesn't help you to predict the next value
  - Knowing x_{t-1} doesn't help to predict x_t
  - More generally, knowing x_{t-1}, x_{t-2}, …, x_{t-p} doesn't help to predict x_t
    - p is the number of previous time step (time lag)
- Dependency of each element at time t (x_t) with the values of elements at previous p time steps (x_{t-1} , x_{t-2}, …, x_{t-p}) is observed using autocorrelation

60

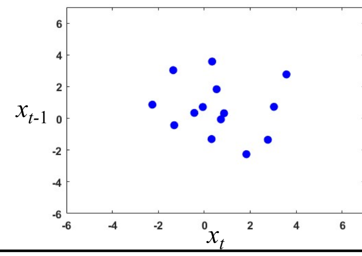
Checking Dependency - Autocorrelation

- The relationship between variables is called correlation
- Autocorrelation: The correlation calculated between the variable and itself at previous time steps
- Example: Data series in i.i.d
  - Autocorrelation between x_t and $x_{t\text{-}1}$ – Pearson correlation coefficient

x_t	0.54	1.83	-2.26	0.86	0.32	-1.31	-0.43	0.34	3.58	2.77	-1.35	3.03	0.73	-0.06	0.71
$x_{t\text{-}1}$		0.54	1.83	-2.26	0.86	0.32	-1.31	-0.43	0.34	3.58	2.77	-1.35	3.03	0.73	-0.06

  - Autocorrelation:

	x_t	$x_{t\text{-}1}$
x_t	1	**-0.1242**
$x_{t\text{-}1}$	**-0.1242**	1



61

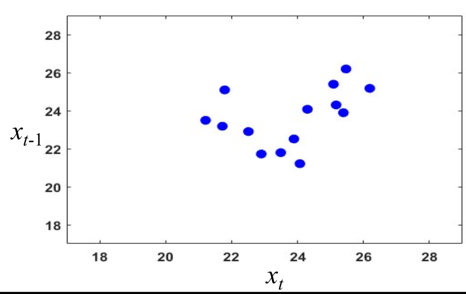
Checking Dependency - Autocorrelation

- The relationship between variables is called correlation
- Autocorrelation: The correlation calculated between the variable and itself at previous time steps
- Example: Daily temperature at Kamand
  - Autocorrelation between x_t and $x_{t\text{-}1}$

x_t	25.47	26.19	25.17	24.3	24.07	21.21	23.49	21.79	25.09	25.39	23.89	22.51	22.9	21.72	23.18
$x_{t\text{-}1}$		25.47	26.19	25.17	24.3	24.07	21.21	23.49	21.79	25.09	25.39	23.89	22.51	22.9	21.72

  - Autocorrelation:

	x_t	$x_{t\text{-}1}$
x_t	1	**0.4054**
$x_{t\text{-}1}$	**0.4054**	1



62

Autoregression (AR) Model

- Autoregression (AR) is a linear regression model that uses observations from previous time steps as input to predict the value at the next time step
- An autoregression (AR) model makes an assumption that the observations at previous time steps are useful to predict the value at the next time step
- The autocorrelation statistics help to choose which lag variables (p) will be useful in a model
- Interestingly, if all lag variables (x_{t-1} , x_{t-2}, …, x_{t-p}) show low or no correlation with the output variable (x_t), then it suggests that the time series problem may not be predictable
- This can be very useful when getting started on a new dataset

63

Autoregression (AR) Model

- AR(1) model: AR model using one time lag (p=1)
  - uses x_{t-1} i.e. value of previous time step to predict x_t
- Given: Time series data: $\mathbf{X} = (x_1, x_2, \ldots, x_t, \ldots, x_T)$
  - x_t is the observation at time t
  - T be the number of observations
- AR(1) model is given as: $x_t = f(x_i, w_0, w_1) = w_0 + w_1\, x_{t-1}$
  - The coefficients w_0 and w_1 are parameters of straight-line (regression coefficients) ***- Unknown***
- The regression coefficients are obtained as seen in simple linear regression (straight-line regression) using least square method

