Explaining what learned models predict: In which cases can we trust machine learning models and when is caution required?

Prabhupad pradhan
prabhupad26@protonmail.com

October 19 2021

1. Introduction

Over the past two decades there has been a significant amount of progress in the world of artificial intelligence and machine learning, by making the use of available data the machine learning models have been able to achieve near human level results. This has led to its widespread adoption in solving many complex problems using available data. However these machine learning models mostly remain black boxes, so in order to establish trust on the predictions certain performance metrics are defined which highlights the robustness and generalization ability of a model. These performance metrics are then used to decide whether to rely on these models. There are several factors which influence these metrics, the majority of which is the quality and quantity of the data used for training the model. In this essay we will discuss the performance metrics which helps us decide when to trust the model and when to be cautious about the predictions.

2. Metrics to evaluate a machine learning model

Evaluating a machine learning model is an essential part of a machine learning pipeline, it helps us decide whether to deploy the model or optimize it further to make it better.
A machine learning task can be divided into two categories : Regression or Classifier, there are many metrics for both the categories.

2.1 Regression metrics

The regression models have a continuous output, so usually the distance between the predicted truth and the ground truth is used as the metric to determine the model performance. There are various techniques to calculate the distance, some of them are : Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error(RMSE), R^2 (R-Squared).
Depending upon the available data the selection of these metrics are done to ensure the model is predicting as per the expectation, for example when there are outliers in the data MSE penalizes small errors by squaring them which leads to an overestimation of how bad the model is, on the other hand MAE is more robust towards outliers than MSE since it doesn’t exaggerate errors [1]. But if the requirement is to emphasize the outliers, MSE should be used. Thus based on problem statements and available data performance metrics are defined.

We performed a small exercise to evaluate a model with RMSE and MAE metrics to see which metric helps us understand the reliability of the model. Figure below shows :

- **A.** Scatter plot showing two variables with a good correlation.
- **B.** Histogram showing the error between Y(observed) and Y(predicted) using normalized RMSE
- **C.** Histogram showing the error between Y(observed) and Y(predicted) using normalized MAE

