# Model Evaluation

Department of Computer Languages and Systems

# Introduction

- model evaluation can be performed in supervised learning as the actual values are available
- there is a fundamental difference between the methods used for evaluating a regressor and a classifier:
  - in regression, we deal with continuous values and the focus is on the error between the actual and predicted values
  - in classification, the focus is on the number of data that are classified correctly

# Classification models

- In classification, we deal with two types of models:
  - in some classifiers (e.g., *k*-NN, SVM), the output is simply the class label
  - in others (e.g., logistic regression, random forest), the output is the probability of a data point belonging to a particular class where through the use of a cut off value we are able to convert these probabilities into class labels

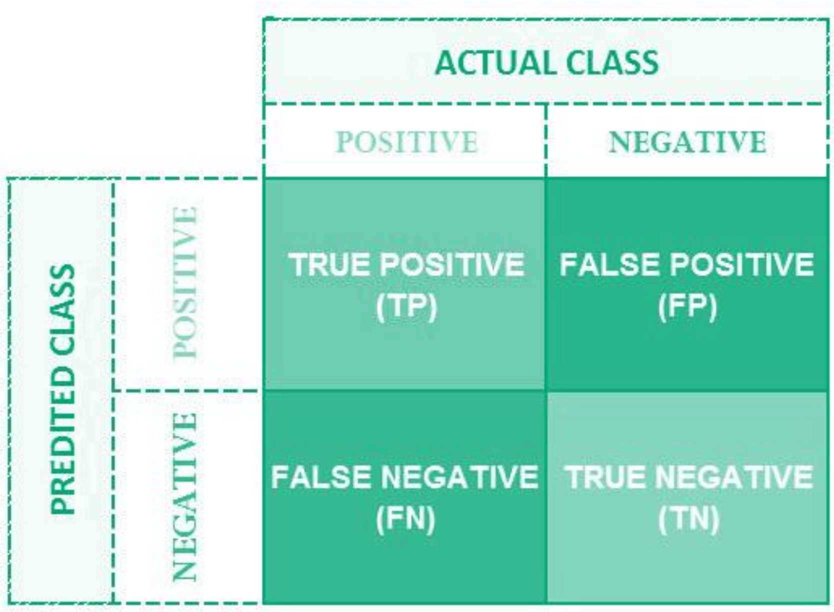
# Confusion matrix

Suppose a data set with class labels 0 (‘negative’ class) and 1 (‘positive’ class), and a classifier that can produce correct and incorrect predictions:

- True positive: the number of positive observations that were correctly predicted by the model
- True negative: the number of negative observations that were correctly predicted by the model
- False positive: the number of negative observations that were incorrectly predicted as positive by the model
- False negative: the number of positive observations that were incorrectly predicted as negative by the model

# Confusion matrix (ii)

For a two-class problem, the confusion matrix will have two rows and two columns where the rows represent the actual values and the columns represent predicted values



| | | ACTUAL CLASS | |
|---|---|---|---|
| | | POSITIVE | NEGATIVE |
| PREDITED CLASS | POSITIVE | TRUE POSITIVE<br>(TP) | FALSE POSITIVE<br>(FP) |
| | NEGATIVE | FALSE NEGATIVE<br>(FN) | TRUE NEGATIVE<br>(TN) |

Our objective is to minimize the FP and FN and maximize the TP and TN, that is, to maximize the diagonal values

# Confusion matrix (iii)

From the confusion matrix, various evaluation metrics can be calculated. The most basic measures are accuracy and classification error:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Error = \frac{FP + FN}{TP + TN + FP + FN} = 1 - Accuracy$$

# Confusion matrix (iv)

**Accuracy and classification error can be very misleading**

For instance, suppose a data sets with 100 instances where 90 belong to class ‘0’ and the remaining 10 belong to class ‘1’. A model that predicts all instances to class ‘0’ will have an accuracy of 90%, but the accuracy here provides us with little information on how actually the model is performing

