1 Learning Outcomes

- Describe the relationship between models and the phenomena that they represent.
- Describe the relationship between independent and dependent variables in a model.
- Describe the relationship between model input and model output.
- Explain what model parameters are and how they are different than model input.

2 What is a Model?

A model is a mathematical tool used to represent or explain some phenomena, or to predict a certain behavior. Mathematical models are used in many science and engineering disciplines. For example, the model of Gaussian (normal) distribution was first used to analyze the errors of measurement made in astronomical observations: it was observed that these errors were symmetric and that small errors occurred more frequently than large errors. The formula of normal distribution was developed to represent the distribution of these errors. In telecommunication engineering, models are used to represent the channel or medium through which speech signals are transmitted. In this way, an appropriate receiver that extracts the information from the transmitted signal could be designed. More recently, many models that represent the time evolution of COVID-19 infection rates were proposed. The main goal was to forecast the transmission of the disease and to better understand what policies to devise.

Representing an observed phenomenon with a mathematical model requires making some assumptions, this is why models are just approximations of the real phenomenon and are seen as an abstraction or simplification of the real-world. However, they are useful tools that increase our understanding of the real-world:

"*All models are wrong, but some are useful.*" - George Box

One way to mathematically represent a model is through an input-output relationship. In other words, a model can be described as a mathematical formula that takes in an input variable and transforms it into an output variable.

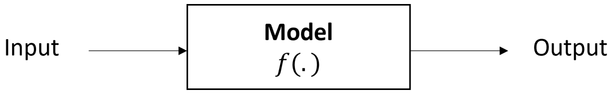

Figure 1: A model can be described with a mathematical function that shown an input-output relationship.

For instance, assume an automotive industry group wanted to identify over- and under-performing car models in terms of sales; it came up with a model that predicts car sales given some car characteristics (Engine size, Vehicle type, Horsepower, Wheelbase, etc). The model represents the relationship between car characteristics and car sales: the input to the model is the car characteristics, the output is the car sales and the model is the mathematical equation that processes the car characteristics and computes its sales.

3 Machine Learning Models

In supervised machine learning, we assume that there is an underlying relationship between a set of features and its corresponding label. Using the training data, we try to approximate this relationship with a machine learning model, which is a mathematical mapping or function that maps a vector of features to its corresponding label. The goal of the training phase is to find this mathematical mapping. The goal of the evaluation phase is to understand how well this mapping approximates the real relationship. The goal of the prediction phase is to use this model to predict the label of future feature vectors.

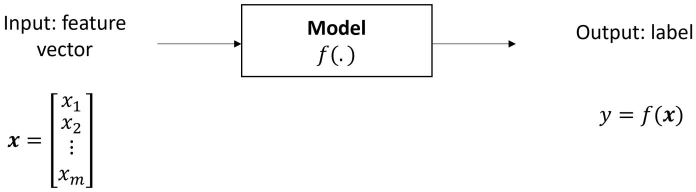
Input: feature vector
Model
$f(.)$
Output: label
$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$
$y = f(\boldsymbol{x})$

Figure 2: A machine learning model is a mathematical function that maps a feature vector to its predicted label.

Parametric vs non-parametric machine learning models

Different machine learning models make different assumptions about the form of the mapping function and how it can be learned. **Parametric** machine learning models assumes a specific functional form or structure for the mathematical mapping. This functional form depends on a finite number of parameters, where parameters can be thought as the building blocks of a parametric model, so that designing a parametric model reduces to findings its parameters.

"*A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model.*
No matter how much data you throw at a parametric model, it won't change its mind about how many parameters it needs." - Artificial Intelligence: A Modern Approach, page 737.

Linear models that are either used for classification or regression are examples of parametric models. In order to compute the label for a given vector of features, linear models computes a weighted sum of the features and then uses this weighted sum in the computation of the predicted label. Training a linear model means using the training dataset to find the appropriate weight for each features; the weights are the parameters of the linear models. Linear regression (regression model) and logistic regression (classification model) are both examples of linear models. We will discuss linear models with more details in the next section.