64

AR(1) Model - Training

- The regression coefficients are obtained as seen in simple linear regression (straight-line regression) using least square method
- Minimize the squared error between the actual data (x_t) at time t and the estimate of linear function (predicted variable ($\hat{x}_t$)) i.e. the function $f(x_{t-1}, w_0, w_1)$

$$\hat{x}_t = f(x_{t-1}, w_0, w_1) = w_0 + w_1 \, x_{t-1}$$

$$\underset{w, w_0}{\text{minimize}} \;\; E(w_0, w_1) = \frac{1}{2}\sum_{t=2}^{T}(\hat{x}_t - x_t)^2$$

- The optimal $\hat{w}_0$ and $\hat{w}_1$ is given as

$$\hat{w}_1 = \frac{\sum_{t=1}^{T}(x_{t-1}-\mu_{t-1})(x_t-\mu_t)}{\sum_{t=1}^{T}(x_{t-1}-\mu_{t-1})^2}$$

$$\hat{w}_0 = \mu_t - w_1 \mu_t$$

- μ_{t-1}: sample mean of variables at time t-1, x_{t-1}
- μ_t: sample mean of variables at time t, x_t

65

AR(1) Model: Testing

- For any test example at time t-1, x_{t-1}, the predicted value at time t, $\hat{x}_t$ is given by:

$$\hat{x}_t = f(x_{t-1}, w_0, w_1) = \hat{w}_0 + \hat{w}_1 \, x_{t-1}$$

- The prediction accuracy is measured in terms of squared error:

$$E = (\hat{x}_t - x_t)^2$$

- Let T_{test} be the total number of test samples
- The prediction accuracy of regression model is measured in terms of root mean squared error:

$$E_{\text{RMS}} = \sqrt{\frac{1}{T_{test}}\sum_{t=1}^{T_{test}}(\hat{x}_t - x_t)^2}$$

66

Autoregression Model

- AR(p) model: AR model using p time lags ($p < T$)
  - uses x_{t-1} , x_{t-2}, …, x_{t-p} i.e. value of previous p time step to predict x_t
- Given: Time series data: $\mathbf{X} = (x_1, x_2, \ldots, x_t, \ldots, x_T)$
  - x_t is the observation at time t
  - T be the number of observations
- AR(p) model is given as:

$$x_t = f(x_i, w_0, w_1, \ldots, w_p) = w_0 + w_1\, x_{t-1} + \ldots + w_p\, x_{t-p}$$

$$x_t = f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{p} w_j x_{t-j} = \mathbf{w}^{\mathsf{T}}\mathbf{x}$$

$$\text{where } \mathbf{w} = [w_0, w_1, \ldots, w_p]^{\mathsf{T}} \text{ and } \mathbf{x} = [1, x_{t-1}, x_{t-2}, \ldots, x_{t-p}]^{\mathsf{T}}$$

  - The coefficients w_0, w_1, …, w_p are parameters of hyperplane (regression coefficients) ***- Unknown***

67

AR (p) Model - Training

- The regression coefficients are obtained as seen in multiple linear regression with p input variables using least square method
- Minimize the squared error between the actual data (x_t) at time t and the estimate of linear function (predicted variable ($\hat{x}_t$)) i.e. the function $f(\mathbf{x},\mathbf{w})$

$$\hat{x}_t = f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{p} w_j x_{t-j} = w_0 + \mathbf{w}^{\mathsf{T}}\mathbf{x}$$

$$\underset{\mathbf{w}}{\text{minimize}} \;\; E(\mathbf{w}) = \frac{1}{2}\sum_{t=p+1}^{T} (\hat{x}_t - x_t)^2$$

- The autocorrelation statistics help to choose which lag variables (p) will be useful in a model

