Something about Machine Learning

writer : Yitong Lu

Nowadays machine learning is a popular and renowned field that is widely applied in manufacturing, retail, healthcare and life sciences, travel and hospitality, financial services, and energy, feedstock, and utilities. Machine learning is a part of artificial intelligence (AI) and computer science that concentrates on the use of data and algorithms to study the human's behavior. It will become more accurate and better in the process of repeating itself in study. Undoubtedly, data is an important part of machine learning. The observations of the data will be absorbed and interpreted by the machine learning model in its own way. It will be treated as training and testing data in deciding machine-learning models. There are many forms of data such as numerical, categorical, time-series, and so on. Algorithms use data to learn patterns and relationships between independent variables and dependent variables, which is used for prediction or classification tasks.

There are two different types of data which is one with outcome (dependent variables) and one without it. We called it labeled data and unlabeled data which also extend to supervised and unsupervised learning. Supervised machine learning depends on labeled input and output training data, whereas unsupervised learning processes unlabelled or raw data. They have different approaches to the data and the model.

Supervised learning has algorithms such as simple linear regression, decision tree regression, logistic regression and random forest. Data types mainly focus on binary

and numeric. The input and output of the data will be labeled, so the model can understand which features can classify an object better or data point with different class labels. Linear regression is commonly used in analyzing survey data. For example, a company uses linear regression to figure out how likely people will buy their product again after the first purchase.

Unsupervised learning is commonly used in using clustering. Clustering is the grouping together of data points into a determined number of categories depending on similarities (or differences) between data points. This way raw and unlabelled data can be processed and clustered depending on the patterns within the dataset. There are K-means clustering and Gaussian Mixture Models. In K-means clustering, K represents the count of clusters. A higher count of clusters means more granular groupings, and a lower count of clusters means less granular groupings.Gaussian Mixture Models is an example of an approach to probabilistic clustering, in which data points are grouped based on the probability that they belong to a defined grouping.

Machine learning takes a crucial role in the field of data science. Most tasks that can be defined as data's pattern can be automated by machine learning. Machine learning will relate and close more and more to daily life and daily things in the future. The self-driving car uses image-detection techniques to fulfill human's desires. The image-detection will recognize and identify an object or feature in an image. An image will split to many individual pixels. The system will create a network that saves as many pre labeled images as we want, creating a process to make the system recognize similar images. AI techniques are definitely making our life better

and easier. However, we still need to utilize and apply AI in a proper way. In the process of exploring and evolving AI techniques, we still need to keep our humanization and treat AI rationally.

Example of Machine Learning(-KNN classifier):

```
breastcancer <- read_csv("C:/Users/guoto/Downloads/BreastCancer.csv")
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 569 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): diagnosis
## dbl (11): ...1, radius_mean, texture_mean, perimeter_mean, area_mean, smooth...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dim(breastcancer)
```

```
## [1] 569  12
```

```
## Warning: package 'class' was built under R version 4.1.3
```

```
breast.knn1 <- knn(X.train,X.test,y.train,k=1)
breast.knn2 <- knn(X.train,X.test,y.train,k=3)
breast.knn3 <- knn(X.train,X.test,y.train,k=5)
breast.knn4 <- knn(X.train,X.test,y.train,k=7)
breast.knn5 <- knn(X.train,X.test,y.train,k=9)
breast.knn6 <- knn(X.train,X.test,y.train,k=11)
```

(d)

```
#knn with variables
comparison.knn <- list("breast.knn1"=c(mean(breast.knn1!= y.test),mean(breast.knn1==y.test)),
"breast.knn2"=c(mean(breast.knn2!= y.test),mean(breast.knn2==y.test)),
"breast.knn3"=c(mean(breast.knn3!= y.test),mean(breast.knn3==y.test)),
"breast.knn4"=c(mean(breast.knn4!= y.test),mean(breast.knn4==y.test)),
"breast.knn5"=c(mean(breast.knn5!= y.test),mean(breast.knn5==y.test)),
"breast.knn6"=c(mean(breast.knn6!= y.test),mean(breast.knn6==y.test)))
comparison.knn
```

```
## $breast.knn1
## [1] 0.1420118 0.8579882
##
## $breast.knn2
## [1] 0.112426 0.887574
##
## $breast.knn3
## [1] 0.1005917 0.8994083
##
## $breast.knn4
## [1] 0.1005917 0.8994083
##
## $breast.knn5
## [1] 0.07692308 0.92307692
##
## $breast.knn6
## [1] 0.08284024 0.91715976
```

from the misclassfication rate we can see for best predictors when k=9 is the best. The misclassification rate is 0.07692308.