Non-parametric models are models that do not assume a strict structure for the mathematical mapping. The term "non-parametric" does not mean that these models do not have parameters, but it means that they cannot be described with a finite number of parameters which could be affected by the number of training examples.

"*A nonparametric model is one that cannot be characterized by a bounded set of parameters*"-

Artificial Intelligence: A Modern Approach, page 737.

Examples of non-parametric models include K-nearest neighbors and decision trees, which will be explained later in the course.

4 Linear Models

Here we discuss what we mean by linear models, how to geometrically interpret them and what training linear models mean.

4.1 What are linear models?

Assume $\mathbf{x} \in \mathbb{R}^m$ a vector of m features:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

A linear model consists of computing a weighted sum of the features: it weights each feature x_i with a weight $w_i \in \mathbb{R}$ and then sums the weighted features as in:

$$\sum_{i^1}^{m} w_i x_i + w_0$$

We can collect the weights into a vector: the weight vector $\mathbf{w}$:

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}$$

The weighted sum can then be expressed as the dot product between the weight vector and the feature vector:

$$\sum_{i=1}^{m} w_i x_i + w_0 = \mathbf{x}^T \mathbf{w} + w_0$$

By converting the higher dimensional feature vector into a scalar value, linear models try to compress the information contained in each feature in one scalar:

- With linear regression (regression model), this scalar is directly used as the predicted label or response:

  $$y = \mathbf{x}^T \mathbf{w} + w_0$$

- With linear classification models, a threshold is applied to this scalar value to predict a class. With binary classification, the label does not represent a continuous value, the label represents one of two possibilities i.e., $y \in \{0, 1\}$. A linear classification model applies a threshold to the weighted sum so that if the weighted sum is greater than the threshold, the label is predicted to be one of the two possibilities. Otherwise (the weighted sum is less than the threshold), the label is predicted to be the other possibility. For example, the weights of the linear model can be found so that:

$$y = \begin{cases} 0 & \text{if } \mathbf{x}^T\mathbf{w} + w_0 < 0 \\ 1 & \text{if } \mathbf{x}^T\mathbf{w} + w_0 \geq 0 \end{cases}$$

In classification, we have more than one linear models. Logistic regression, linear discriminant analysis, linear SVM (support-vector machines) are all examples of linear classification models. Each one of this model motivates the problem of finding the weights in a different way.

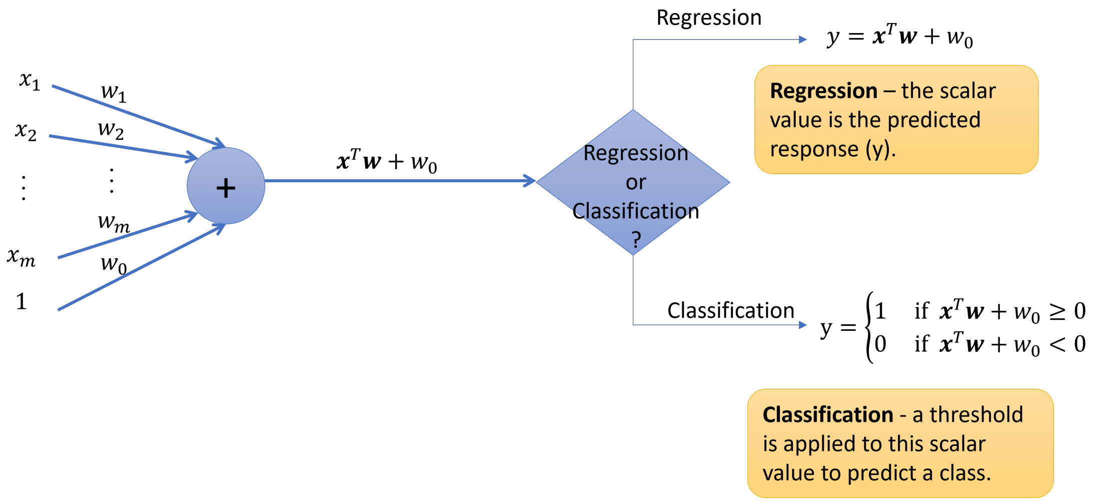
x_1
x_2
$\vdots$
x_m
1
w_1
w_2
$\vdots$
w_m
w_0
+
$\boldsymbol{x}^T\boldsymbol{w} + w_0$
Regression or Classification ?
Regression
$y = \boldsymbol{x}^T\boldsymbol{w} + w_0$
Regression – the scalar value is the predicted response (y).
Classification
$y = \begin{cases} 1 & \text{if } \boldsymbol{x}^T\boldsymbol{w} + w_0 \geq 0 \\ 0 & \text{if } \boldsymbol{x}^T\boldsymbol{w} + w_0 < 0 \end{cases}$
Classification - a threshold is applied to this scalar value to predict a class.

