wipro holmes

wipro

The Question of Accuracy

When it pertains to classification problems in Machine Learning (ML), the definition of accuracy is straightforward, it is “the ratio of the number of predictions that the model got correct to the total number of predictions made”. Technically, accuracy is defined as above only in classification problems in ML. Though one may argue that that there needs to be an equivalent measure of “accuracy” of a regression model, the term and quantification of the prediction error in a regression model is different. We use concepts like mean squared error, root mean squared error, correlation coefficient, absolute error and mean absolute error among others, in the case of regression to get a sense of the accuracy of prediction.

For unsupervised algorithms like clustering, the measure of model performance can be done using modularity[i]. This is a measure of how closely the points in a cluster are clumped together and how apart clusters are from each other.

We also have the concepts of train-accuracy and test-accuracy. Once a model has been built using training data, we can measure the prediction accuracy of the model on a subset of the training data itself - this is the train-accuracy. The measurement of accuracy on data that has not been seen during training is the test-accuracy. Ideally, one would like to see that the test accuracy is as high as possible for a given model once the training is done and it is deployed.

With the exception of reinforcement learning, one should keep in mind that the accuracy of an ML model will not improve over time as it “sees” more data. The only way one could improve the performance of a model in production would be to retrain it on more data or more relevant data and redeploy the new model if it has better performance. This is done using MLOps - Machine Learning Operations. There is a tendency to call this process continuous learning, but, in reality, the current model is merely replaced with a better one. The model itself does not change for the better.

Test accuracy is what matters in production, using the right dataset for determining test accuracy is the first step in determining its value.

What affects the test accuracy?

ML model choice

Given a classification problem, clearly the model architecture itself is the primary contributor to the accuracy of predictions since, for example, some neural net architectures are better suited for certain problems than others. In a deep neural network, the model architecture that correctly captures the relevant features and transforms for the problem at hand, is of paramount importance for the model to have good generalizing capacity and hence good test-accuracy. Convolutional Neural Networks[ii] are best suited for extraction of visual features that lead to the right "internal" features and maps in the neural network, leading to effective classification of images. Recurrent neural networks[iii] have the ability to remember, process and capture the essence of sequence information, which is most relevant for tasks like language translation.

Training data volume and diversity

The quantity and diversity of the training data has a profound impact on the training of the model as one might expect. Too little data and the model will not be able to converge, leading to large error and poor accuracy. If the quantity is large enough but the diversity is low meaning high-class imbalance, then the model is skewed towards the most populous class.

Training data assumptions

The assumptions made on the training data is equally important, for example, a classifier that can distinguish between the images of dogs and cats, if it has not been also been trained with the images of neither dogs or cats as a third "none of the above" category, then the trained model is making an assumption that only cat and dog images need to be provided for inference. If we provide a third non-cat/dog image for prediction, then the model will be unable to predict "none of the above". Other assumptions may be much more subtle, for example, a model for the detection of faces in train image may only have faces in the image that are at-least 50x50 pixels in size. Similarly, a sentiment analysis model built using plain text without emoticons may be less effective when used on social media posts, where emoticons are plentiful.

Synthetic training data

Many ML practitioners are using synthetically generated training data due to the lack of or sparse availability of real world training data. It must be taken with caution that the synthetic data should be as good as the real, in terms of all possible parameters, like for signature detection, we could paste cropped signature images on images of documents and generate synthetic data. This is a good use of the concept, but in most other cases, it is hard to mimic the real world data, in which case, the trained model will have very low-test accuracy.

Data augmentation

In most vision based ML problems, data augmentation is done regularly to increase the train data volume and the diversity of the data. This has an effect on the final accuracy of the trained model. Care has to be taken that the augmentation technique does not introduce data that violates the train data assumptions.

Hyper-parameters

The hyper-parameters used during training can also affect accuracy. These parameters are not learnt during training but either set manually or searched from an exhaustive list. Some examples of hyper-parameters in deep neural nets include, learning rate, mini batch size, number of epochs, momentum, dropout values and network initialization schemes among others. These chosen hyper-parameters have an effect on the weights of the ML model results that impact the test accuracy. In shallow ML, the degrees of freedom allowed in the ML model, for instance in non-linear regression models, are also an example of a hyper-parameter. The incorrect value of this parameter may result in overfitting or under-fitting.

Test dataset quantity and diversity

Clearly, since the test accuracy is derived from the test data, the quantity and the diversity of this data has an impact on the assessment of the accuracy of the model. If we choose too little data, then we may get a skewed picture of the accuracy. Similarly, if we choose data which has low diversity meaning which has high-class imbalance, then, the estimation of model accuracy may be incorrect. Techniques like k-fold cross validation are very good at estimating the model performance when the quantity of data is not too high, but there is enough representation for each of the classes.

Test data assumptions

The exact assumptions made by the training data are to be followed by the test data. If this is not the case, then, the measurement of the model performance based on accuracy has no meaning.

Erroneous labelling of train or test data

Usually when manual labelling is involved, in supervised ML, there is a good chance that errors are introduced in the labelling of samples. This has a detrimental effect on the model performance if there are errors in the train data labelling, or wrong conclusions on the performance if there are errors in the test data labelling.