68

AR (p) Model - Training

- The optimal $\hat{\mathbf{w}}$ is given as $\hat{\mathbf{w}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{x}^{(t)}$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{t-1} & x_{t-2} & \dots & x_{t-p} \\ 1 & x_t & x_{t-1} & \dots & x_{(t+1)-p} \\ - & - & - & - & - \\ 1 & x_{t+n-1} & x_{t+n-2} & \dots & x_{(t+n)-p} \\ - & - & - & - & - \\ 1 & x_{T-1} & x_{T-2} & \dots & x_{T-p} \end{bmatrix} \quad \mathbf{x}^{(t)} = \begin{bmatrix} x_t \\ x_{t+1} \\ - \\ x_{t+n} \\ - \\ x_T \end{bmatrix}$$

$\mathbf{X}$ is data matrix with time lag

- The autocorrelation statistics help to choose which lag variables (p) will be useful in a model

AR (p) Model: Testing

- The value at time t, $\hat{x}_t$ is predicted by taking values from past p time steps (x_{t-1} , x_{t-2}, …, x_{t-p}) as input:

$$\hat{x}_t = f(\mathbf{x}, \hat{\mathbf{w}}) = \hat{w}_0 + \sum_{j=1}^{p} \hat{w}_j x_{t-j} = \hat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}$$

- The prediction accuracy is measured in terms of squared error:

$$E = \left(\hat{x}_t - x_t\right)^2$$

- Let T_{test} be the total number of test samples
- The prediction accuracy of regression model is measured in terms of root mean squared error:

$$E_{\text{RMS}} = \sqrt{\frac{1}{T_{test}} \sum_{t=1}^{T_{test}} \left(\hat{x}_t - x_t\right)^2}$$

Illustration AR(1) Model – Prediction of Temperature: Training

Date	Temp (x_t)
Sept 1	25.47
Sept 2	26.19
Sept 3	25.17
Sept 4	24.30
Sept 5	24.07
Sept 6	21.21
Sept 7	23.49
Sept 8	21.79
Sept 9	25.09
Sept 10	25.39
---	---
Oct 29	23.06
Oct 30	23.72
Oct 31	23.02

- T, the number of observations = 61

71

Illustration AR(1) Model – Prediction of Temperature: Training

Date	Temp (x_{t-1})	Temp (x_t)
Sept 1		25.47
Sept 2	25.47	26.19
Sept 3	26.19	25.17
Sept 4	25.17	24.30
Sept 5	24.30	24.07
Sept 6	24.07	21.21
Sept 7	21.21	23.49
Sept 8	23.49	21.79
Sept 9	21.79	25.09
Sept 10	25.09	25.39
---	---	---
Oct 29	22.76	23.06
Oct 30	23.06	23.72
Oct 31	23.72	23.02

- T, the number of observations = 60

$$\hat{w}_1 = \frac{\sum_{t=1}^{60}(x_{t-1}-\mu_{t-1})(x_t-\mu_t)}{\sum_{t=1}^{60}(x_{t-1}-\mu_{t-1})^2}$$

$$\hat{w}_0 = \mu_t - w_1\mu_{t-1}$$

- μ_{t-1}: 22.81
- $\hat{w}_1$: 0.523
- μ_t : 22.85
- $\hat{w}_0$: 10.861

72

Illustration AR(1) Model – Prediction of Temperature: Test

- $\hat{w}_1$: 0.523
- $\hat{w}_0$: 10.861

Date	Temp (x_{t-1})	Temp (x_t)
Nov 2	22.30	-

- Predicted Temperature for Nov 2 : 22.52
- Actual Temperature on Nov 2 : 21.43
- Squared error : **1.19**

73

Illustration AR(p) Model – Prediction of Temperature: Checking Dependency

Date	Temp (x_t)
Sept 1	25.47
Sept 2	26.19
Sept 3	25.17
Sept 4	24.30
Sept 5	24.07
Sept 6	21.21
Sept 7	23.49
Sept 8	21.79
Sept 9	25.09
---	---
Oct 28	22.76
Oct 29	23.06
Oct 30	23.72
Oct 31	23.02