Figure 3: Linear machine learning models.

4.2 Geometric interpretation of linear models

For linear regression, the equation $y = \mathbf{x}^T\mathbf{w} + w_0$ represents a hyperplane's equation in the space with y and the feature vector $\mathbf{x}$ (i.e., if we plot y vs $\mathbf{x}$). For linear classification models, the equation $\mathbf{x}^T\mathbf{w}+w_0 = 0$ represents a hyperplane's equation in the features space. Let's look at some examples for regression and classification.

Regression

Simple Linear Regression - Let's assume we have one feature, then the linear regression model is given by:

$$y = w_1x_1 + w_0$$

This is known as simple linear regression. Given that we're interested in knowing how y is related to x_1, the above equation represents a line equation, and the parameters w_1 and w_0 represent the slope and y-intercept respectively of the line.
As shown in figure 4, a simple linear regression model fits the observed training data with a line. The example shown on the right shows if we can predict car fuel consumption in miles per gallon

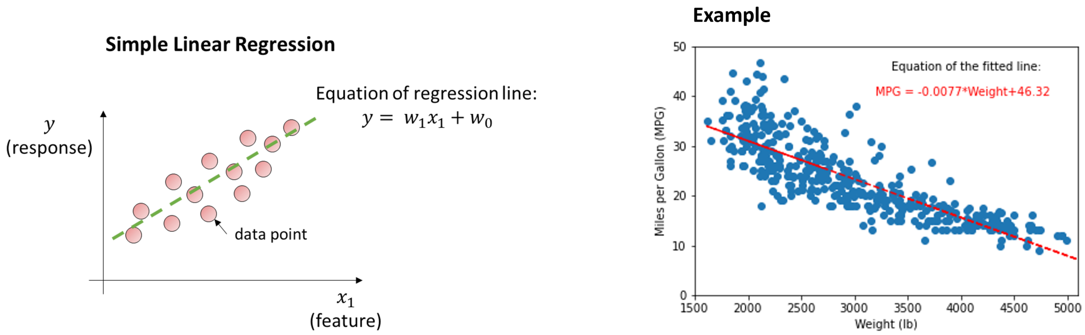

Figure 4: Simple Linear Regression (data source for the example shown on the right)

(MPG) from the car's weight. The line plotted is the line obtained if we train the data with a simple linear regression model. Although the relationship between MPG and weight is not perfectly linear, the linear fit captures the essence of the relationship. As already mentioned, models are always approximations of real-life.

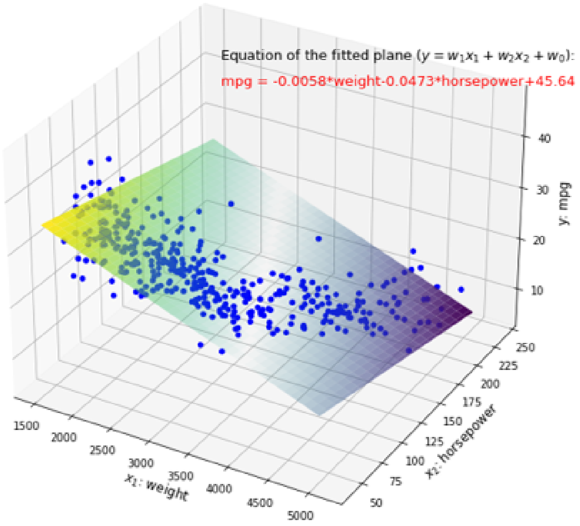

Figure 5: Linear Regression in a three-dimensional setting, with two features and one response. (data source)

