Reading: Evaluating Machine Learning Models

Define train/test split

Understanding the train/test split is fundamental in machine learning, particularly in supervised learning scenarios. This concept involves dividing your data set into two parts: the training and the test sets. While the training set educates the model on patterns within the data, the test set evaluates its ability to generalize to new, unseen data. For instance, imagine you're predicting house prices based on features like size, bedrooms, and location. You'd split your data into training and testing sets, allowing the model to learn from one and be tested on the other. This process ensures you can assess the model's performance accurately.

In this scenario, 80% of the data set is allocated for training (X_train and y_train), while the remaining 20% is allocated for testing (X_test and y_test). By specifying the random_state parameter, the split becomes reproducible. This means that executing the code with the same random_state value will consistently produce the same split, ensuring consistency across multiple runs.

Evaluate classification models using accuracy, a confusion matrix, precision, and recall

Accuracy: Accuracy quantifies the proportion of accurate predictions among all predictions generated by the model. For instance, in forecasting house prices, accuracy indicates the percentage of accurately forecasted prices among all prices within the data set. It's calculated by determining the proportion of instances that were predicted correctly out of the total instances.

Accuracy=Total number of predictionsNumber of correct predictionsAccuracy=Number of correct predictionsTotal number of predictions

Confusion matrix: A confusion matrix offers a concise overview of how well a classification model performs. By comparing the model's predictions to the actual outcomes, we determine the counts of true positives, true negatives, false positives, and false negatives. While directly applying a confusion matrix to predicting house prices may seem unconventional due to their continuous nature, you can discretize prices into categories (e.g., cheap, moderate, expensive), and then utilize the confusion matrix to compare predicted categories against actual ones.

A confusion matrix provides a summary of prediction results on a classification problem. The matrix shows the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

Actual/predicted	High price	Low price
High price	TP	FN
Low price	FP	TN

Precision: Precision is how many correct positive predictions the model made out of all its positive predictions. In the realm of house price prediction, precision reflects the accuracy of predicted house prices among all predictions. However, given the continuous nature of house prices, precision is commonly defined within a specific tolerance level. For instance, you might establish a threshold (e.g., within 5% of the actual price) and compute precision based on predictions falling within that range.

Precision=$TPTP+FP$Precision=*TP+FPTP*

Recall: Recall measures how many true positive predictions were made out of all actual positive instances. In the scenario of house price prediction, recall indicates the percentage of accurately predicted house prices among all actual house prices. Typically, recall is defined within a specified tolerance level to accommodate the continuous nature of house prices.

Recall=$TPTP+FN$Recall=*TP+FNTP*

Here is an illustrative example:

Assume your data set has been split into training and testing sets, and you have built a classification model. You run the model on the test set and obtain the following confusion matrix:

Actual/predicted	High price	Low price
High price	80	20
Low price	10	90

From this confusion matrix:

True positives (TP) = 80 (High price predicted as High price) In the provided data, out of 100 expensive houses, the model accurately identified 80 as high-priced.

True negatives (TN) = 90 (Low price predicted as Low price) In the given data, out of 100 nonexpensive houses, the model correctly identified 90 as low-priced.

False positives (FP) = 10 (Low price predicted as High price) In the given data, out of 100 non-expensive houses, the model incorrectly classified 10 as high-priced.

False negatives (FN) = 20 (High price predicted as Low price) In the given data, out of 100 expensive houses, the model incorrectly classified 20 as low-priced.

Now calculate the evaluation metrics:

Accuracy=$TP+TNTP+TN+FP+FN$=80+9080+90+10+20=170200=0.85=85%Accuracy=*TP*+*TN*+*FP*+*FNTP*+*TN*=80+90+10+2080+90=200170=0.85=85%

Precision=$TPTP+FP$=8080+10=8090≈0.89=89%Precision=*TP*+*FPTP*=80+1080=9080≈0.89=89%

Recall=$TPTP+FN$=8080+20=80100≈0.80=80%Recall=*TP*+*FNTP*=80+2080=10080≈0.80=80%

These metrics help you understand the performance of your classification model, allowing you to make informed decisions about model improvements or adjustments.

Evaluate a regression model using mean squared error and other kinds of error terms

Evaluating a regression model involves assessing how well the model predicts the target variable. Here, we'll discuss how to evaluate a regression model using mean squared error (MSE) and other error metrics like mean absolute error (MAE), root mean squared error (RMSE), and R-squared (R^2).

- **Mean squared error (MSE):**
    - MSE computes the mean of squared errors, representing the average squared disparity (difference) between the estimated values and the actual value.
    - Formula:
      MSE=1n∑i=1n(yi−yi^)2MSE=n1∑i=1n(yi−yi^)2
    - Lower MSE indicates better model performance.

