DSC 40A Class Notes

Supplement on Spread

So far in this class, we have made predictions based on data. We do this by mathematically representing the loss $L(h, y)$ of a prediction h on a data value y, and then minimizing the average loss $R(h)$ across the data set. We have used calculus, gradient descent, and other methods to minimize the average loss $R(h)$. Our goal with all of these methods was to find the input h that minimizes $R(h)$, and we have found that these inputs, for various different definitions of loss, represent the **center** of the data in some way. We have seen the mean, median, and mode emerge as optimal predictions h^* when we use certain loss functions.

In our effort to minimize $R(h)$, we have focused on the input h^* that minimizes $R(h)$, but we have ignored the minimum value of the output, $R(h^*)$. Next we will consider what the minimum value of the output of $R(h)$ tells us about our data set. We will see that for each loss function, the minumum value of the output of $R(h)$ represents the **spread** of the data in some way. Consider the following loss functions.

1) One loss function we used was the **absolute loss**,

$$L_{\text{abs}}(h, y) = |y - h|.$$

The empirical risk for this loss function is

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|.$$

We found that $R(h)$ was minimized at $h^* = \text{median}(y_1, y_2, \ldots, y_n)$. Therefore the minimum value of $R(h)$ is

$$R(h^*) = R(\text{median}(y_1, y_2, \ldots, y_n)) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \text{median}(y_1, y_2, \ldots, y_n)|.$$

$R(h^*)$ is the average distance of each data point from the median. This is called the **mean absolute deviation from the median**, a term which we can break down as follows. Reading from right to left: for each data point, take the deviation (difference) from the median, put this in absolute value, and then take the mean over all points in the data set. For example, consider the data set $3, 4, 4, 4, 6, 9$. The median is 4 and

the absolute deviation of each data point from the median is $1, 0, 0, 0, 2, 5$. The mean absolute deviation from the median is the average of these numbers, or $\frac{8}{6} = \frac{4}{3}$.

The mean absolute deviation from the median measures how spread out the data is from its center, the median in this case. Higher values of the mean absolute deviation from the median says that the data is further from its median on average, so more spread out. Notice that since the median is the measure of center that corresponds to the loss function L_{abs}, the spread is measured relative to that measure of center.

2) Another loss function we used was the **square loss**,

$$L_{\text{sq}}(h, y) = (y - h)^2.$$

The empirical risk for this loss function is

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2.$$

We found that $R_{\text{sq}}(h)$ was minimized at $h^* = \text{mean}(y_1, y_2, \ldots, y_n)$. Therefore the minimum value of $R_{\text{sq}}(h)$ is

$$R_{\text{sq}}(h^*) = R_{\text{sq}}(\text{mean}(y_1, y_2, \ldots, y_n)) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \text{mean}(y_1, y_2, \ldots, y_n))^2.$$

$R_{\text{sq}}(h^*)$ is the average squared distance of each data point from the mean. You may recognize this measure of spread as the **variance** of the data set. The variance, or its square root, the **standard deviation**, is probably the most ubiquitous way to measure the spread of a data set. Notice that it measures spread about the mean, because $R_{\text{sq}}(h)$ is minimized at the mean.

For example, if the data set is $3, 4, 4, 4, 6, 9$, the mean is 5 and the squared deviation of each data point from the mean is $4, 1, 1, 1, 1, 16$. Therefore the variance is the average of these numbers, $\frac{24}{6} = 4$.

3) Another loss function we saw was the **0-1 loss**,

$$L_{0,1}(h, y) = \begin{cases} 0, & \text{if } h = y \\ 1, & \text{if } h \neq y \end{cases}.$$

This corresponds to the empirical risk

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^{n} L_{0,1}(h, y_i),$$

which is just a count of the number of data points y_i not equal to h. We found that $R_{0,1}(h)$ was minimized at $h^* = \text{mode}(y_1, y_2, \ldots, y_n)$. Therefore the minimum value of $R_{0,\,1}(h)$ is

$$R_{0,\,1}(h^*) = R_{\text{sq}}(\text{mode}(y_1, y_2, \ldots, y_n)),$$

which is just a count of the number of data points y_i not equal to mode$(y_1, y_2, \ldots, y_n)$. For example, on the data set $3, 4, 4, 4, 6, 9$, there are 3 data points not equal to the mode. This value, 3, says something about the spread of the data relative to the mode. For example, if fewer of the data points had been equal to the mode, then this value would have gone up, indicating that the data is less clustered at one particular value, or more spread out. Just as the mode was a crude way of measuring the center of a data set, this measurement of counting the number of data points that are different from the mode is a very basic way of measuring spread. It is not especially useful and it doesn't even have a name. Notice that the more sophisticated our loss function, the more useful the resulting measures of spread and center.

We have now seen several different measures of center and of spread and how they correspond to different loss functions. These measures of center and spread are often called **descriptive statistics** because they summarize characteristics of a data set at a high level. When learning from a data set, we are primarily concerned with the value h^* that minimizes the empirical risk function $R(h)$, but we have also seen that the value $R(h^*)$ gives meaningful information about the variation in the data set.