Accuracy from different points of view

Even though the model performance is fixed after training, the accuracy measured can be viewed from different viewpoints to get a holistic sense of the performance of the ML model, for example, in the classic case of extraction of text from document image using OCR, we can define at least two accuracy measures - the first being character level accuracy, which is the ratio of the total number of characters correctly recognised to the total number of characters in the test dataset. Then comes word level accuracy where we want the ratio of the total number of words correctly recognised to the total number of words in the test dataset. Similarly, we can define these measures at different levels like paragraph, page and document.

Similarly, in the case of extraction of key value pairs from invoices using an ML model, we can define the accuracy at a field level where we are talking about the ratio of the total number of fields (key value pairs) correctly extracted to the total number of fields in the dataset. The accuracy at the invoice level would be the ratio of the number of invoices which have all fields correctly extracted to total number of invoices in the test dataset.

It is very important that we are specific on which definition of accuracy we are referring to when qualifying the accuracy of an ML model.

ML model accuracy decay

Post the training, once the model is in production, the measured accuracy may decrease over a period of time. This is called model decay. There are two major reasons for model decay – the first being data drift, which is the introduction of new varieties of test data which the model was never trained on. Similarly, concept drift that happens over a period of time is the change in interpretation of the data as having a predicted value that is different from what was used earlier for training. In both the cases mentioned, we will need to retrain the model with newly labelled data so that the decay is addressed. The astute reader will immediately recognize that the decay is due the violation of the assumptions made on the initial train data.

Committing on accuracy

Once the accuracy of the trained ML model has been accurately measured using a validation set approach of k-fold cross validation, we should diligently document the assumptions on which the test data was accumulated and labelled. This set of assumptions along with the quantity and diversity should be well documented. This documentation should find its way to any agreement on model performance with the client using the ML model.

What does 80% accuracy mean?

In traditional software powered by algorithms, it is a tendency to prove the efficacy of a software module, with a handful of “extreme cases”. This approach will not work for ML models. Let us take the example of Optical Character Recognition - OCR, if we were trying to measure the accuracy of the model, then a handful of poor document image samples to judge the performance would be an incorrect approach. The right approach would be to assemble and label several hundred records, then predict on these new records with the ML model, arriving at a good estimate for the accuracy. So, if a product claims 80% accuracy, then it is based on the data assumptions and the volume and diversity of the test data. When we say test data volume, the idea would be that the accuracy converges to a fixed value once there is sufficient test data volume. Too little is not good enough and too much may be unnecessary.

Due diligence during solution design

When designing an end-to-end solution for a customer, we take special care in assessing the data situation at the customer so that we can theoretically assess the performance of the ML models with this data. One such assessment table for a sample problem of entity extraction from invoices is detailed below:

Assessment Table

Invoice Language	Document Types	Min DPI: Max DPI	Latin Script	Embedding Function	Approximate Number of Vendors	Handwritten Text Present?	Multiple Languages in Document?
English	Images, Text PDFs, Scanned PDFs, Excel files, Email body	96:290	Y	English, case sensitive Text Embedding Model	40,000	Yes	No
German		200:300	Y	Multilingual, case insensitive Text Embedding Model	20,000	Yes	Yes
Spanish		150:290	Y		5,000	Yes	No
Dutch		200:300	Y		5,000	Yes	Yes
Arabic		200:290	N		2,000	No	Yes

A novel approach to handling data drift – The data assumptions filter

A method which can be used to prevent data drift, discussed previously, is as follows. The central idea would be to have an assumptions pre-condition module, which may be rule-based ML based, for example, assuming we process only A4 page scans in an ML model, we can state that the minimum required DPI is 300. This would be a business rule which checks that the image size is at-least 2480 pixels x 3508 pixels. An example of an ML-based assumptions filter could be a binary classifier which checks for artefacts like stamps and seals among others, for instance, if we are processing English invoice images, then once we know the document is an invoice, an ML-based binary classifier would check for artefacts in the image like stamps, seals or handwritten signatures which affect the invoice's OCR quality. The classifier would pronounce a fit or unfit vote on each image. If an incoming image fails to pass through the filter, then we log the data record and flag it as "not fit" for the model. This helps in auditing the performance of the model over a period of time and taking actions that improve overall system performance.

Conclusion

We have looked at what is accuracy in the context of ML models and what are the ingredients that affect its test accuracy. Any ML model makes assumptions on the data that was used for its training. These assumptions would have to be mandatorily adhered to while preparing the test data which estimates the accuracy of the model. While in production, we recommend the usage of a "data assumptions filter", which can act as a doorkeeper which allows only valid data to the prediction engine.

References

[1]https://bit.ly/2OUsomS

[2]https://bit.ly/2sjHH0x

[3]https://bit.ly/2rrf6pQ

About the author

Prithvi S Javgal
ML Architect, HOLMES,
Wipro limited.

Prithvi has over 18 years of software design and development experience in various domains including healthcare, embedded, security, SaaS, robotic automation and other domains of software engineering. He also has deep expertise in ML applied to document understanding, insurance claims fraud, computer vision problems in the retail and deep learning powered video surveillance.

As a member of Wipro HOLMES, he is engaged in developing ML-powered automation solutions for insurance claims processing, including adjudication, auditing, severity prediction and fraud detection.

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035,
India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 175,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**

IND/B&T/DEC2019-MAY2020