**Thus, we need other more complex measures to evaluate the models in some problems (e.g., class imbalance)**

# Other evaluation measures

- False positive rate (false alarm): the proportion of negative observations that were wrongly (as positive) predicted by the model

$$FPrate = \frac{FP}{FP + TN}$$

- False negative rate (miss rate): the proportion of the positives that were classified as negative by the model

$$FNrate = \frac{FN}{FN + TP}$$

# Other evaluation measures (ii)

- True positive rate (recall, sensitivity): the proportion of positives that were classified correctly by the model. It is the opposite of FNrate

$$TPrate = \frac{TP}{TP + FN} = 1 - FNrate$$

- True negative rate (specificity): the proportion of negatives that were classified correctly by the model. It indicates how good the model is at avoiding the misclassification of the negatives

$$TNrate = \frac{TN}{TN + FP} = 1 - FPrate$$

# Other evaluation measures (iii)

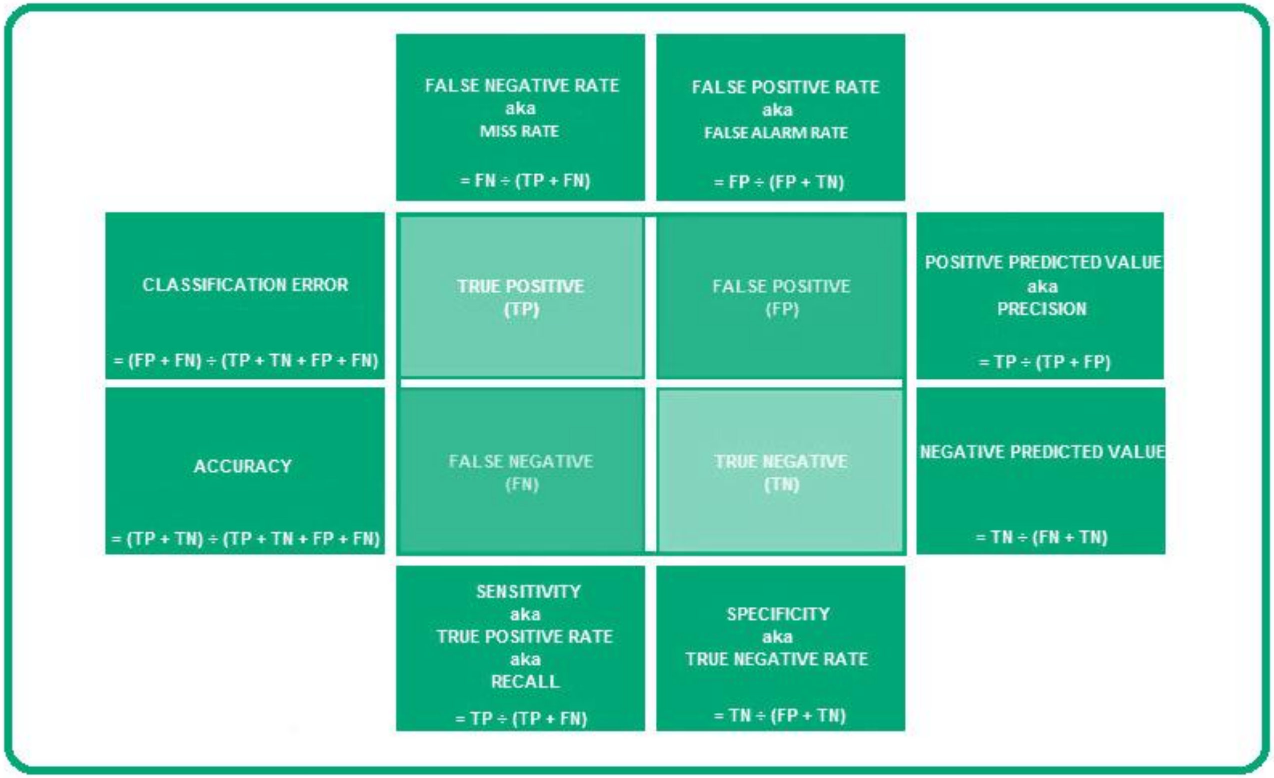
- Precision (positive predicted value): the proportion of the positive out of what the model predicted as positive. Unlike TPrate, Precision provides us with the percentage of TP out of all predicted positives