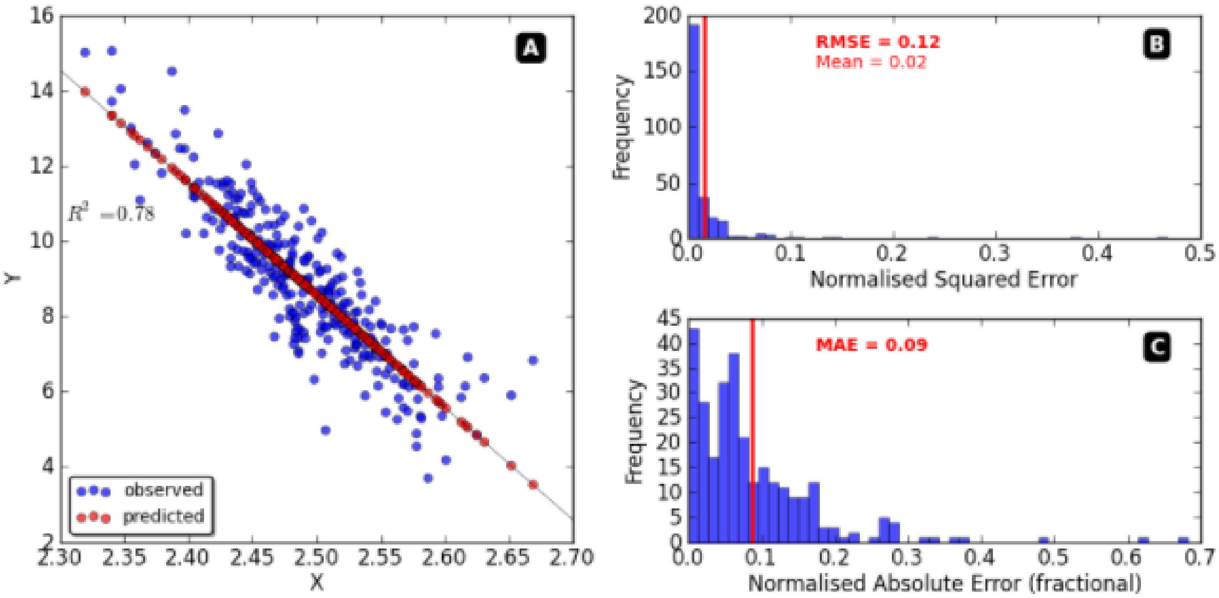


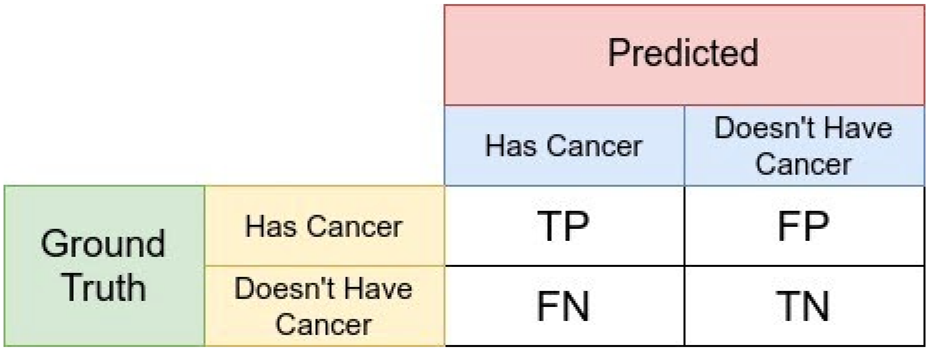
Figure shows a comparison between root mean square error and mean absolute error for a regression problem

In scenarios when it makes more sense to give more weight to points further away from the mean i.e., being off by 10 is more than twice as bad as being off by 5 in such cases RMSE is a more appropriate measure of error. If being off by ten is just twice as bad as being off by 5 then MAE is more appropriate [7]. This is just one of the instances where a model qualifies with these metrics and ensures that it can perform a regression task as per our expectation, several researches have been conducted to explore various techniques in determining performance of a model such as using a combination of different metrics [2].

2.2 Classification metrics

The classification models have a discrete output, so usually the metrics are chosen to compare these discrete classes to evaluate the model's performance and tell you how good or bad the classification is. There are various evaluation metrics which are used to determine the model performance such as Accuracy, Precision and Recall, F1-score, AU-ROC, there is a visualization technique called Confusion matrix which is used to visualize ground truth versus model predictions, based on this matrix other metrics such as

accuracy, precision and recall evaluate the model performance. Problems like binary classification, multiclass classification
Depending upon the requirement and the available data these metrics are used for model evaluation, for instance if we are building a cancer prediction system which takes in some information about the patient and predicts if the patient has cancer or not, the below figure shows the confusion matrix for this problem.



		Predicted	
		Has Cancer	Doesn't Have Cancer
Ground Truth	Has Cancer	TP	FP
	Doesn't Have Cancer	FN	TN

Figure shows the layout for confusion matrix for a cancer prediction system

Every cell in this matrix represents an evaluation factor, here the positive sample means the patient has cancer and negative sample means doesn't have cancer (so the null hypothesis here is assumes as "*The patient has cancer*"):

- True Positive(**TP**) signifies positive samples correctly predicted by the model
- True Negative(**TN**) signifies negative samples correctly predicted by the model
- False Positive(**FP**) signifies positive samples incorrectly predicted by the model. This factor is also referred to as Type-I error in statistical nomenclature. This error positioning in the confusion matrix depends on the choice of the null hypothesis.
- False Negative(**FN**) signifies negative samples incorrectly predicted by the model. This factor is also referred to as Type-II error in statistical nomenclature. This error positioning in the confusion matrix also depends on the choice of the null hypothesis.

Now using these factors from the confusion matrix other metrics such as precision, recall etc. are derived.

