2.6)

Robust regression methods are designed to not be strongly affected by violations in underlying assumptions of the data as well as lower sensitivity to outliers or noise in the dataset, unlike the more conventional least squares methodology which is inefficient and potentially bias in the presence of outliers.

To overcome the shortfalls of the least squares regression, a simple manipulation of the least squares method can be utilized. A common to optimization technique is to minimize the sum of the *absolute* errors. This can be accomplished by summing the absolute values, shown in Eq. 1, or summing the absolute value of the residuals, shown in Eq.2.

$$S = \sum_{i=1}^{n} | y_i | \qquad \text{(Eq. 1)}$$

$$S = \sum_{i=1}^{n} |y_i - f(x)| \qquad \text{(Eq. 2)}$$

Using a least absolute error provides a more robust method to approximate a dataset in that the method is resistant to outliers or noise that might plague the data. Furthermore, since the least absolute error is not squared, each data point is given equal emphasis, compared to conventional least squares introducing more weight to larger residuals (e.g. large variations due to data noise).

2.7)

$$E(w_1, w_0|X) = \frac{1}{N}\sum_{t=1}^{N}[r^t - (w_1 x^t + w_0)]^2$$

Partial derivative of E with respect to w_0:

$$\frac{\partial E}{\partial w_0} = 0 = \sum_{t}^{N}[r^t - (w_1 x^t + w_0)] = \sum_{t}^{N} r^t - w_1 \sum_{t}^{N} x^t - w_0 \sum_{t}^{N} 1$$

$$\sum_{t}^{N} r^t = w_1 \sum_{t}^{N} x^t + (N * w_0) \rightarrow w_0 = \frac{1}{N}\left[\sum_{t}^{N} r^t - w_1 \sum_{t}^{N} x^t\right]$$

Where, $\frac{1}{N}\sum_{t}^{N} r^t$ is the mean of r, or $\bar{r}$, and $\frac{1}{N}\sum_{t}^{N} x^t$ is the mean of x, or $\bar{x}$.

$$\therefore w_0 = \bar{r} - w_1 \bar{x}$$

Partial derivative of E with respect to w_1:

$$\frac{\partial E}{\partial w_1} = 0 = \sum_t^N [r^t - (w_1 x^t + w_0)]\, x^t$$

where, $\frac{\partial}{\partial w_1} w_1 x^t = x^t$, thus the x^t outside the brackets in the previous equation. Now, multiplying x^t through:

$$0 = \sum_t^N [r^t x^t - (w_1 (x^t)^2 + w_0 x^t)] \rightarrow 0 = \sum_t^N r^t x^t - \sum_t^N w_1 (x^t)^2 - w_0 \sum_t^N x^t$$

Substituting w_0 equation previously found into equation:

$$0 = \sum_t^N r^t x^t - \sum_t^N w_1 (x^t)^2 - (\bar{r} - w_1\, \bar{x}) \sum_t^N x^t$$

$$0 = \sum_t^N r^t x^t - \bar{r} \sum_t^N x^t - w_1 \sum_t^N (x^t)^2 + (w_1\, \bar{x}) \sum_t^N x^t$$

Recall that $\sum_t^N x^t = N\bar{x}$ and we can factor w_1 out of the two remaining terms:

$$0 = \sum_t^N r^t x^t - \bar{r} N \bar{x} - w_1 \left(\sum_t^N (x^t)^2 - \bar{x} N \bar{x} \right)$$

$$w_1 \left(\sum_t^N (x^t)^2 - \bar{x} N \bar{x} \right) = \sum_t^N r^t x^t - \bar{r} N \bar{x}$$

$$\therefore w_1 = \frac{(\sum_t^N r^t x^t - \bar{r} N \bar{x})}{(\sum_t^N (x^t)^2 - \bar{x} N \bar{x})}$$

2.9)

If we assume all points of the dataset reside on circle and are equidistant apart from each other, we can show that seven points can be shattered by a triangle for all-possible labeling. To provide context, find the following Figures to illustrate this concept.

