Introduction to Fundamentals of Machine learning

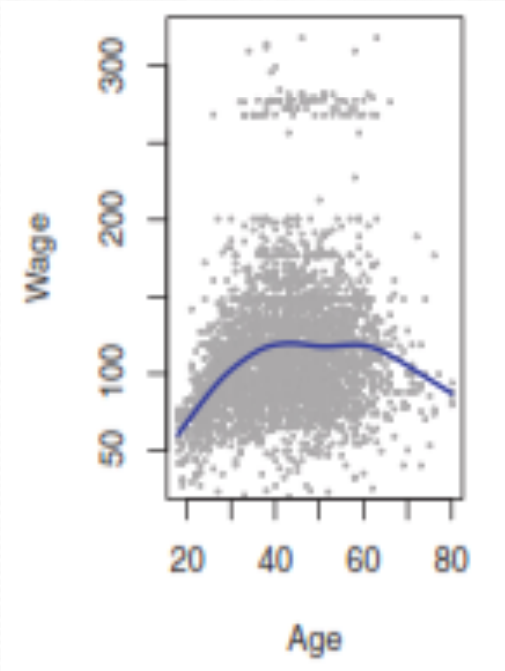
Introduction to some basic terms

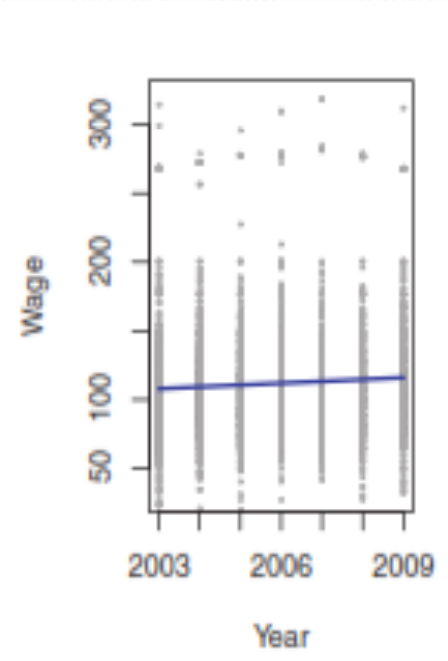
- Feature Vector and Matrix : An array of elements which are quantitative , qualitative, or well defined, qualitative features. e.g. a column vector $\underline{X}$
- $\underline{X}$ ={Author, Paper, Publication}
- Supervised learning: Supervised learning is the Data mining task of inferring a function from **labeled training data**.The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the **supervisory signal**). A **supervised learning algorithm** analyzes the training data and produces an inferred function, which can be used for mapping new examples.

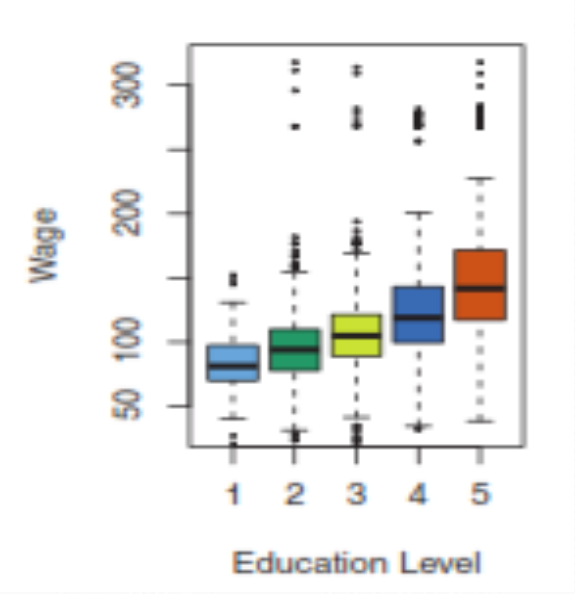
- Example of supervised learning :Classifying different fruits
- Unsupervised learning: **Unsupervised learning** is that of trying to find hidden structure in unlabeled data.
- Example: clustering of fruits into some no. of categories

Wage data

- Different factors that relate to wages for a particular place are: age , education , year







Variation of wage vs age, year, education level

- Left: wage as a function of age. On average, wage increases with age until about 60 years of age, at which point it begins to decline.
- Center: wage as a function of year . There is a slow but steady increase of approximately $10,000 in the average wage between 2003and 2009.
- Right: Boxplots displaying wage as a function of education,with 1 indicating the lowest level (no high school diploma) and 5 the highest level (an advanced graduate degree). On average, wage increases with the level of education.

- Given an employee's age, we can use this curve to predict his wage . From the figures there is a significant amount of variability associated with this average value, so age alone is unlikely to provide an accurate prediction of a particular man's wage.
- The most accurate prediction of a man's wage will be obtained by combining his age , his education , the year.

Other examples

- In the **stock market** data to predict whether index will increase or decrease on given day using the past 5 days' percentage changes in the index that is to predict whether Smarket performance will fall into the Up bucket or the down bucket. This is a classification problem.
- **Gene Expression data**: Consider the NCI60 data set, which consists of 6,830 gene expression measurements for each of 64 cancer cell lines. Instead of predicting a particular output variable, we are interested in determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements. This is a clustering problem.

Notations and Simple Matrix Algebra

- *n* to represent the number of distinct data points
- *p* denote the number of variables that are available for use in making predictions
- x_{ij} represent the value of the jth variable for the ith observation, where i = 1, 2, . . ., n and j = 1, 2, . . . , p

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

- The matrix X can be written as

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{pmatrix},$$

- Or

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

- y_i to denote the ith observation of the variable on which we wish to make predictions. set of all *n* observations in vector form as:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

- our observed data consists of {(x1, y1), (x2, y2), . . . , (xn, yn)}