Multiple Linear Regression - If we now have two features, then the linear regression model is given by:

$$y = w_1x_1 + w_2x_2 + w_0$$

This equation represents the equation of a plane as shown in figure 5. In this example, in addition to weight, we also consider horsepower as an additional feature that can be useful in predicting MPG. The plane shown is obtained if we train the data with a linear regression model.

More generally, given m features, the linear regression model : $y = \mathbf{x}^T\mathbf{w}+w_0$ represents the equation of the hyperplane in the m+1-dimensional space given by the response y and the m features.

Classification

In linear classification, finding the weights such that

$$y = \begin{cases} 0 & \text{if } \mathbf{x}^T\mathbf{w} + w_0 < 0 \\ 1 & \text{if } \mathbf{x}^T\mathbf{w} + w_0 \geq 0 \end{cases}$$

means finding the hyperplane

$$\mathbf{x}^T\mathbf{w} + w_0 = 0$$

that divides the feature space (space of all possible feature vectors) into two half-spaces so that all vector features that satisfy $\mathbf{x}^T\mathbf{w} + w_0 < 0$ are predicted to belong to class 0, and that all vector features that satisfy $\mathbf{x}^T\mathbf{w} + w_0 \geq 0$ are predicted to belong to class 1. Note that $\mathbf{w}$ represents the normal vector that defines the orientation of the hyperplane.

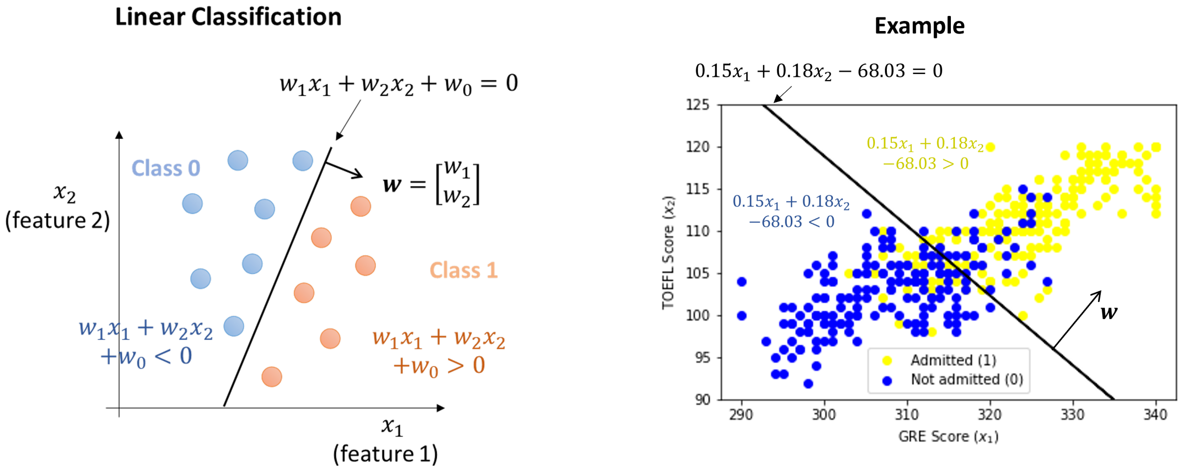

Figure 6: Example of Linear Classification in a 2-dimensional feature space (data source).

Example of two features - assume we have two features, then as seen in figure 6, the line given by the following equation:

$$\mathbf{x}^T\mathbf{w} + w_0 = 0 \text{ or } w_1x_1 + w_2x_2 + w_0 = 0$$

divides the 2-dimensional feature space into two half-planes defined as follows:

- any point that is on the same side of the normal vector $\mathbf{w}$ should satisfy

  $$\mathbf{x}^T\mathbf{w} + w_0 > 0$$

  and is assigned to class 1;

- any point that is on the other side of the normal vector $\mathbf{w}$ should satisfy

  $$\mathbf{x}^T\mathbf{w} + w_0 < 0$$

  and is assigned to class 0.

