Artificial Intelligence for Medicine II

Spring 2025

Lecture 71: Supervised Learning
Regression

(Many slides adapted from Bing Liu, Han, Kamber & Pei; Tan, Steinbach, Kumar and the web)

What is regression?

- Regression is a statistical method used to understand the relationship between variables.
- It helps predict or explain the behavior of a dependent variable based on one or more independent variables.
- For example, in linear regression, the goal is to find the best-fit line that represents the relationship between the variables.

Different types of regression

There are different types of regression, such as:

- **Linear Regression**: Models a straight-line relationship between variables.
- **Logistic Regression**: Used for binary classification problems.
- **Polynomial Regression**: Captures non-linear relationships by fitting a polynomial curve.
- **Multiple Regression**: Involves multiple independent variables to predict a dependent variable.
- Regression is widely used in fields like economics, finance, machine learning, and scientific research to make predictions and analyze trends.

Prediction vs Classification

- Prediction focuses on estimating future or unknown values based on a model. It uses the relationships identified by regression (or other methods) to make forecasts.
- Prediction is similar to classification
  - First, construct a model
  - Second, use model to predict unknown value
    - Major method for prediction is regression
      - Linear and multiple regression
      - Non-linear regression
- Prediction is different from classification
  - Classification refers to predict categorical class label
  - Prediction models continuous-valued functions

PREDICTIVE REGRESSION

- The prediction of continuous values can be modeled by a statistical technique called *regression*.
- The objective of regression analysis is to determine the best model that can relate the output variable to various input variables.
- More formally, regression analysis is the process of determining how a variable Y is related to one or more other variables X1, X2, …, Xn.
- Y is usually called the response output or dependent variable, and X_1 ... X_n are called inputs, regressors, explanatory variables, or independent variables

Source: Data Mining: Concepts, Models, Methods, and Algorithms
by Mehmed Kantardzic

Regression Equation

- The relationship that fits a set of data is characterized by a prediction model called a *regression* equation.
- The most widely used form of the regression model is the general linear model formally written as

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \ldots + \beta_n \cdot X_n$$

- Applying this equation to each of the given samples we obtain a new set of equations

$$y_j = \alpha + \beta_1 \cdot x_{1j} + \beta_2 \cdot x_{2j} + \beta_3 \cdot x_{3j} + \ldots + \beta_n \cdot x_{nj} + \varepsilon_j \quad j = 1, \ldots, m$$

where ϵ_j's are errors of regression for each of m given samples. The linear model is called linear because the expected value of y_j is a linear function: the weighted sum of input values.

Linear regression

- Linear regression with one input variable is the simplest form of regression.
- It models a random variable Y (called a response variable) as a linear function of another random variable X (called a predictor variable).
- Given n samples or data points of the form (x_1, y_1) , (x_2, y_2), …, (x_n, y_n), where $x_i \in X$ and $y_i \in Y$, linear regression can be expressed as

$$Y = \alpha + \beta \cdot X$$

where α and β are regression coefficients.

- With the assumption that the variance of Y is a constant, these coefficients can be solved by the method of least squares; which minimizes the error between the actual data points and the estimated line.
- The residual sum of squares is often called the sum of squares of the errors about the regression line and it is denoted by SSE:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i')^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

where y_i is the real output value given in the data set, and y_i' is a response value obtained from the model

Linear regression

Differentiating SSE with respect to α and β, we have

$$\partial(SEE)/\partial\alpha = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)$$

$$\partial(SEE)/\partial\beta = -2\sum_{i=1}^{n}((y_i - \alpha - \beta x_i)\cdot x_i)$$

Setting the partial derivatives equal to zero (minimization of the total error) and rearranging the terms, we obtain the equations

$$\alpha\sum_{i=1}^{n}x_i + \beta\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_i\ y_i$$

$$n\alpha + \beta\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i$$

which may be solved simultaneously to yield computing formulas for ɑ and β

Linear regression

- Using standard relations for the mean values, regression coefficients for this simple case of optimization are

$$\beta = \left[\sum_{i=1}^{n} (x_i - mean_x) \cdot (y_i - mean_y)\right] / \left[\sum_{i=1}^{n} (x_i - mean_x)^2\right]$$