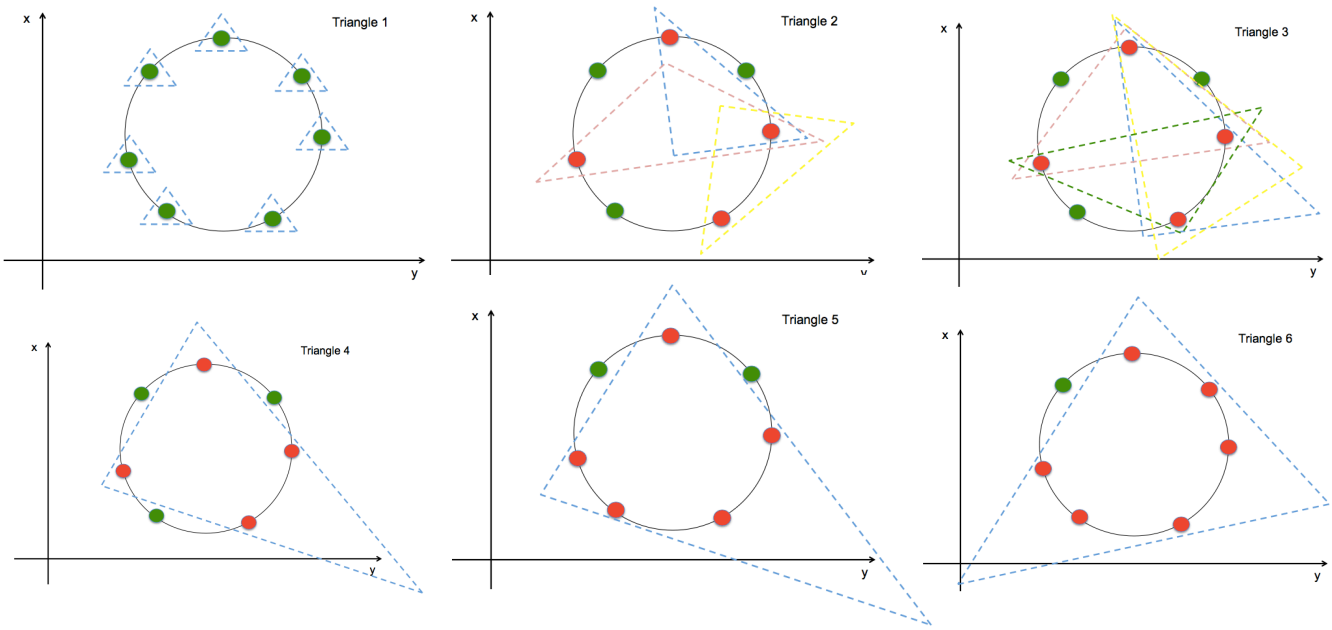

Figure 2.9a: Example illustration of triangle's ability to shatter seven points.

If we wish to increase the data points to eight and maintain the circular distribution of data, we can show that a triangle cannot shatter the points for a difficult alternating labels grouping. That is, for a set of data that alternates the category or labels, such as the list [+,-,+,-,+,-,+,-,+,-], the triangle cannot shatter this case. Thus, the VC dimension is 7 due to the inability to shatter 8 points.

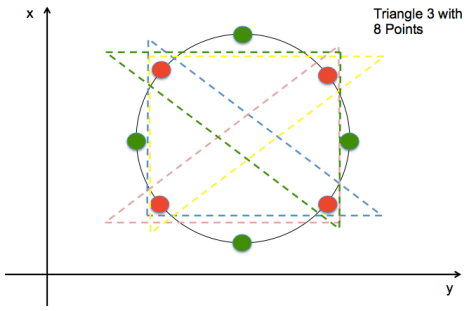

Figure 2.9b: Alternating category of eight points illustrating inability to contain all similar points in one triangle.

2.10)

Say we have a set of points given in the following Figure 2.10a. If we wish to separate the data into two groups using a line, there is a clear area where the line can be placed to accomplish the task. If we wish to find the optimal line that minimize the misclassification of the two groups, we could .

