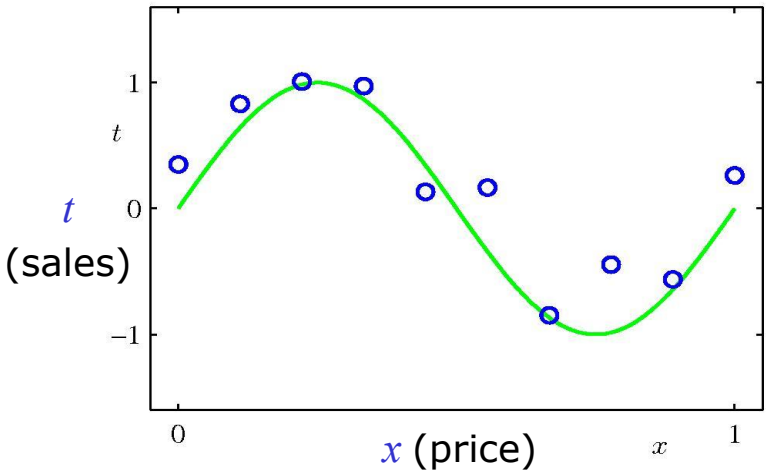
Supervised Machine Learning: Regression

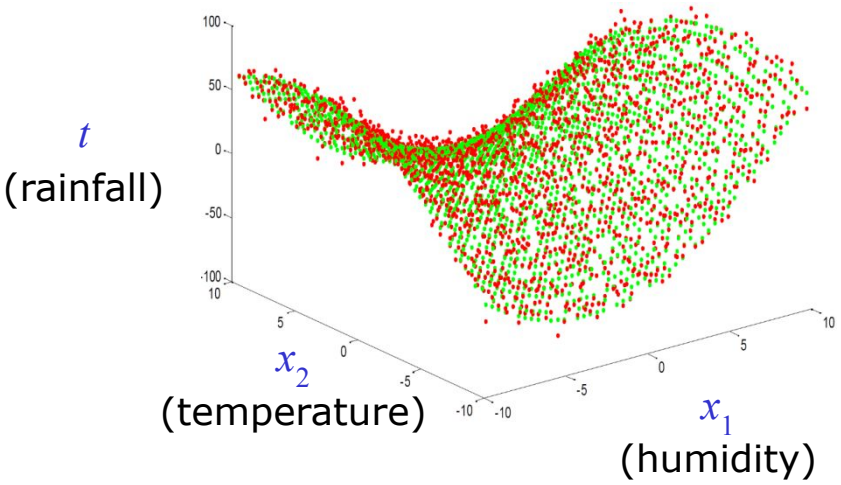
Numeric Prediction (Regression)

- Numeric prediction: Task of predicting continuous (or ordered) values for given input
- Example:
  - Predicting potential sales of a new product given its price



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

  - Predicting amount of rainfall given the temperature and humidity in the atmosphere



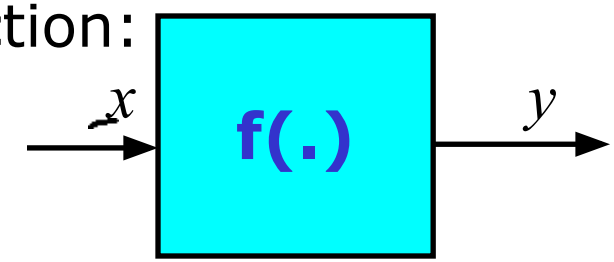
$$t = \mathrm{f}(\mathbf{x})$$

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

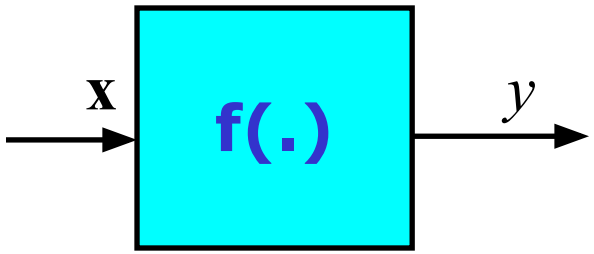
Numeric Prediction (Regression)

- Regression analysis is used to model the relationship between one or more independent (input) variable and a dependent (output) variable
  - Dependent variable is always continuous valued or ordered valued
  - Example: Dependent variable: Rainfall

    Independent variable(s): temperature, humidity
- The values of independent variables are known
- The dependent 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 ($\mathbf{x}$) $\in R^d$
- Single dependent variable (y)

Numeric Prediction (Regression)

- Regression is a two step process
  - Step1: Building a regression model
    - Learning from data (training phase)
    - Supervised learning: In supervised learning, each example is a *pair* consisting of an input example (independent variables) and a desired output value (dependent variable)
    - Regression model is built by analysing or learning from a training data set made up of one or more independent variables and their dependent labels

$$y_n = f(\mathbf{x}_n)$$

      - $\mathbf{x}_n$ is the n^{th} input example and y_n is the corresponding output 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