$$Precision = \frac{TP}{TP + FP}$$

- Negative predicted value: the proportion of the negative out of what the model predicted as positive

$$NPV = \frac{TN}{TN + FN}$$

# Other evaluation measures (iv)



| | | | |
|---|---|---|---|
| | FALSE NEGATIVE RATE<br>aka<br>MISS RATE<br>= FN ÷ (TP + FN) | FALSE POSITIVE RATE<br>aka<br>FALSE ALARM RATE<br>= FP ÷ (FP + TN) | |
| CLASSIFICATION ERROR<br>= (FP + FN) ÷ (TP + TN + FP + FN) | TRUE POSITIVE<br>(TP) | FALSE POSITIVE<br>(FP) | POSITIVE PREDICTED VALUE<br>aka<br>PRECISION<br>= TP ÷ (TP + FP) |
| ACCURACY<br>= (TP + TN) ÷ (TP + TN + FP + FN) | FALSE NEGATIVE<br>(FN) | TRUE NEGATIVE<br>(TN) | NEGATIVE PREDICTED VALUE<br>= TN ÷ (FN + TN) |
| | SENSITIVITY<br>aka<br>TRUE POSITIVE RATE<br>aka<br>RECALL<br>= TP ÷ (TP + FN) | SPECIFICITY<br>aka<br>TRUE NEGATIVE RATE<br>= TN ÷ (FP + TN) | |

All these metrics are useful, but they alone cannot provide us with a good evaluation score. Thus, we need to consider these values in pairs such as Recall-Precision, FPrate-FNrate, TPrate-TNrate, etc.

# Other evaluation measures (v)

- Geometric mean of accuracies: it achieves the maximum value when the accuracy on each of the classes is maximum while keeping these accuracies balanced

$$Gmean = \sqrt{TPrate \cdot TNrate}$$

- Mean class weighted accuracy:

$$CWA = w \times TPrate + (1 - w) \times TNrate$$

where $w \in [0,1]$

# Other evaluation measures (vi)

- *F*-measure:

$$F - measure = \frac{(1+\beta^2)\cdot recall \cdot precision}{\beta^2 \cdot recall + precision}$$

where $\beta \in [0,1]$. If precision and recall are equally important, then $\beta$ is set to 1, and *F*-measure is known as the $F_1$-measure

- Area under the ROC curve:

$$AUC = \frac{TPrate + TNrate}{2}$$

- Gini coefficient:

$$Gini = 2 \times AUC - 1$$

# Regression models

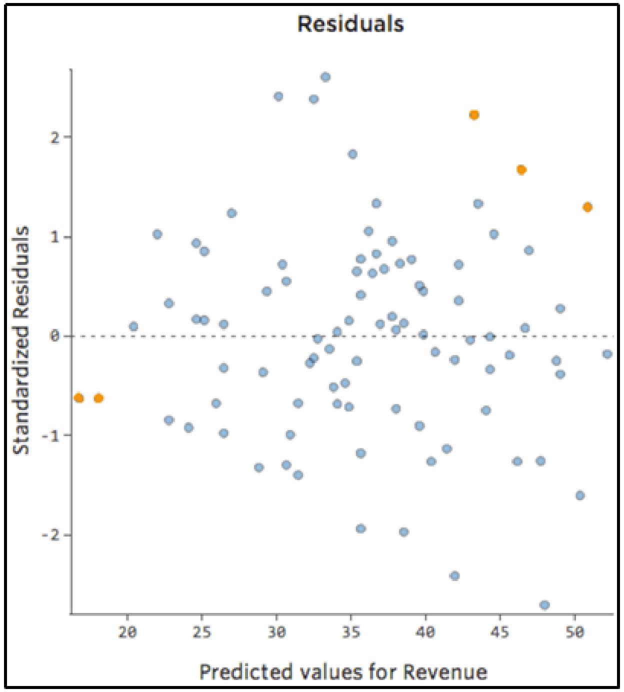
- In regression, evaluation methods depend on the ways of calculating the residual (difference between the actual and predicted values):
  - some methods (e.g., sum squared error, mean squared error and root mean squared error) just calculate this difference
  - others (e.g., adjusted coefficient of determination) also take into account the problem of overfitting