$$\alpha = mean_y - \beta \cdot mean_x$$

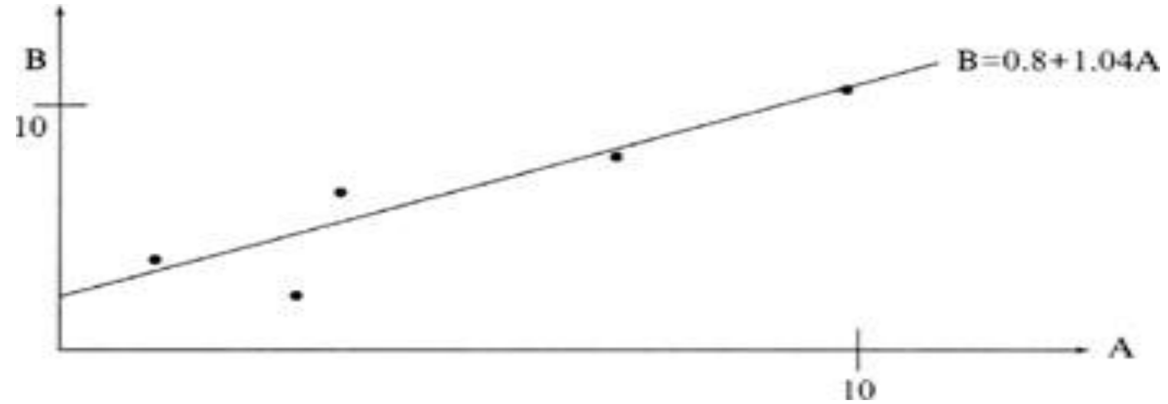
where $mean_x$ and $mean_y$ are the mean values for random variables X and Y given in a training data set.

- It is important to remember that our values of α and β, based on a given data set, are only-estimates of the true parameters for the entire population.

- The equation $y = \alpha + \beta x$ may be used to predict the mean response y_0 for the given input x_0, which is not necessarily from the initial set of samples.

Linear regression

A	B
1	3
8	9
11	11
4	5
3	2



For example, if the sample data set is given in the form of a table, and we are analyzing the linear regression between two variables (predictor variable A and response variable B), then the linear regression can pe expressed as

$$B = \alpha + \beta \cdot A$$

where α and β coefficients can be calculated based on previous formulas (using $mean_A$ = 5, and $mean_B$ = 6), and they have the values

$$\alpha = 0.8$$

$$\beta = 1.04$$

The optimal regression line is

$$B = 0.8 + 1.04 \cdot A$$

Multiple regression

- *Multiple regression* is an extension of linear regression with one response variable, and involves more than one predictor variable.
- The response variable Y is modeled as a linear function of several predictor variables.
- For example, if the predictor attributes are X_1, X_2, and X_3, then the multiple linear regression is expressed as

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$$

where α, β_1 β_2, β_3 are coefficients that are found by using the method of least squares.

Multiple regression

- For a linear regression model with more than two input variables, it is useful to analyze the process of determining β parameters through a matrix calculation:

$$Y = \beta \cdot X$$

where $\beta = \{\beta_0, \beta_1, \ldots, \beta_n\}$, $\beta_0 = \beta$, and X and Y are input and output matrices for a given training data set. The residual sum of the squares of errors SSE will also have the matrix representation

$$SSE = (Y - \beta \cdot X)' \cdot (Y - \beta \cdot X)$$

and after optimization

$$\partial(SSE)/\partial\beta = 0 \;\Rightarrow\; (X' \cdot X)\beta = X' \cdot Y$$

the final β vector satisfies the matrix equation

$$\beta = (X' \cdot X)^{-1}(X' \cdot Y)$$

where β is the vector of estimated coefficients in a linear regression. Matrices X and Y have the same dimensions as the training data set.

Multiple regression **with Nonlinear functions**

•There is a large class of regression problems, initially nonlinear, that can be converted into the form of the general linear model.

For example, a polynomial relationship such as

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1 X_3 + \beta_4 \cdot X_2 X_3$$

can be converted to the linear form by setting new variables $X_4 = X_1 \cdot X_3$ and $X_5 = X_2 \cdot X_3$.

• Also, polynomial regression can be modeled by adding polynomial terms to the basic linear model. For example, a cubic polynomial curve has a form

$$Y = \alpha + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot X^3$$

By applying transformation to the predictor variables ($X_1 = X$, $X_2 = X^2$, and $X_3 = X^3$), it is possible to linearize the model and transform it into a multiple-regression problem, which can be solved by the method of least squares.

Multiple regression

- The major effort, on the part of a user, in applying multiple-regression techniques lies in identifying the *relevant* independent variables from the initial set and in selecting the regression model using only relevant variables. Two general approaches are common for this task:

  1. *Sequential search approach*-which consists primarily of building a regression model with an initial set of variables and then selectively adding or deleting variables until some overall criterion is satisfied or optimized.

  2. *Combinatorial approach*-which is, in essence, a brute-force approach, where the search is performed across all possible combinations of independent variables to determine the best regression model.

- Irrespective of whether the sequential or combinatorial approach is used, the maximum benefit to model building occurs from a proper understanding of the application domain.

