# Problem-Solving With Machine Learning

**Project Part One: Frame a Machine Learning Problem**

**Instructions:** Think of a problem that you want to solve with machine learning. Frame this problem like a data scientist by answering the following questions. Please limit your answers to 100 words or less.

|  |
| --- |
| What is the **context** of the problem? Why is it important to solve this problem? |
| I am now working on a machine learning project which the scope is to identify Store Hourly associates (employees) that have a high risk of leaving the company by using internal and external data. |
| What inputs (**features**) would you include? Describe **data types** and possible feature **transformations**. |
| There are many columns in the internal dataset, and the features I am going to use are:  1. Region number – int – define stores are in which region  2. Time in role – float – how long does associate stay in the current role 3. Tenure Category – float – how long does associate join Walmart 4. Better Wage – string – if associate has a higher wage than store average 5. Associate type – string – including part time and full-time employee status only |
| What are the potential outputs (**labels**)? |
| I want to predict on the employee status  Employee status – string – showing if it is an active or terminated employee |
| Are **observations dependent** anyhow? Often, we have either temporal dependency or multiple observations per patient or device or event. This is important for train/test splitting. |
| There are some features that might related other features. For example, in my dataset, I have zip code, store number, region number and state code. These are related features in the dataset and might cause issue later in the model. Also, I have job family and job description, they are often showing similar descriptions in one job family. I think these features may also causing some dependency problem. |
| Is this a **regression**, or **classification** (**binary** or **multiclass**)? Explain. |
| This is a classification because I want to have the result in three brackets – high risk, medium risk and low risk. |
| How might you split the observations into **train** and **test** sets? Are there potential **biases** to look out for? |
| I want to split in 70-10-20. This data has 1.7M rows so I think I can train on 70%, validate on 10% and test on the 20% data. |
| What type of **loss** function might be appropriate for quantifying the error of your algorithm? |
| I think I will use zero-one loss because I want to feed the model I choose to train the model and the output will be one of three categories. |

# Project Part Two: Application and Limitations of k-NN

**Instructions:** Answer the following questions about the k-Nearest Neighbors algorithm. Please limit your answers to fewer than 100 words.

|  |
| --- |
| You have implemented k-NN to learn a decision boundary from a training data set that you know to be noisy. To address the issue of noise, you increase this number of nearest neighbors to nine, which has seemingly reduced the impact of the noise on your decision boundary. However, you notice that significant groups of data points in your training data set are now misclassified by this new decision boundary. **Why** might this happen? **How** can you adjust your algorithm **to** **improve** its **accuracy**? |
| When K increases to a larger number, that means it is looking for more neighbors close to each data point. This increases the range that similar data is selecting. As mentioned in the course, the little ‘island’ gets wash out, the boarder that distinguish different data points will become more smoother, which may include the wrong data points. Also, it may happen because the distance between different data points are the same. In this situation, we can change K to even number, then it may increase the accuracy. |
| When implementing k-NN, you must choose an appropriate distance function. **What** is the role of the distance function and **why** is it important to the accuracy of your machine learning algorithm? |
| A distance is important for the KNN accuracy and the classifier because it will influence the neighbor in the model. Different distance calculation may generate different result for the model. Therefore, choosing the right distance function is essential. |
| **Why** does increasing the number of observations also increase the computing **power** and **time** necessary to run k-NN? **What** possible solution is available to help improve the efficiency of k-NN in such a scenario? |
| Because KNN has to look for each data point and also its neighbor, if the number of observations is big, it can take so much time to run through every observation and is cost a lot of computing power and time to complete it. The speed could increase by using data structure like k-d trees or ball trees. For example, k-d tree will separate data points to different boxes, then it will only look for the distance within the box. Then, it can save time to compare everything in the who dataset. |
| **Describe** the curse of dimensionality. **Why** does k-NN break down in high-dimensional space? |
| When the curse increases, it may not able to find the exact the same result for the model, which can lead prediction become failure. Take human being as an example, even a twins has small different by looking their features. Twins may have the same gender, blood type and eye color. However, Twins may have different height and weight. So, if the input only has 3 dimensional such as gender, blood type and eye color, the model may show this Twins as result. But, if the input increases to 5 dimensional, it may not have result generated because they have different weight and height. That’s why KNN breaks down in high-dimensional space. |