Lazy Learning

Decision tree induction, Bayesian classification, classification by backpropagation are all examples of eager learning.

Eager learning → when given a set of training tuples, will constructed a classification model before receiving test tuples to classify.

Lazy learner → when given a training tuple, it simply stores it and waits until it is given a test tuple. They also known as instance based learners. Ex → K-NN classifier.

prediction

It is the task of predicting continuous values for a given input.

The most widely used approach for numeric prediction is Regression.

It is a statistical methodology, developed by Sir Francis Galton (1822-1911), a mathematician.

Regression analysis can be used to model the relationship between one or more independent or predictor variables and a dependent or response variable.

predictor variables are attributes of interest describing the tuple. The values are known.

The response variable is what we want to predict.

Given a tuple described by predictor variables, we want to predict the associated value of the response variable.

Types of Regression

(1) Linear Regression →

It involves a response variable, y and a single predictor variable, x.

It is the simplest form of regression and models y as a linear function of x.

i.e. $y = b + wx$ ——— (1)

b and w are regression coefficients.

The regression coefficients can also be thought of as weights, so we can write.

$y = w_0 + w_1 x$ ——— (2).

These coefficients can be solved by the method of least squares, which estimates the best-fitting straight line as the one that minimizes the error between the actual data and the estimate of the line.

Let D be a training set consisting of values of predictive variable, x, for some population and their associated values for response variable y.

The training set contains |D| data points of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_{|D|}, y_{|D|})$.

The regression coefficients can be estimated using the following eqns.

$$w_1 = \frac{\sum_{i=1}^{|D|} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|} (x_i - \bar{x})^2} \quad \text{---} (3)$$

$$w_0 = \bar{y} - w_1 \bar{x} \quad \text{---} (4)$$

where $\bar{x}$ = mean of $x_1, x_2, \ldots, x_{|D|}$

$\bar{y}$ = " " $y_1, y_2, \ldots, y_{|D|}$

Ex: Salary data.

x, years of experience.	y, salary (in 3K)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

we model the relationship that salary may be related to the no. of years of experience with the eqn

$$y = w_0 + w_1 x.$$

From the data, $\bar{x} = 9.1$

$\bar{y} = 55.4$

substituting these values into eqn (3) & (4)

we get

$$w_1 = \frac{(3-9.1)(30-55.4) + (8-9.1)(57-55.4) + \dots + (16-9.1)(83-55.4)}{(3-9.1)^2 + (8-9.1)^2 + \dots + (16-9.1)^2}$$

$$= 3.5.$$

$$w_0 = 55.4 - (3.5)(9.1) = 23.6.$$

Hence, $y = 23.6 + 3.5x.$

The salary of a person with 10 years of experience can be predicted as

$$y = 23.6 + 3.5 \times 10 = \$58,600$$

Multiple linear regression

It is an extension of straight-line regression so as to involve more than one predictor variable.

It allows response variable, y to be modeled as a linear function of n predictor variables or attributes, describing a tuple X.

i.e. $X = (x_1, x_2, \ldots x_n)$

our training data set, D, contains data of the form

$(X_1, y_1), (X_2, y_2) \ldots (X_{|D|}, y_{|D|})$.

Where X_i are the n-dimensional training tuples with associated response variables y_i.

Ex → A multiple linear regression model based on two predictor attributes or variables, A_1 and A_2 is

$$y = w_0 + w_1 x_1 + w_2 x_2$$

where x_1 and x_2 are the values of attributes A_1 and A_2 respectively in X.

Nonlinear Regression

of a given response variable and predictor variable have a relationship which is nonlinear or modeled by a polynomial function.

polynomial regression → is used when there is just one predictor variable. It can be modeled by adding the polynomial terms to the basic linear model. By applying transformations to the variables, we can convent the nonlinear model into a linear one that can be solved by the method of least squares.

Transformation of a polynomial regression model to a linear regression model.

Let a cubic polynomial is given by

$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

Let $x_1 = x$

$x_2 = x^2$

$x_3 = x^3$.

So, we get $y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$ which is easily solved by the method of least squares.

Ex

x mid-term exam	y final exam
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74
81	90

$|D| = 12$

$\bar{x} = 866/12 = 72.167$

$\bar{y} = 888/12 = 74$

$w_1 = 0.5816$

$w_0 = 32.028$

$y = 32.028 + 0.5816 x$

$y = 32.028 + 0.5816(86)$

$= 82.045 \approx 82$ on the final exam.

(1) Plot the data. Do x and y have a linear relationship?

(2) Use the method of least squares to find an eqn for the prediction of a student's final exam grade based on the student's midterm grade in the course.

(3) Predict the final exam grade of a student who received an 86 on the mid-term exam.