Correlation analysis

• Additional postprocessing steps may estimate the quality of the linear-regression model.
• Correlation analysis attempts to measure the strength of a relationship between two variables (in our case this relationship is expressed through the linear regression equation).
• One parameter, which shows this strength of linear association between two variables by means of a single number, is called a *correlation coefficient r*. Its computation requires some intermediate results in a regression analysis.

$$r = \beta \cdot \sqrt{(S_{xx}/S_{yy})} = S_{xy}/\sqrt{(S_{xx} \cdot S_{yy})}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - mean_x)^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - mean_y)^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - mean_x)(y_i - mean_y)$$

LOGISTIC REGRESSION

- Linear regression is used to model continuous-value functions.
- Generalized regression models represent the theoretical foundation on which the linear regression approach can be applied to model categorical response variables..
- A common type of a generalized linear model is *logistic regression*. Logistic regression models the probability of some event occurring as a linear function of a set of predictor variables.

- Rather than predicting the value of the dependent variable, the logistic regression method tries to estimate the probability p that the dependent variable will have a given value.
- We use logistic regression only when the output variable of the model is defined as a binary categorical. On the other hand, there is no special reason why any of the inputs should not also be quantitative; and, therefore, logistic regression supports a more general input data set.

LOGISTIC REGRESSION

- Suppose that output Y has two possible categorical values coded as 0 and 1.
- Based on the available data we can compute the probabilities for both values for the given input sample: $P(y_j = 0) = 1 - p_j$ and $P(y_j = 1) = p_j$.
- The model that we will fit these probabilities is accommodated linear regression:

$$\log(p_j/(1-p_j)) = \alpha + \beta_1 \cdot X_{1j} + \beta_2 \cdot X_{2j} + \beta_3 \cdot X_{3j} + \ldots + \beta_n \cdot X_{nj}$$

- This equation is known as the *linear logistic model*.
- The function $\log(p_j(1 - p_j))$ is often written as logit(p).

LOGISTIC REGRESSION

Suppose that the estimated model, based on a training data set and using the linear regression procedure, is given with a linear equation

$$\text{logit}(p) = 1.5 - 0.6 \cdot x_1 + 0.4 \cdot x_2 - 0.3 \cdot x_3$$

and also suppose that the new sample for classification has input values $\{x_1, x_2, x_3\} = \{1, 0, 1\}$. Using the linear logistic model, it is possible to estimate the probability of the output value 1, (p(Y = 1)) for this sample. First, calculate the corresponding logit(p):

$$\text{logit}(p) = 1.5 - 0.6 \cdot 1 + 0.4 \cdot 0 - 0.3 \cdot 1 = 0.6$$

and then the probability of the output value 1 for the given inputs:

$$\log(p/(1 - p)) = 0.6$$

$$p = e^{-0.6}/(1 + e^{-0.6}) = 0.35$$

Based on the final value for probability p, we may conclude that output value Y = 1 is less probable than the other categorical value Y = O.

LOG-LINEAR MODELS

- Log-linear modeling is a way of analyzing the relationship between categorical (or quantitative) variables.
- The log-linear model approximates discrete, multidimensional probability distributions.
- It is a type of generalized linear model where the output Yi is assumed to have a Poisson distribution, with expected value μ_j. The natural logarithm of μ_j is assumed to be the linear function of inputs

$$\log(\mu_j) = \alpha + \beta_1 \cdot X_{1j} + \beta_2 \cdot X_{2j} + \beta_3 \cdot X_{3j} + \ldots + \beta_n \cdot X_{nj}$$

- Since all the variables of interest are categorical variables, we use a table to represent them, a frequency table that represents the global distribution of data.
- The aim in log-linear modeling is to identify associations between categorical variables.

Regression Summary

- **Regression** is a statistical method used to model the relationship between a **dependent variable** and one or more **independent variables**. It helps us **predict continuous outcomes**.
- **Purpose**: Predict or explain a numerical value.
- **Used in**: Economics, medicine, machine learning, etc.

Regression Summary

Types of Regression:

1. **Linear Regression**
   - Models a straight-line relationship between the variables.
   - Example: Predicting house price based on size.
2. **Multiple Linear Regression**
   - Like linear regression, but with multiple input variables.
3. **Polynomial Regression**
   - Models a non-linear relationship by introducing polynomial terms.
4. **Logistic Regression**
   - Despite the name, it's used for classification (binary outcomes), not regression.
5. **Ridge and Lasso Regression**
   - Regularized versions of linear regression to prevent overfitting.
6. **Non-linear Regression**
   - Models more complex relationships that are not linear or polynomial.
7. **Log-linear Models**
   - **Log-linear models** are a type of **statistical model** used to analyze the **relationships between categorical variables** by modeling the **logarithm of expected cell counts** in a contingency table.