Answers for Assessment of Data and Business Analytics Lab. Industrial & Management Engineering HUFS

Muhammad Rasyid Ridha

26 September 2018

Answer 1.a.

What do you know about supervised, unsupervised, and semi-supervised learning? How do they relate each other?

Supervised learning is a type of machine learning where the target variable (also known as dependent, response or outcome variable) is available. Most of machine learning problem is supervised learning. The purpose is to predict the target variable by training the data given the predictor variables (also known as features, independent or input variable).

There are two types of supervised learning: regression and classification. Regression is used when the target variable is continous, e.g. height, temperature, amount of sales, currency, etc. In other hand, classification is used when the target variable is categorical, e.g. gender, yes or no, class, etc. There are many techniques of supervised learning algorithm. The popular one are regression-based and tree-based algorithm. Some use case examples of supervised learning problems are:

- Gender prediction
- House price prediction
- Credit scoring
- Demand forecasting
- Churn prediction
- Image classification

In other hand, when the target variable is not exist or unknown, unsupervised learning can be used. The purpose is for exploratory data analysis, group similar data, features generation and understand the hidden pattern of the data. It will form clusters from the data based on similarity or dissimilarity metrics.

The popular unsupervised learning technique are k-means and hierarchical clustering. When using k-means, user needs to specify number of clusters based on the context, problems, purpose and business understanding. Different with k-means, hierarchical clustering does not require number of clusters. Some use case examples of unsupervised learning problems are:

- Customer segmentation
- Text clustering
- Geospatial clustering

Supervised and unsupervised learning can be related each others. For example, we can find clusters from the data using unsupervised learning algorithm and use the clusters as the target variable that will be predicted using supervised learning classification algorithm.

In addition, unsupervised learning can be used as features generation or data preprocessing that might be useful to improve prediction result of a supervised learning problem.

The combination between supervised and unsupervised learning is semi-supervised learning. It is a class of supervised learning that can be used when the target variable has very limited labeled data. One of semi-supervised learning technique is by utilizing unsupervised learning to generate clusters or pseudo-label of the data, known as cluster-and-label method (Zhu 2008).

In a real application, labeling the data can be expensive and unlabeled data can be obtained easily. Semi-supervised learning technique can be used to tackle this limitation with careful assumptions and sufficient samples of data.

Answer 1.b.

What would be the best learning approach (in accordance to question 1.a.) to do activity recognition research if you have only two devices; smartphone and smartwatch?

It depends on the data that we collect. If the data has target label, it is clear that the best learning approach is supervised learning. It is a classification problem with multiclass target variable.

Using only smartphone and smartwatch, it is possible to label the activity of the user manually. There are some android apps that can be used to log activity. One of the open-source app to log activity is made by Lathia (2016).

However, sometimes the target label is not exist or limited. In this case, the label of activity is not exist or limited. Hence, using unsupervised or semi-supervised learning approach can be the alternative. Unsupervised learning can be used to identify hidden pattern from the data and form the cluster of activity pattern. In addition, when we have very limited target label, we can use semi-supervised approach.

In the end, even though we use unspervised or semi-supervised learning, the purpose is to predict activity based on the data generated by the sensor. It is important to have the labeled dataset and use supervised learning for this case.

Answer 2.a

Features extraction

Answer 2.b

The modeling is done using R with new currently developed machine learning package, `tidymodels`, created by Max and Wickham (2018). The steps of reproducing the result are:

1. Install R, RStudio and dependency of packages using `install_pkg.R`
2. Run `model.R` to preprocess, split train and test data, train the data and save the model and prediction in the `output` folder. The algorithm used in `model.R` is linear SVM.
3. Run `eval.R` to see the evaluation metrics (accuracy and F1-score) by running this in terminal.

```
$ Rscript eval.R
```

Notes

The answer presented here is the simple one. There are many ways to improve the model, such as:

1. Experiment with different features extraction (dimension reduction, different sliding window, clustering)
2. Try different models
3. Tune hyperparameters
4. Use cross-validation for more robust evaluation metrics

References

Lathia, Neal. 2016. "Mining Smartphone Sensor Data with Python." https://github.com/nlathia/pydata_2016/tree/master/App/AccelerometerCollector.

Max, Kuhn, and Hadley Wickham. 2018. *Tidymodels: Easily Install and Load the 'Tidymodels' Packages*. https://CRAN.R-project.org/package=tidymodels.

Zhu, Xiaojin. 2008. "Semi-Supervised Learning Literature Survey." http://legacydirs.umiacs.umd.edu/~hal/courses/2011F_ML/out/ssl_survey.pdf.