# Regression models (ii)

- In regression, unlike classification (where we can count the number of observations correctly classified), our predictions are either bigger or smaller than the actual value (rarely it is same as the actual value)
- Thus, we are not concerned with how many times we were wrong but rather what matters is the quantum of residuals
- To validate a regression model, you must use residual plots to visually confirm the validity of the model

# Regression models (iii)

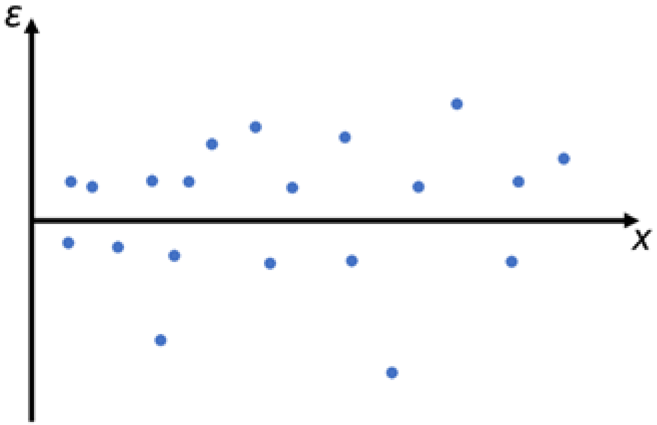
Residual plots: the residual values on the y-axis and the predicted values on the x-axis.

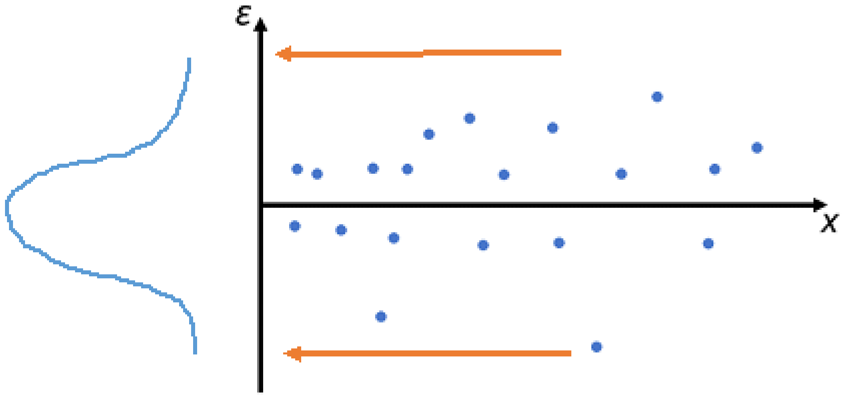


- if it has a high density of points close to the origin and a low density of points away from the origin and they are symmetrically distributed around the horizontal axis, the regression model is appropriate

# Regression models (iv)

How to know whether a residual plot is good based on the characteristics just presented?





If we project all the residuals onto the y-axis, we end up with a normally distributed curve. This satisfies the assumption that regression model **residuals are independent and normally distributed**

# Measures to compare multiple data sets

- Mean/median of prediction: we can understand the bias in prediction between two models using the arithmetic mean of the predicted values. If there are outliers, it is better to use the median
- Standard deviation of prediction: it is a measure of the amount of variation or dispersion of a set of values

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

A low SD indicates that the values tend to be close to the mean (also called the expected value) of the set. In contrast, a high SD indicates that the values are spread out over a broader range.

