# Things to look into

Accuracy = F1 score when classes are balanced?

R2 score for regression

# Classification

## Confusion matrix

<https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5>

A lot of the information is from ChatGPT.

A confusion matrix is an matrix used for evaluating the performance of a classification model, where is the number of target classes.

For example, for or binary classification, we have

A chart with different colored squares

Description automatically generated

A good model has high TP and TN and low FP and FN.

## Accuracy

Accuracy measures the rate of correct predictions. It’s a valid metric when your data has balanced classes but not when your data is imbalanced.

For binary classification,

In general, accuracy is the sum of the diagonal elements in the confusion matrix divided by the sum of all elements.

The accuracy of this confusion matrix is

A graph on a black background

Description automatically generated

## Precision

Precision measures the accuracy of predictions for each class. That is, given class ,

Where is the number of samples correctly predicted to be class , and is the number of samples incorrectly predicted to be class .

Precision measures the accuracy of the confusion matrix columns and is an appropriate metric when it’s important to identify false positives – for example, in spam detection, where we want to reduce the number of false positives.

In our example,

A screenshot of a black screen

Description automatically generated

## Recall, aka sensitivity or true positive rate (TPR)

Recall measures the ability of the classifier to find all the samples of a class. That is, for a class ,

Where is the number of class samples correctly predicted to be class , and is the number of class samples that are misclassified.

Recall looks at the rows of the confusion matrix and is an appropriate metric when it’s important to identify false negatives – for example, when identifying serious illnesses in healthcare, it’s important to reduce the number of false negatives (patient has the illness but the test results say they’re fine).

In our example,

A screen shot of a graph

Description automatically generated

## F1 score

The F1 score is the harmonic mean of precision and recall, thus capturing both metrics in one and providing a measure of both false positives and false negatives. A high F1 score means the classifier has both high precision and high recall, that is, the model makes correct positive predictions and also correctly identifies most of the actual positive samples.

Some notes –

1. Ideally, precision and recall are both 1, and F1 score is 1
2. When precision and recall are the same, F1 score is equal to precision/recall
3. Arithmetic mean weights the values equally, but harmonic mean penalizes low values. This means that F1 score will be high only if both precision and recall are high, penalizing classifiers that perform well on one but poorly on the other.

Precision and recall generally trade off. For example, a classifier may achieve high precision by being very conservative in its positive predictions but at the cost of low recall (missing many actual positives). Conversely, a classifier might achieve high recall by predicting many positives but at the cost of low precision (many false positives).

## Specificity, aka true negative rate (TNR)

Specificity measures a model’s ability to correctly identify negative samples (in multi-class, this is all samples that are not in the class of interest). That is, for a class ,

Where are samples, not in class , that are correctly predicted not to be in class , and are samples, not in class , that are incorrectly predicted to be in class .

Like precision, specificity is concerned with identifying false positives.

For example,

A black background with white text and red line

Description automatically generated

A screenshot of a test

Description automatically generated

A green rectangle with white text

Description automatically generated

### False positive rate (FPR)

## ROC (receiver operating characteristic) curve and AUC

Like the F1 score, the ROC curve is another metric that measures both false positives and false negatives. While F1 score combines precision and recall, the ROC curve combines recall and false positive rate.

The output of your binary classification model is a probability of 1. The prediction of the model depends on the threshold we choose, which can be anywhere between 0 and 1.

Decreasing the threshold means we’re more likely to predict 1, which raises our sensitivity. However, this also raises our false positive rate.

Increasing the threshold means we’ll have fewer false positives, but this means we’ll have more false negatives, thus decreasing sensitivity.

This is how we generate the ROC curve – sweep different thresholds and measure TPR and FPR on the dataset:

A graph on a lined paper

Description automatically generated

We can calculate AUC (the area under the curve) with a piecewise-trapezoidal approximation.

Ideally, the ROC curve is a vertical line starting from the origin, giving a sensitivity of 1 with an FPR of 0. In this case, AUC is 1.

### ROC curves for multiclass classification

There are two approaches: one-vs-rest (OVR) and one-vs-one (OVO).

For a given class, OVR treats the other classes as one big class. ROC/AUC is calculated per class.

OVO calculates ROC/AUC for each pair of classes.

## Gain and lift curves

<https://www.listendata.com/2014/08/excel-template-gain-and-lift-charts.html>

Gain, lift, and cumulative lift measure how much better your model is at identifying positive examples than randomly guessing.

How it works:

1. Run your test data through your model and get the probabilities of positive class
2. Sort the probabilities from high to low
3. Split the data into even bins. Lower bin index contains higher probabilities. Let be the total number of bins.
4. Let be the number of positive examples in each bin. There should be more positive examples for lower bin index.

Let be the total number of positive examples in your data, i.e.

Let be your dataset size. The baseline positive rate is

It’s your chance of picking a positive example if you’re just picking randomly.

Gain is the percentage of positive examples in the first bins. is the total number of positive examples in the first bins.

Gain tells you how many positive examples your model captures in the first bins. If your gain is 50%, then your model captures 50% of the positive examples in the first decile (bins=10). If your model is guessing randomly, then you would capture 10% in the first decile, 20% in the second, etc.

A graph with a red line and blue line

Description automatically generated

Cumulative lift is how many positive examples your model captures in the first bins divided by how many positive examples a random guesser captures in the first bins. Note that cumulative lift is always equal to 1 for the last bin.

Lift is how many positive examples your model captures in bin divided by how many positive examples a random guesser captures in a bin.

A graph of a lift curve

Description automatically generated

Practical applications (from ChatGPT):

1. Marketing Campaigns: To determine how much more effective a targeted marketing campaign is compared to a non-targeted one. (for example, targeting only the first bins gets you 90% of buyers)
2. Customer Retention: To prioritize efforts on the customers most likely to churn.
3. Fraud Detection: To focus investigations on the most likely fraudulent activities.

## Random guessing equivalents

Gain and lift curves explicitly measure how much better your model is than random guessing.

An AUC of 0.5, corresponding to a straight line from (0, 0) to (1, 1) is equivalent to random guessing. Specifically, random guessing is the point (0.5, 0.5), where TPR = 0.5 and FPR = 0.5.

This means half of your predictions are wrong for both true positives and true negatives.

# Regression

Root mean-square error (RMSE) – aka norm or Euclidean norm (Euclidean distance b/w hypothesis and label):

Mean absolute error (MAE) – aka norm or Manhattan norm (Manhattan distance, i.e. distance as if you were traversing city blocks):

Choosing LSE as the cost function means that you will minimize both LSE and RMSE through optimization. However, keep in mind that LSE is the cost function while RMSE is the performance metric.

You can choose any order of norm as your performance metric.

gives the number of nonzero elements in the vector, and gives the maximum absolute value in the vector (all other terms go to zero).

Higher means more sensitivity to outliers (large errors have more weight). When outliers are exponentially rare (like when samples are Gaussian-distributed), then RMSE gives you a more accurate picture of your prediction performance. This directly follows from generalized linear models and maximum likelihood estimate.