- p = 3
- T, the number of observations = 61

74

Illustration AR(p) Model – Prediction of Temperature: Checking Dependency

Date	Temp (x_{t-3})	Temp (x_{t-2})	Temp (x_{t-1})	Temp (x_t)
Sept 1				25.47
Sept 2			25.47	26.19
Sept 3		25.47	26.19	25.17
Sept 4	25.47	26.19	25.17	24.30
Sept 5	26.19	25.17	24.30	24.07
Sept 6	25.17	24.30	24.07	21.21
Sept 7	24.30	24.07	21.21	23.49
Sept 8	24.07	21.21	23.49	21.79
Sept 9	21.21	23.49	21.79	25.09
---	---	---	---	---
Oct 28	22.83	23.98	24.47	22.76
Oct 29	23.98	24.47	22.76	23.06
Oct 30	24.47	22.76	23.06	23.72
Oct 31	22.76	23.06	23.72	23.02

- p = 3
- T, the number of observations = 61
- Autocorrelation between x_t and x_{t-1} : 0.54
- Autocorrelation between x_t and x_{t-2} : 0.25
- Autocorrelation between x_t and x_{t-3} : -0.08
- An autocorrelation is deemed significant if

$$|\text{autocorrelation}| > \frac{2}{\sqrt{T}} = 0.25$$

- Time lag p=2 is sufficient as x_t is significant with x_{t-1} and x_{t-2}

75

Illustration AR(p) Model – Prediction of Temperature: Training

Date	Temp (x_{t-2})	Temp (x_{t-1})	Temp (x_t)
Sept 1			25.47
Sept 2		25.47	26.19
Sept 3	25.47	26.19	25.17
Sept 4	26.19	25.17	24.30
Sept 5	25.17	24.30	24.07
Sept 6	24.30	24.07	21.21
Sept 7	24.07	21.21	23.49
Sept 8	21.21	23.49	21.79
Sept 9	23.49	21.79	25.09
---	---	---	---
Oct 28	23.98	24.47	22.76
Oct 29	24.47	22.76	23.06
Oct 30	22.76	23.06	23.72
Oct 31	23.06	23.72	23.02

- p = 2
- T, the number of observations = 59
- Multiple linear regression with number of input variables = 2

$$\hat{\mathbf{w}} = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{x}^{(t)} \; ; \quad \hat{\mathbf{w}} \in \mathbf{R}^{3}$$

76

Illustration AR(p) Model – Prediction of Temperature: Test

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{x}^{(t)} \; ; \quad \hat{\mathbf{w}} \in \mathbf{R}^3$$

Date	Temp (x_{t-2})	Temp (x_{t-1})	Temp (x_t)
Nov 2	23.02	22.30	--

- Predicted Temperature for Nov 2 : 22.49
- Actual Temperature on Nov 2 : 21.43
- Squared error : **1.13**

77

Summary: Regression

- Regression analysis is used to model the relationship between one or more independent (predictor) variable and a dependent (response) variable
- Response is some function of one or more input variables
- Linear regression: Response is linear function of one or more input variables
- Nonlinear regression: Response is nonlinear function of one or more input variables
  - Polynomial regression: Response is nonlinear function approximated using polynomial function upto degree p of one or more input variables

78

Summary: Regression

- Autoregression (AR): Regression on the values of same attribute
  - It is a time series model
  - Linear regression model that uses observations from previous p time steps as input to predict the value at the next time step
  - It makes an assumption that the observations at previous time steps are useful to predict the value at the next time step
  - The autocorrelation statistics help to choose which lag variables (p) will be useful in a model
- AR model can be performed on time series data with single variable or with multiple variables
- In this course we are limited only on the time series data with single variable

Text Books

1. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Third Edition, Morgan Kaufmann Publishers, 2011.

2. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.