# Measures to compare multiple data sets (ii)

- Range of prediction: it is the maximum and minimum value in the predicted values, helping to understand the dispersion between models

- Coefficient of variation (relative standard deviation): it simply represents the variance standardized by its mean

$$CV(\%) = \frac{SD}{mean} \times 100$$

SD is the most common measure of variability for a single data set, but it is meaningless when comparing two different data sets. In such cases, CV is the method of choice

# Measures to compare multiple data sets (iii)

For instance, consider two different data sets:

- Data 1: Mean1 = 120000 : SD1 = 2000
- Data 2: Mean2 = 900000 : SD2 = 10000

Let us calculate CV for both datasets:

- CV1 = SD1/Mean1x100 = 1.6%
- CV2 = SD2/Mean2x100 = 1.1%

We can conclude Data 1 is more spread out than Data 2

# Common evaluation measures

- Sum squared error: it calculates the difference between the true ($y_i$) and predicted ($\hat{y}_i$) values across the *n* test observations

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Mean squared error: it is the SSE divided by the number of test observations. This is the most common metric for regression

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

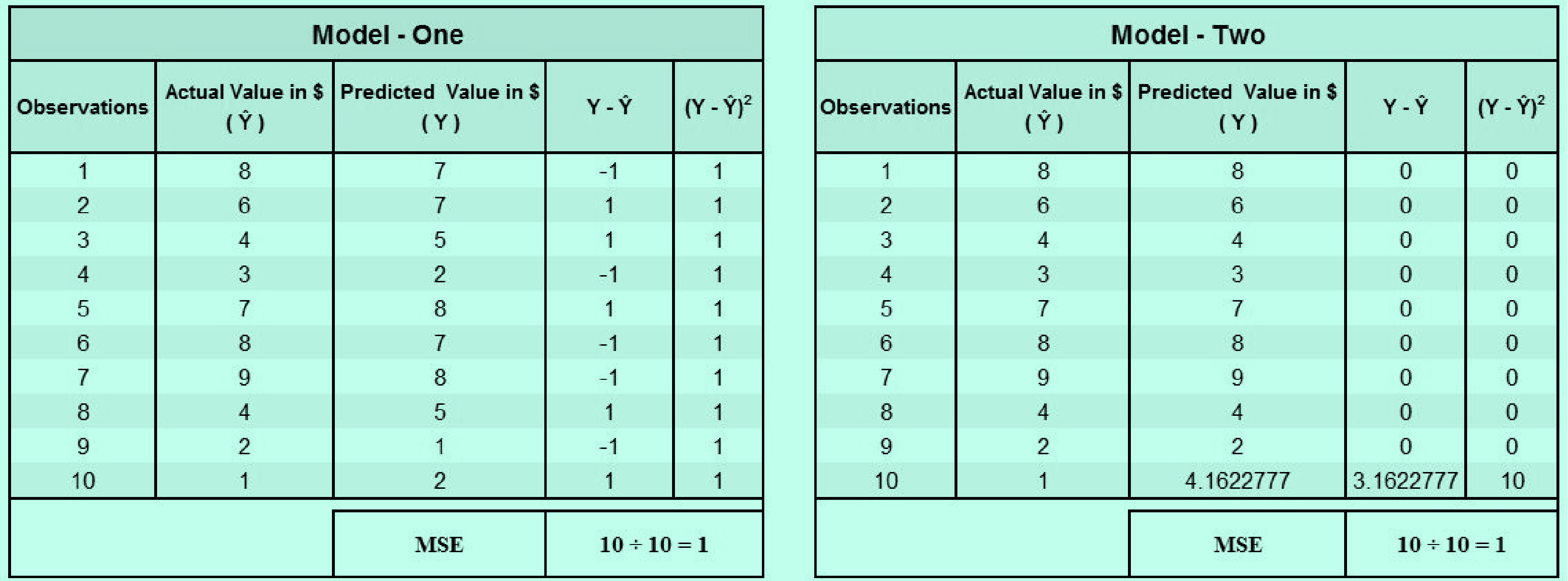
- Root mean squared error: it is the square root of MSE

$$RMSE = \sqrt{MSE}$$

# Common evaluation measures (ii)

- As lower the MSE/RMSE value is, the better the model is with predictions
- A higher MSE/RMSE indicates that there are large deviations between the predicted and actual value
- Drawbacks of MSE/RMSE:
  - sensitive to outliers
  - a single big error will have the same effect as a lot of small errors

