MACHNINE LEARNING WORKSHEET

C

B

C

B

6. B

8. D

9. A,B,D

10. A,C,D

11. In data analytics, outliers are values within a dataset that vary greatly from the others—they're either much larger, or significantly smaller. Outliers may indicate variabilities in a measurement, experimental errors, or a novelty. In a real-world example, the average height of a giraffe is about 16 feet tall. However, there have been recent discoveries of two giraffes that stand at 9 feet and 8.5 feet, respectively. These two giraffes would be considered outliers in comparison to the general giraffe population.

When going through the process of data analysis, outliers can cause anomalies in the results obtained. This means that they require some special attention and, in some cases, will need to be removed in order to analyse data effectively.

There are two main reasons why giving outliers special attention is a necessary aspect of the data analytics process:

Outliers may have a negative effect on the result of an analysis

Outliers—or their behaviour—may be the information that a data analyst requires from the analysis.

IQR = The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by $Q_1$ known as the lower quartile, the second Quartile is denoted by $Q_2$ and the third Quartile is denoted by $Q_3$ known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

Interquartile range = Upper Quartile – Lower Quartile = $Q_3 – Q_1$

12. Bagging and Boosting get N learners by generating additional data in the training stage. N new training data sets are produced by random sampling with replacement from the original set. By sampling with replacement some observations may be repeated in each new training data set.

In the case of Bagging, any element has the same probability to appear in a new data set. However, for Boosting the observations are weighted and therefore some of them will take part in the new sets more often.

13. Linear regression is a common tool that the pharmacokinetics uses to calculate elimination rate constants. Standard linear regression provides estimates for the slope, intercept, and $r^2$, a statistic that helps define goodness of fit. Statistical texts define $r^2$ as the coefficient of determination and it is calculated using the following equation:

$$r^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

where SS = the sum of squares for either the residuals or the total (original data). As the residuals get smaller, $r^2$ gets larger to a maximum value of 1.

Another way to think of $r^2$ is to consider that it expresses the amount of variability in Y that is explained by X, given the selected mathematical model. To put that into pharmacokinetic terms, $r^2$ defines the amount of variability in concentration (Y) that is explained by time (X) using a monoexponential decline equation $(C = C_0 * e^{-kt})$.

When performing linear regression, the slope and intercept parameters are chosen to maximize $r^2$ which defines the “best-fit” of the data. This accepted methodology ensures that combination of slope and intercept parameters explain as much of the variability in concentrations that are observed for the selected data points. While this method is accepted for well-defined datasets, in pharmacokinetic analysis, the terminal slope of a pharmacokinetic curve may include 3, 4, 5, 6, or even more data points. And the option to select the number of datapoints is somewhat arbitrary. The addition of data to the model often increases the $r^2$ value by virtue of simply adding datapoints. As $SS_{total}$ increases, $r^2$ decreases, even if $SS_{residuals}$ does not decrease.

To address this concern, a new statistic called “adjusted” $r^2$ was developed. This new statistic essentially issues a penalty for each additional data point in the analysis. This has the effect of requiring the additional data point to improve the $r^2$ by more than just decreasing $SS_{total}$. The “adjusted” $r^2$ is calculated using the following equation:

$$Adjusted\ r^2 = 1 - (1 - r^2) * \frac{(n-1)}{(n-2)}$$

where n = the number of datapoints used in the regression. At very large values of n, adjusted $r^2$ is equivalent to $r^2$. However, at small values of n that are used in pharmacokinetic analysis (e.g. <10), the adjusted $r^2$ can be significantly different from $r^2$. For example, moving from 4 data points to 5 data points, the adjusted $r^2$ statistic is multiplied by 0.75, or given a penalty of 25%!

Thus the “adjustment” is related to selecting the right amount of data to include in the analysis. This statistic is biased toward selecting the fewest amount of data points while maximizing the coefficient of determination. Maximizing the adjusted $r^2$ when performing terminal slope regressions selects the best set of slope and intercept parameters with the

fewest number of data points. Many consider the use of adjusted $r^2$ as the optimal method for selecting a terminal rate constant for pharmacokinetic data.

To learn about how we've improved Phoenix to make performing NCA and PK/PD modeling even easier, please watch this webinar I gave on the latest enhancements to Phoenix.

14. The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a range of [0,1]. *Standardization* typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

In this blog, I conducted a few experiments and hope to answer questions like:

Should we always scale our features?

Is there a single best scaling technique?

How different scaling techniques affect different classifiers?

Should we consider scaling technique as an important hyperparameter of our model?

I'll analyse the empirical results of applying different scaling methods on features in multiple experiments settings.

15. Cross-Validation (CV) is one of the key topics around testing your learning models. Although the subject is widely known, I still find some misconceptions cover some of its aspects. When we train a model, we split the dataset into two main sets: training and testing. The training set represents all the examples that a model is learning from, while the testing set simulates the testing examples as in Figure 1.