Evaluation Metrics

Picking the Right Metric

In building machine learning models, we need to first need to pick a metric for performance and test how well our model is performing. There are several metrics we will cover that depend on the problem we are trying to solve.

Before we can pick a performance metric, first it is important to recognize is that machine learning is about learning from data to make predictions. Here we focus on Supervised Machine Learning where we have labelled data and our model makes classification or regressive types of prediction.

In addition, it is also important when testing your model to partition your dataset into training and testing data. If the training and test datasets are not partitioned, we run into issues evaluating the model because it has already seen all the data. We need an independent set of data to verify that the model can generalize well rather than just to the training examples. We will go over common sources of model errors in this lesson and cover how to properly split your datasets in the Data Modeling & Validation section.

Classification and Regression

Classification is about making prediction on unseen examples and deciding which category new instants belongs. For example, we can organize objects based on the color blue or red, or whether they are square or round so when we see new objects we can organize them by their features.

In regression, we want to make a prediction on continuous data. For example, we might have a list of different people's height, age, and gender and wish to predict their weight. Or perhaps, like in the final project of this course we have some housing data and wish to make a prediction about the value of a single home.

Depending on the problem at hand will largely determine how we choose to evaluate a model.

Classification vs Regression Metrics

In classification we want to see how often a model correctly or incorrectly identifies a new example, whereas in regression we might be more interested to see how far off the model's prediction is from the real true value.

For the rest of this lesson we will go over several performance metrics. For classification we will go over accuracy, precision, recall and F-score. For regression we will go over mean absolute error and mean squared error.

Classification Metrics

For classification we are dealing with models that make prediction on discrete data. That is to say these models decide if a new instance belongs or does not belong in a given set of categories. In this

case we are measuring if the prediction did or did not get accurately classified the instance in question.

Accuracy

The most basic and common classification metric is accuracy. Accuracy here is described as the number of items classified or labeled correctly over all items in that class.

For instance if a classroom has 15 boys and 16 girls, can a facial recognition software correctly identify all boys and all girls? If the software can identify 10 boys and 8 girls than the software is 66% and 50% accurate respectively for boys and girls:

$$\text{accuracy} = \frac{\text{number of correctly identified instances}}{\text{all instances}}$$

For more information about accuracy and how to use it in sklearn, check out this link here.

Shortcomings of Accuracy

- Not ideal for skewed classes
- Not suitable where one might want to be conservative or aggressive in prediction

F1 Score

Now that we covered precision and recall, one extra metric you might want to consider is using the F1 score. F1 score considers both the precision and recall in order to compute a new score.

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0:

$$F1 = 2\frac{precision * recall}{precision + recall}$$

For more information about F1 score how to use it in sklearn, check out this link here.

Regression Metrics

As mentioned earlier for regression type of problems we are dealing with model that make predictions on continuous data. In this case we care more about how close the prediction is. For example with height & weight predictions we do not care as much if the model can 100% accurately predict someone's weight down to a less than a fraction of the pound, but perhaps how consistently the model can make a close prediction (perhaps with 3-4 pounds of the individual).

Mean Absolute Error

As you may have recalled in statistics, we can measure error using absolute error to find the predicted distance from the true value. The mean absolute error takes the total absolute error of each example and averages the error based on the number of data points. By adding up all the

absolute values of a model we can avoid canceling out errors from being too high or below the true values and get an overall error metric to evaluate the model on.

For more information about mean absolute error and how to use it in sklearn, check out this link here.

Mean Squared Error

Mean squared is another very common metric to measure model performance. In contrast with absolute error, the residual error (the difference between predicted and the true value) is squared.

Some benefits of squaring the residual error is that it automatically converts all the errors as positives, emphasizes larger errors rather than smaller errors, and from calculus is differentiable which allows us to find the minimum or maximum values.

For more information about mean squared error and how to use it in sklearn, check out this link here.