ML Algorithm from Scratch

1. **Output in C++**

- **For program1:**

```
Open file titanic_project.csv
Closing file titanic_project.csv

Elapsed time for training algorithm: 500.758s

Coefficients:
w0: 0.999877
w1: -2.41086

Accuracy: 0.785425
Sensitivity: 0.863636
Specificity: 0.695652
```

- **For program2:**

```
Open file titanic_project.csv
Closing file titanic_project.csv

Prior probability:
Survived=no: 0.61
Survived=no: 0.39

Likehood values for p(pClass|survived)
0.159836 0.840164 0.602459
0.679487 0.320513 0.320513

Likehood values for p(sex|survived)
0.159836 0.840164
0.679487 0.320513
ageMean:
30.4182 28.8261
ageVar:
205.153 209.155
```

2. **Analyze the result of algorithms**

- The result is very similar to what we have in R, there is only slightly different because of the rounding issue form C++. More interation, the closer we have between C++ result and R result. However, the code in C++ run significantly slower than in R.

3. **Compare and contrast Generative Classifer vs Discriminative Classifer**
   - **Generative Classifer:** such as Naïve Bayes usually outperform the discriminative classifer such as logistic regression. This algorithm directly estimates the parameters of for P(Y) and P(X|Y). It is also work better with small data sets. The Naïve Bayes algorithm usually have high bias and low variance compared to discriminative such as logistic regression. However, if the training set keep growing to infinitive and the independence assumption hold, both Naïve Bayes and logistic regressions will get the same result as they converge toward each other.
   - **Discriminative Classifer:** such as logistic regressions estimates the paremeter of P(Y/X) directly. When the size of training data grow, the logistic regression perform also increase more compare to Naïve Bayes. When the number of factor and level on data increase, the logistic regression tend be overwhelmed while the Naïve Bayes consider this as a favor.

     Note: All information above was collected from ML Handbook by Dr. Karen Mazidi. (Section 7.9.5)

4. **Reproducible Research**
   - **Reproducible Research:** mean which the same data, software and code. Anyone can reproduce same results such as: tables, graphs, reports, etc by following a list of instruction on the research. The reproducible ablity is important because it is usually the research was perform correctly, it can be easily reference or verify. It also mean it can be extended easier because the code/instruction is clearly written, it can increase the collaboration efficency. The reproducibility can be implemented when when the research was conducted on transparency, the instruction/code is easy to understand and follow. A set of requirements was clearly stated.

     Note: Sources:

     1. Bock, Tim. "What Is Reproducible Research?" *Displayr*, 7 Dec. 2020, https://www.displayr.com/what-is-reproducible-research/.
     2. "Research Reproducibility: About." *Subject and Course Guides*, https://researchguides.uic.edu/reproducibility.
     3. "Research Reproducibility: About." *Subject and Course Guides*, https://researchguides.uic.edu/reproducibility.