- **Mean absolute error (MAE):**
    - MAE calculates the average size of the errors in a set of predictions, without taking into account their direction.
    - Formula:
      MAE=1n∑i=1n|yi−yi^|MAE=n1∑i=1n|yi−yi^|
    - Lower MAE indicates better model performance.
- **R-squared (R^2):**
    - R^2 tells us how much of the dependent variable's changes are explained by changes in the independent variables.
    - Formula:
      R2=1−∑i=1n(yi−y¯)2R2=1−∑i=1n(yi−y¯)2∑i=1n(yi−yi^)2
    - Higher R^2 values, which range from 0 to 1, signify better model performance.
- **Root mean squared error (RMSE):**
    - RMSE, as the square root of MSE, provides an insight into the typical size of errors.
    - Formula:
      RMSE=MSERMSE=MSE
- **Mean absolute percentage error (MAPE):**
    - MAPE quantifies accuracy as a percentage of the error.
    - Formula:
      MAPE=1n∑i=1n|yi−yi^yi|×100MAPE=n1∑i=1n||$yiyi$−yi^||×100
- **Median absolute error:**
    - Median absolute error is robust to outliers because it uses the median instead of the mean.
    - Useful when your data has outliers that could skew the error metrics.

Data set example

Here's a small dataset for illustration:

Size (sqft)	Number of bedrooms	Location	Actual price ($)	Predicted price ($)
1500	3	1	300000	310000
1600	3	2	320000	315000
1700	4	1	350000	345000
1800	4	3	370000	375000
1900	5	2	400000	390000

Mean squared error (MSE):

Calculation:

MSE = (1/5) * [$(300000 - 310000)^2$ + $(320000 - 315000)^2$ + $(350000 - 345000)^2$ + $(370000 - 375000)^2$ + $(400000 - 390000)^2$]

MSE = (1/5) * [$(-10000)^2$ + $(5000)^2$ + $(5000)^2$ + $(-5000)^2$ + $(10000)^2$]

MSE = (1/5) * [100,000,000 + 25,000,000 + 25,000,000 + 25,000,000 + 100,000,000]

MSE = 275,000,000 / 5

MSE = 55,000,000

Mean absolute error (MAE):

Calculation:

MAE = (1/5) * (|300000 - 310000| + |320000 - 315000| + |350000 - 345000| + |370000 - 375000| + |400000 - 390000|)

MAE = (1/5) * (10000 + 5000 + 5000 + 5000 + 10000)

MAE = 35000 / 5

MAE = 7000

Root mean squared error (RMSE):

Calculation:

MSE = (1/5) * [(300000 - 310000)^2 + (320000 - 315000)^2 + (350000 - 345000)^2 + (370000 - 375000)^2 + (400000 - 390000)^2]

MSE = (1/5) * [(-10000)^2 + (5000)^2 + (5000)^2 + (-5000)^2 + (10000)^2]

MSE = (1/5) * [100,000,000 + 25,000,000 + 25,000,000 + 25,000,000 + 100,000,000]

MSE = 275,000,000 / 5

MSE = 55,000,000

RMSE = sqrt(55,000,000)

RMSE ≈ 7,416.20

More about R-squared (R^2):

R-squared, or the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable (house price) that can be predicted from the independent variables (size, number of bedrooms, and location). It serves as a key indicator of how well a regression model fits the data.

Formula

$R2=1-\sum i=1n(yi-yi\hat{})2\sum i=1n(yi-\bar{y})2R2=1-\sum i=1n(yi-\bar{y})2\sum i=1n(yi-yi\hat{})2$
where,

- $(yi)(yi)$ = Actual value of the dependent variable (house price)
- $(y\hat{}i)(y\hat{}i)$ = Predicted value of the dependent variable
- $(\bar{y})(\bar{y})$ = Mean of the actual values of the dependent variable
- $(n)(n)$ = Number of observations
- $(\sum i=1n(yi-y\hat{}i)2)(\sum i=1n(yi-y\hat{}i)2)$ = Sum of squares of residuals (SSR) or residual sum of squares (RSS)
- $(\sum i=1n(yi-\bar{y})2)(\sum i=1n(yi-\bar{y})2)$ = Total sum of squares (TSS)

Interpretation

- **$R^2 = 1$**: The model explains all the variability of the response data around its mean. The predictions perfectly match the actual data.
- **$R^2 = 0$**: The model doesn't account for any variation in the response data around its mean. The predictions are as good as simply using the mean of the actual data.
- **$0 < R^2 < 1$**: The model explains a portion of the variability, with higher values indicating a better fit.
- **$R^2 < 0$**: This can occur if the model is worse than a horizontal line (mean of actual values), which typically indicates an incorrect model.

Numerical calculation

Let's assume the following hypothetical test set and predictions:

- Actual prices
  ($y$test)($y$test) = [370000, 320000]
- Predicted prices
  ($y$pred)($y$pred) = [360000, 310000]

Steps and calculations:

1. Calculate the mean of ($y$test):[$\bar{y}=\sum y\text{test}n$]($y$test):[$\bar{y}=n\sum y$test]

Given ($y$test)($y$test) = [370000, 320000]
$\bar{y}$ $\bar{y}$ = (370000 + 320000)/2 = 345000

2. Calculate the total sum of squares ((SStot)): $SS\text{tot}=\sum i=1n(yi-\bar{y})2$(($S$Stot)): SStot =$\sum i=1n(yi-\bar{y})2$
   SStot SStot = (370000 - 345000)^2 + (320000 - 345000)^2
   SStot SStot = (25000)^2 + (-25000)^2
   SStot SStot = (625000000 + 625000000) = 1250000000

3. Calculate ($R2$):[$R2=1-SS\text{res}SS\text{tot}$]($R2$):[$R2=1-S$Stot$S$Sres]
   $R2$ $R2$= 1 - (200000000/1250000000)
   $R2$ $R2$= 1 - 0.16
   $R2$ $R2$= 0.84