# Common evaluation measures (iii)



| Model - One | | | | |
|---|---|---|---|---|
| Observations | Actual Value in \$ ( $\hat{Y}$ ) | Predicted Value in \$ ( Y ) | $Y - \hat{Y}$ | $(Y - \hat{Y})^2$ |
| 1 | 8 | 7 | -1 | 1 |
| 2 | 6 | 7 | 1 | 1 |
| 3 | 4 | 5 | 1 | 1 |
| 4 | 3 | 2 | -1 | 1 |
| 5 | 7 | 8 | 1 | 1 |
| 6 | 8 | 7 | -1 | 1 |
| 7 | 9 | 8 | -1 | 1 |
| 8 | 4 | 5 | 1 | 1 |
| 9 | 2 | 1 | -1 | 1 |
| 10 | 1 | 2 | 1 | 1 |
| | | MSE | 10 ÷ 10 = 1 | |

| Model - Two | | | | |
|---|---|---|---|---|
| Observations | Actual Value in \$ ( $\hat{Y}$ ) | Predicted Value in \$ ( Y ) | $Y - \hat{Y}$ | $(Y - \hat{Y})^2$ |
| 1 | 8 | 8 | 0 | 0 |
| 2 | 6 | 6 | 0 | 0 |
| 3 | 4 | 4 | 0 | 0 |
| 4 | 3 | 3 | 0 | 0 |
| 5 | 7 | 7 | 0 | 0 |
| 6 | 8 | 8 | 0 | 0 |
| 7 | 9 | 9 | 0 | 0 |
| 8 | 4 | 4 | 0 | 0 |
| 9 | 2 | 2 | 0 | 0 |
| 10 | 1 | 4.1622777 | 3.1622777 | 10 |
| | | MSE | 10 ÷ 10 = 1 | |

The predictions of model-1 had an error of 1 for all 10 samples while for model-2, 9 out of 10 samples were predicted correctly but one prediction was off by 4.16 → the single big error made by model-2 has the same impact as the 10 small errors made by model-1

# Other evaluation measures

- Relative squared error: it can be used to compare between models whose errors are measured in different units. It takes the total squared error of our model and normalizes by the total squared error of a model that uses the mean as the predictor

$$RSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

RSE > 1 indicates that our model is not even as good as a model that simply predicts the mean as the prediction for each observation. RSE = 0 corresponds to the ideal case.

# Other evaluation measures (ii)

- Mean absolute error: it is similar to MSE, but instead of squaring the difference between the actual and predicted values, this takes an absolute value of it

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Unlike MSE/RMSE, here a big error does not overpower a lot of small errors → the output provides us with a relatively unbiased understanding of how the model is performing

# Other evaluation measures (iii)

- Median absolute error: it is the median of all absolute differences between the actual and predicted values

$$MedAE = median_{i=1}^{n}\left(|y_i - \hat{y}_i|\right)$$

- Mean absolute percentage error: it is calculated as the absolute difference between the actual and predicted values divided over every observation

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100$$

# Other evaluation measures (iv)

- Relative absolute error: like RSE, it is also used to compare between models whose errors are measured in different units

$$RAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

For a perfect fit, the numerator is equal to 0 and $\mathrm{RAE} = 0$. So, the RAE ranges from 0 to infinity, with 0 corresponding to the ideal case.

# Other evaluation measures (v)

- Coefficient of determination (*R*$^2$): this is a statistical measure that determines how well a regression model predicts an outcome. This means how good is a model for a data set

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The numerator is the sum of squared residuals and the denominator is the total sum of squares.

*R*$^2$ lies between 0 and 1. *R*$^2$ = 0 indicates that the model fails to accurately model the data at all. The closer its value to 1, the better a model is.

# Other evaluation measures (vi)

- Adjusted coefficient of determination: it penalizes adding more independent variables which do not increase the explanatory power of the regression model. It is always less than or equal to the value of $R^2$

$$adj\_R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where *p* is the number of independent variables and *n* the sample size.