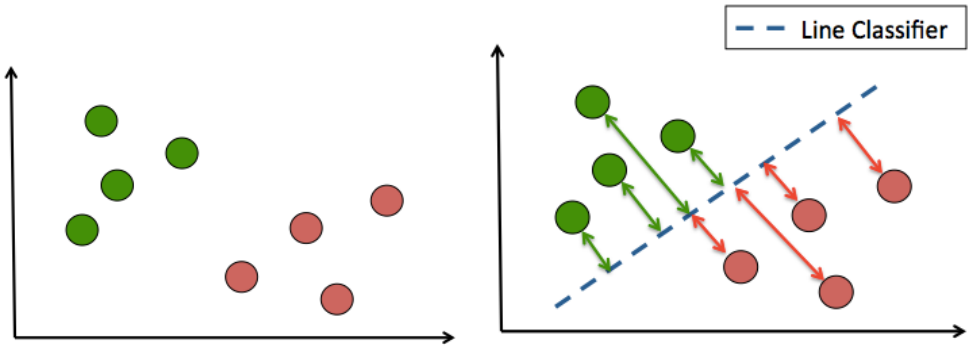

Figure 2.10a (Left): Example data set of two groups, green and red.
Figure 2.10b (Right): Illustration of margin between line classifier and the datum.

We know a linear function can be expressed in three variables (y, x, b) and we know the position of each datum, we can solve for the distance between the line and the datum. Let's say a data point is positioned at (x_0, y_0) and we are using a line expressed as $\mathbf{A}x + \mathbf{B}y + \mathbf{C} = 0$. The distance between the datum and the line, illustrated in Figure 2.10, can be calculated by Eq. 1.

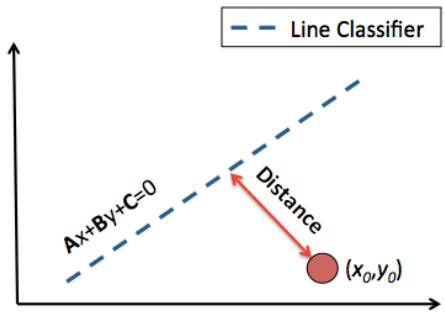

Figure 2.10: Illustration of distance between a datum and line classifier.

$$\text{Distance} = \frac{|\mathrm{A}x_0 + \mathrm{B}y_0 + \mathrm{C}|}{\sqrt{\mathrm{A}^2 + \mathrm{B}^2}} \qquad \textbf{Eq.1}$$

For multiple points, we can maximize the summed distance to find the optimal line that separates the data.

$$\text{argmax}\left(\sum_{(x,y)}\left[\frac{|\mathrm{A}x+\mathrm{B}y_0+\mathrm{C}|}{\sqrt{\mathrm{A}^2+\mathrm{B}^2}}\right]\right) \quad \textbf{Eq.2}$$

Unfortunately, this equation alone does not account for misclassifications, thus we require a parameter to validate the classification. We can use a piecewise function with a binary output (i.e., indicator vector), positive or negative value, to account for the classification status by multiplying said value by the distance.

$$w_i = \begin{cases} 1 & \text{if correct} \\ -1 & \text{if incorrect} \end{cases} \quad \textbf{Eq.3}$$

$$\text{argmax}\left(\sum_{(x,y)}\left[\frac{|\mathrm{A}x+\mathrm{B}y_0+\mathrm{C}|}{\sqrt{\mathrm{A}^2+\mathrm{B}^2}} w_i\right]\right) \quad \textbf{Eq.4}$$

This function will obtain the greatest distances between the line and points and assign a correction value for misclassifications to minimize the false positives or negatives.

2.11)

Labeling error can occur due to subjectivity of the labeling procedure, an error during data entry, or lack of information prior to labeling the training data. Creating an automated method to identify mislabeled data can be difficult, as true outliers can exist in datasets. An easy method would be to apply a Gaussian distribution to a data class and use a large standard deviation threshold to identify high-likelihood mislabeled data points. For example, if a data cluster has an assigned Gaussian distribution, say after a fit function is applied where the center of the data is the mean, the data analyst can apply a large threshold value, say 10 standard deviations, to dismiss any data points 10 standard deviations away from the distribution mean. This method can detect data anomalies, such as mislabeled data points.