Assignment #: 1

Venkata Sai Chelagamsetty

September 14th, 2022

1. (a) Generalization:It refers to how well your model will work with fresh, previously unobserved data that comes from the same distribution as the model's initial data.

   (b) Overfitting: It occurs when a model absorbs the detail and noise in the training data to the point where it has a negative effect on the model's performance on new data. This implies that the model learns concepts from the noise or random oscillations in the training data..

   (c) Underfitting: It describes a model that is unable to generalize to new data or model the training set of data.

   An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

   (d) Regularization: By punishing the coeficients when they attempt to take big values in order to match the outliers, we may prevent the model from overfitting the data.

   (e) No free lunch theorem: It suggests that there isn't just one top optimization algorithm. It also means that there isn't a single ideal machine learning method for predictive modeling issues like classification and regression because of how closely optimization, search, and machine learning are related.

   (f) Occam's razor:It implies that instead of using complex models like ensembles in machine learning, we should use simpler models with fewer coefficients.

   (g) Independent and Identically Distributed Data – A characteristic of a series of random variables where each component is mutually independent and has the same probability distribution as the other values.

   (h) Cross-validation is a method for testing the effectiveness of machine learning models that involves training multiple models on subsets of the input data and then comparing the results. To identify overfitting, or failing to generalize a pattern, use cross-validation.

   (i) Degrees of Freedom: One of the statistic's properties is the ability to vary in as many ways as the number of unrestricted, independent factors that determine its value. This is degree of freedom.

2. Following the frequentist approach, we can determine the probability of getting a head after tossing a coin by simply counting the number of heads and dividing that by the total number of total number of coin tosses.

   (a) In the first cases the number of heads are 12 and number of tails are 5. So the probability of getting a head on the toss is $\frac{12}{17}$ which is greater than 0.5. Therefore, if we were to make guess, we would guess it as a head

   (b) In the second case,the number of heads are 6, whereas the number of tails are 11. Therefore the probability of getting a tail is $\frac{11}{17}$ which is greater than 0.5. Therefore, if we were to make guess, we would guess it as a tail.

3. Consider the following matrices.

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \cdots & x_1^M \\ 1 & x_2 & x_2^2 \cdots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 \cdots & x_N^M \end{bmatrix}, \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix}, \quad \boldsymbol{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$$

So we have $E(w) = \frac{1}{2}\sum_{i=1}^{N} (y(x_n, w) - t_i)^2$, which can be written down in a matrix form as,

$$\begin{aligned} &= \frac{1}{2} \left\| \begin{bmatrix} 1 & x_1 & x_1^2 \cdots & x_1^M \\ 1 & x_2 & x_2^2 \cdots & x_2^M \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 \cdots & x_N^M \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix} - \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} \right\|_2^2 \\ &= \frac{1}{2} \|\boldsymbol{X}w - t\|_2^2 \end{aligned}$$

$$\begin{aligned} \frac{1}{2}\|Xw - t\|_2^2 \\ = \frac{1}{2}(Xw - t)^T (Xw - t) \\ = \frac{1}{2} \left((Xw)^\top - t^\top\right)(Xw - t) \\ = \frac{1}{2} \left(w^\top X^\top - t^\top\right)(Xw - t) \\ = \frac{1}{2} \left(w^\top X^\top (Xw - t) - t^\top (Xw - t)\right) \\ = \frac{1}{2} \left(w^\top X^\top Xw - w^\top X^\top t - t^\top Xw + t^\top t\right) \\ = \frac{1}{2} w^\top X^\top Xw - w^\top X^\top t + \frac{1}{2} t^T t \end{aligned} \tag{1}$$

Now to find the w that minimizes the error, we take the derivative of the above equation with w and equate it to zero. We get

$$\begin{aligned} X^\top Xw - X^\top t + 0 = 0 \\ w_\star = (X^\top X)^{-1} X^\top y \end{aligned} \tag{2}$$