Precision metric represents the Type-I errors(**FP**) which occur when we reject a true null Hypothesis, so in this case the model labels a cancerous patient to be non-cancerous. The value of precision ranges from 0 to 1, a value towards 1 represents that the model didn't miss any true positives and is able to correctly identify cancerous patients which is exactly how we would want this model to behave. On the other hand if the value is near 0 then this model should not be used for prediction and should be fine tuned to bring its precision near

1.

Recall/Sensitivity metric represents the Type-II errors(**FN**) which occur when we accept the false null Hypothesis, so in this case the model labels a non-cancerous patient to be cancerous. The value of recall ranges from 0 to 1, a value towards 1 represents that the model didn't miss any true positives and is able to correctly identify cancerous patients which is exactly how we would want this model to behave. On the other hand if the value is near 0 then this model should not be used for prediction and should be fine tuned to bring its precision near 1.

These metrics play a vital role in accepting or rejecting a machine learning model for building such disease prediction systems. However in order to improve this model one can improve either precision or recall but not both of them. For example in our example above if we try to reduce the cases of non-cancerous patients being labeled as cancerous it will have no effect on cancerous patients being labeled as non-cancerous. So in this case a combination of both the metrics are used to calculate the F1-score (which is the harmonic mean of precision and recall).

So these are some of the metrics that are used for deciding whether to accept or reject a model for classification problems with just two classes. For classification problems involving more than just two classes there are various metrics along with the metrics we just discussed that are used to determine the performance of a model [3]. In the next section we will discuss the various factors which cause the poor performance of learning models and how we can diagnose them.

3. Factors affecting the model performance

There are several factors which lead to poor performance, in this section we will be discussing some of those factors and try to propose possible solutions to diagnose them.

3.1 Data unavailability

In certain cases the data that we are feeding to the learning model doesn't represent the entire targeted population and due to this insufficiency incorrect predictions may happen. So there should be enough examples present in the training set otherwise the generalization error may increase, virtual creation of data has been proved useful in handling this kind of issues [4].

3.2 Imbalance data

This kind of problem occurs when the examples of a few of the classes in the dataset are present in huge amounts, outnumbering the other classes in the dataset. This issue could lead to models performing good results on the dominant classes which could result in misleading the evaluation result as if the model accuracy was high. In order to overcome this problem several performance measures are analyzed to discover the imbalance ratio in the dataset [5], another way to handle this issue is to over-sample or under-sample the data.

3.3 Missing value

We often come across data which has missing data in them, although there are some learning algorithms which are capable of handling such scenarios. There are some scenarios where the data is missing partially or completely, so based on the way these values are missing there are different techniques to handle the missing data, for example if the data is partially missing then that data point can be completely deleted or replaced with some average value [6].

4. Conclusions

Trusting a machine learning model to perform a task is completely dependent on understanding the reason behind those predictions and defining a metric with that understanding. Those metrics are then used as a prerequisite for that learning model which decides whether to accept it or reject it. This essay has discussed various performance metrics in a technical context and has explained the scenario in which caution should be taken, with this information one can build a high performing machine learning model which can be trusted for its generalization and accuracy.

References

1. Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology , Available at https://arxiv.org/ftp/arxiv/papers/1809/1809.03006.pdf
2. Root mean square error (RMSE) or mean absolute error (MAE). Available at: https://www.researchgate.net/publication/262980567_Root_mean_square_error_RMSE_or_mean_absolute_error_MAE
3. Metrics for Multi-Class Classification : An overview , available at : https://arxiv.org/pdf/2008.05756.pdf
4. Incorporating prior information in machine learning by creating virtual examples available at : https://ieeexplore.ieee.org/document/726787
5. Classification performance metric for imbalance data based on Recall and Selectivity normalized in class labels, Available at : https://arxiv.org/pdf/2006.13319.pdf
6. Imputation of missing data using machine learning https://www.aaai.org/Papers/KDD/1996/KDD96-023.pdf
7. https://stats.stackexchange.com/questions/48267/mean-absolute-error-or-root-mean-squared-error