The right plot of figure 6 shows the GRE scores and TOEFL scores of students who applied to graduated school and whether they were admitted or not. Given this data, we're interested in predicting if a student will be admitted or not based on their GRE and TOEFL scores. The line shown is obtained by training the data with logistic regression (a linear classification model). In figure 8, the colors shown are the actual values of the weighted sum (with no thresholding) in the feature space. We will learn later that with logistic regression, the higher (lower) this value is, the more the model is certain that the feature vector corresponds to class 1 (0). From this example, one can directly see a limitation for linear models: not all data is linearly-separable and finding a more complex model (non-linear or adding more features) might be sometimes more helpful than using a linear model.

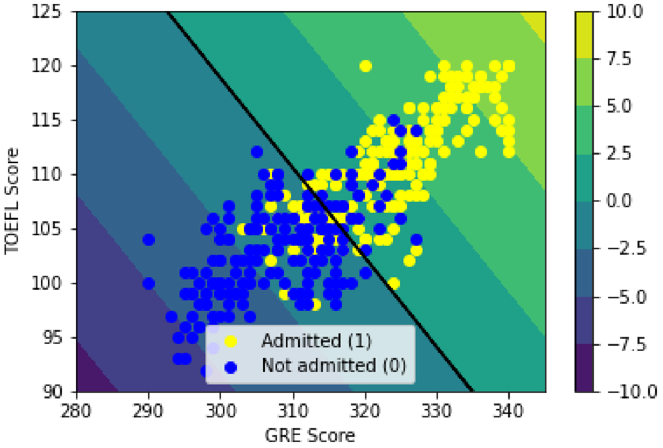

Figure 7: Possible values of the weighted sum in the feature space.

Example of three features - assume we have three features, then as seen in figure 8, the plane given by the following equation:

$$\mathbf{x}^T\mathbf{w} + w_0 = 0 \text{ or } w_1x_1 + w_2x_2 + w_3x_3 + w_0 = 0$$

divides the 3-dimensional features space into two half-spaces. In this figure, three physiological features (depth, intensity, radius) were extracted from breast thermograms to classify normal and abnormal breast thermograms[1].

5 Training Linear Models

As already mentioned, linear models are parametric models that have the feature weights as the model's parameters. These weights are the essential components of linear models, so that knowing them is enough for us to know everything about the linear model. How to choose these weights? This is the goal of the training phase. The weights are learned from the set of training pairs of feature vectors and corresponding labels: $\{\mathbf{x}^{(i)}, y^{(i)}\}$.

[1] Thermography uses a special camera to measure the temperature of the skin on the breast's surface.

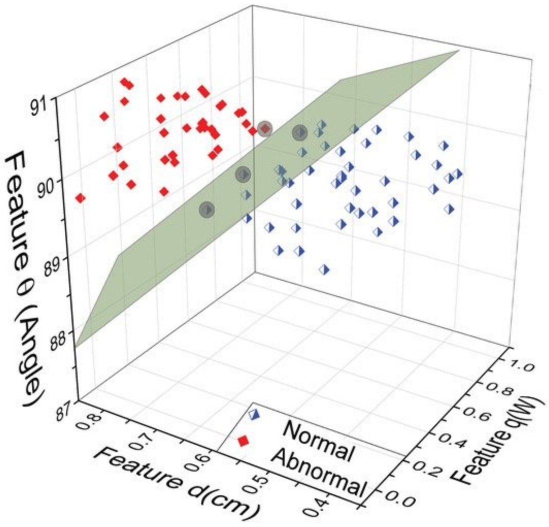

Figure 8: Linear classification of normal and abnormal breast thermograms using three physiological features. (Image source).