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

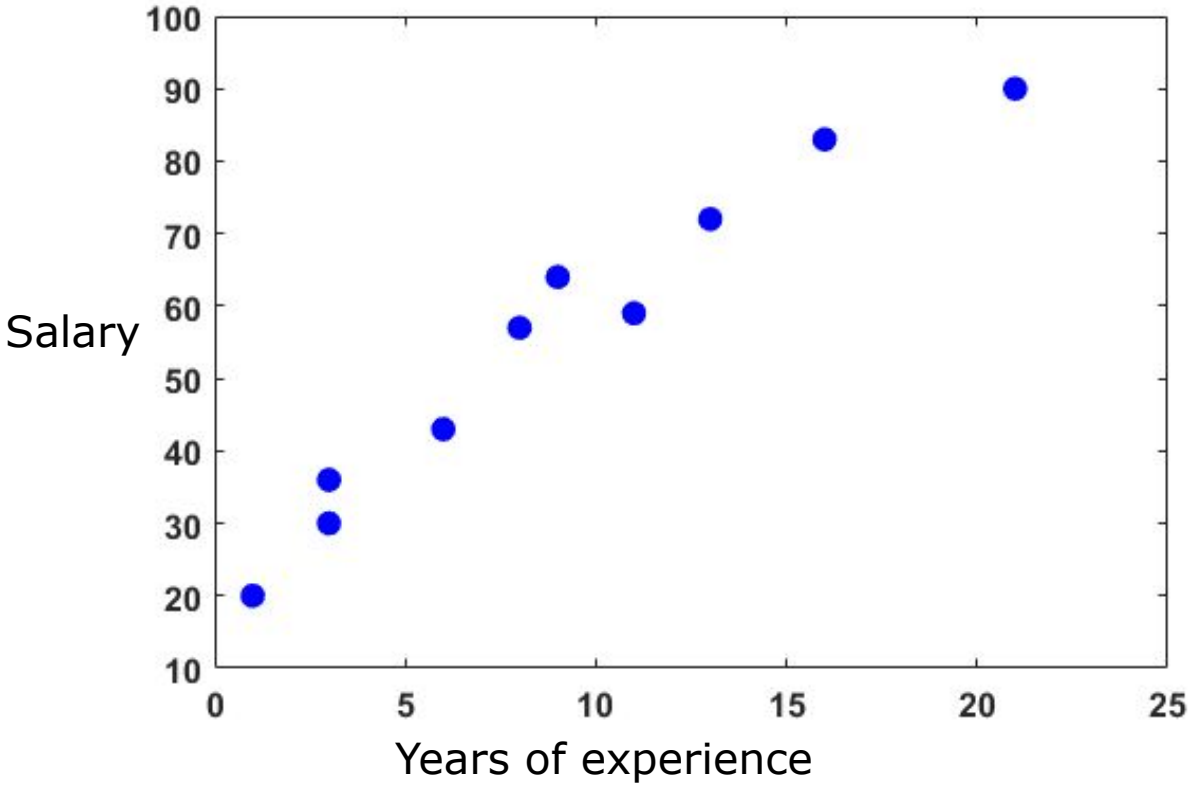


Illustration of Training Set: Temperature Prediction

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)

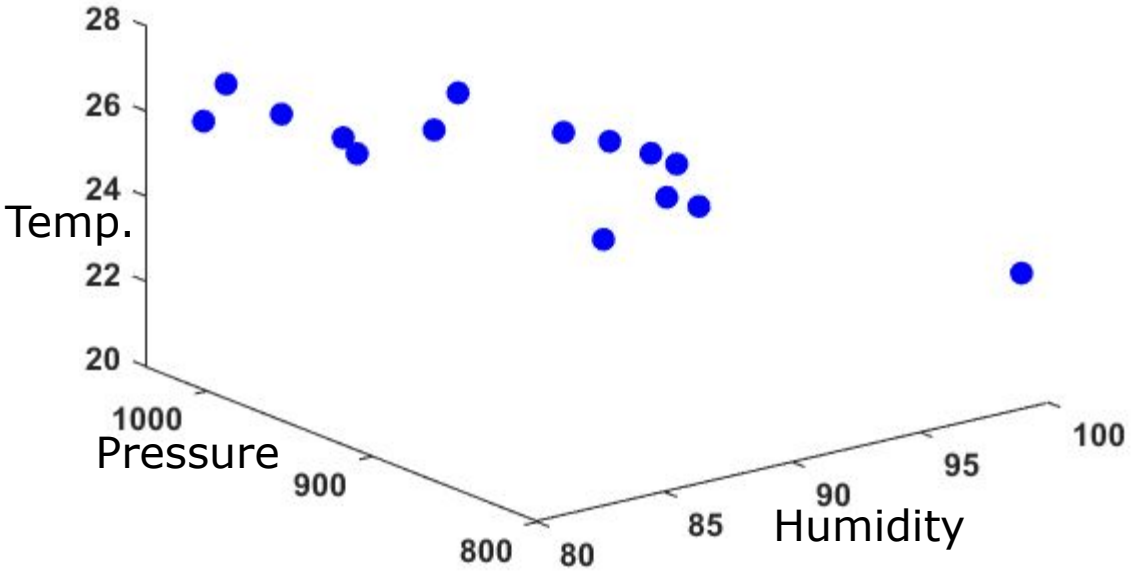


Illustration of Training Set: Wine Quality Prediction [1]

Fixed Acidity (x_1)	Volatile Acidity (x_2)	Citric acid (x_3)	Residual Sugar (x_4)	Chlorides (x_5)	Free SO_2 (x_6)	Total SO_2 (x_7)	Density (x_8)	pH (x_8)	Sulphates (x_9)	Alcohol (x_{10})	Quality (y)
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5.42
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5.57
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5.17
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6.65
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5.68
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5.63
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5.32
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7.16
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7.2
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5.18

- Number of independent variable: 10
- Dependent variable: Quality

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. "*Modeling wine preferences by data mining from physicochemical properties*," In Decision Support Systems, Elsevier, vol. 47, issue 4, pp. 547-553, 2009.

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.