In simple linear regression, given some training dataset depicted in figure 9, we see a strong linear relationship that can be modeled using a line. There's an infinite number of possibilities for choosing that line; in other words we have an infinite number of choices for w_1 and w_0: varying w_1 changes the orientation of the line and varying w_0 moves the line vertically. The best w_1 and w_0 are chosen so that the overall error made by the model is minimized. The red vertical lines shown in figure 9 on the right, represent the error between each observed label and its predicted label. The fitted line is chosen to minimize the sum of the squared vertical distances between each observation and the line.

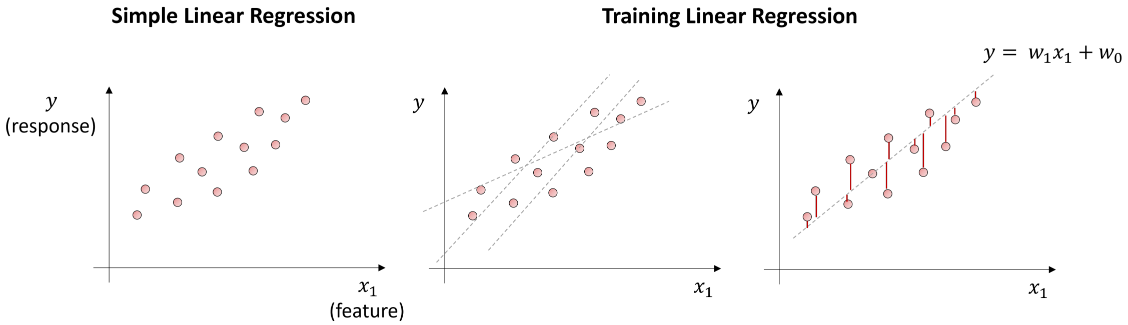

Figure 9: The fitted line is chosen by minimizing the sum of the squared vertical distances between each observation and the line.

In classification, we know that the normal vector is what define the orientation of the separating hyperplane. We can have an infinite number of possibilities for the orientation of the separating hyperplane. We choose the one that best minimize the overall classification error int he training dataset. Each linear classification model (logistic regression, linear SVM, linear discriminant analysis) motivates the problem in a different way but in all cases finding the parameters of a linear model is formulated as an optimization problems that tries to find the best separating hyperplane.

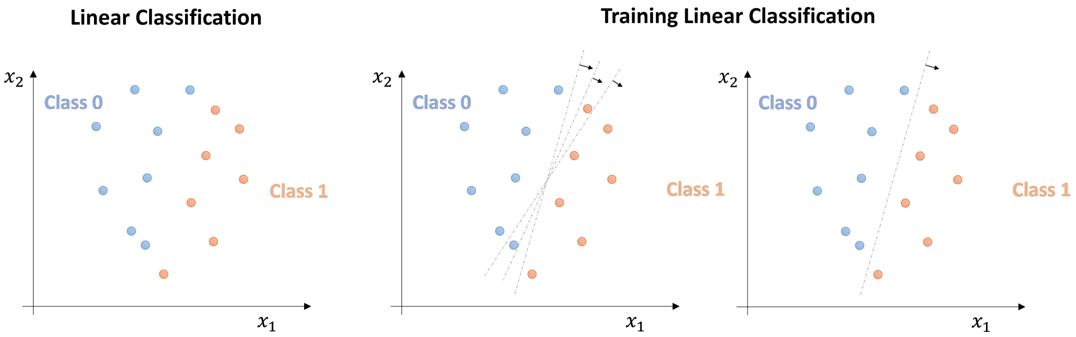

Figure 10: The separating hyperplane is chosen by minimizing the overall miss-classification.

6 Advantages and disadvantages of linear models

Linear models are easy to understand and interpret models. Each feature reflects the contribution of each feature on the overall prediction. Linear models could be extremely useful when data shows strong linear relationship. However, this is not true for all data. In regression, the features might be related to the response in a more complex way so that summarizing it with a hyperplane might not be enough. In classification, not all data is linearly separable. We will learn more about other non-linear models.

7 Resources