Final Reflection

DA 5030

Dauby, Ray

2024-12-11

Introduction:

Over the course of this semester, I have learned astronomical amounts about machine learning and data mining, and improved my skills in R, while preparing myself for work in industry. I spent countless hours reading the textbook, watching lectures, applying my knowledge to practice problems and weekly coding assignment, and brought it all together for my final project. This course was excatly as challenging as I had hoped it would be, and I am thrilled to have had the chance to engage with it in my last semester of grad school.

1. Personal Learning Journey

Initial Understanding:

When I started this course, I was almost completely unfamiliar with machine learning concepts. I had taken stats for psychology during my undergrad and was familiar with evaluation metrics like mean deviation, MSE, residuals, and accuracy, but almost all of the techniques, models, and algorithms I worked on this semester were new to me. I had some experience at my last co-op doing various kinds of regression with dose-response data, but this course added a whole new level of understanding to that data analysis and now I wish I could go back and update my apps that I made for them!

Key Learnings:

Having had that previous experience with linear, logistic, and polynomial regression, I was thrilled to be engaging with similar models in our course. The units that I was most interested in were decision trees and multiple regression, because their interpretability/visibility/development made sense to me, which made them highly engaging. I enjoyed learning about xgboost/boosting models during my final project, and feel that the practicums were very effective at solidifying my understanding of the modules.

Growth and Development:

I feel that my general understanding of machine learning concepts and techniques has grown immensely over the past four months, primarily due to the enthusiasm and commitment to learning I maintained throughout the semester. Given that I started at a solid zero understanding of machine learning, I feel that my growth has been exponential across the board. I would say that model evaluation and comparison has been the most intriguing aspect of my learnings so far, because I loved learning about how different algorithms perform differently on similar datasets and how we can optimize our model usage for improved predictive capacity.

2. Course Content and Key Concepts

Core Algorithms and Techniques:

KNN was one of the first models we learned about in this course and for that reason it was one of the most memorable for me. It helped that this model is easily visualized, so the process of building and evaluating it felt very accessible to me. I applied it first in the KNN weekly assignment, where we processed and analyzed a prostate cancer dataset to predict a patient diagnosis, and later applied it in the Practicum 1 assignment to predict the development status of a country based on life expectancy data. KNN only had the K hyperparameter to tune, which is typically best at the sqrt(n cases), so I felt that this model was a great starting point for my ML journey.

Naive Bayes was also one of the first few models we examined in this course, but was a bit less visually understandable than KNN. That being said, the underlying mathematics of the model were accessible and as a model for binary classification using conditionally independent features, it made sense in the way that it was taught. I applied Naive Bayes in the weekly assignment, where it was used to predict the party of a member of the House of Representatives based on how they voted on certain measures using an ensemble model. This was a super engaging application and one I would not have thought of on my own!

SVM was one of the later models we learned in the course, but again I was engaged by it because it was visually understandable. I employed this model in my final project and used it to make a binary prediction on my heart failure data, to examine whether or not a patient would survive the follow-up period. SVM has two main hyperparameters, C and sigma, which define the cost parameter and the point influence. I enjoyed getting to work with this model because it felt like a more complex application of KNN in some ways, and comparing the different kernels gave me a better perspective on the dimensionality and divisions of the dataset.

Practical Applications:

The most real-world applicable models I built for this course were during the final project, where my dataset was relevant to healthcare outcomes and patient survival. Theoretically, this understanding of the linkage between certain clinical factors could help healthcare providers make informed decisions about patient care and resource allocation. This feels super relevant in today's world where we have an abundance of healthcare data as a significant untapped resource for predicting individual outcomes. I used a pretty clean dataset, so the data preprocessing steps such as missing data and encoding were not necessary steps. Model selection was also pretty intuitive because the dataset was entirely numeric and I had a binary prediction task, which indicates models like logistic regression, random forest, and SVM (the three models I used). Hyperparameter tuning proved to be more complicated than I had previously imagined due to the computational intensity of broad grid searches, which slowed my processing time significantly, and also did not improve model performance by much at all. I would like to learn more about each of the hyperparameters used in all of the models we studied and how much optimizing them truly affects the model performance.

Model Evaluation and Bias:

One of the things I came into this course hoping to understand more was bias in machine learning, and I was interested in this topic as an opportunity to delve into the ethics of ML models and their predictions. When we were using test sets like 'concrete' and 'iris,' I had a hard time understanding how bias would affect prediction outcomes, but when it came to healthcare and demographics, it became increasingly evident that the ethics of the data scientist is incredibly relevant to model development. Discussion of proxy variables and the omission/inclusion of demographic variables enlightened me to the importance of ethical data preprocessing. Apart from ethics, I also learned about the importance of performance and evaluation metrics such as cross-validation and confusion matrices to help me compare and assess model performance and reliability. These valuation steps helped me take a step back from the thick of the code and actually examine what I was doing, which was crucial for my understanding of the topics.

3. Challenges and Problem-Solving Strategies

Key Challenges:

The most challenging aspect of this course for me was getting into the groove of how to best absorb and apply the course content. The first few weeks were hard for me because I had yet to figure out that I learned best from the textbook and not as well from lecture videos. After making this discovery, I spent more time with the textbook before watching lecture videos, which significantly improved my understanding of the unit topics before I got to the weekly assignment. The theory of each algorithm actually came easier to me than I was expecting, because I am a very visual learner and most of the techniques could be explained by graphical models. By the end of the course, my weekly time management was much improved and I was able to spend more energy working on my final project.

Critical Thinking and Problem-Solving:

One of the problems I faced in my final project was with my ensemble models, I had used the original xgboost package for boosting my logistic and svm models, but when it came time to create an ensemble model, the caret package required that I use xgboost from caret, which then caused me to go back and rework the boostong sections for those two models. I had modified the data to fit the original xgboost requirements (matrix format and encoded target variable), but I had to undo that and deal with the repercussions of the un-encoding later on in the code. To fix all of these issues and interpret the errors I was encountering, I did lots of further research and employed data science discussion forums to solve my problems. I tend to get frustrated easily with this kind of issue, especially when I am on a time crunch, so I approached this problem by working on it for an hour or so at a time and then working on something else to hold off the frustration.

4. Collaboration and Communication

Group Work and Peer Learning:

I chose not to work with others on any of the assignments because I have had a lot of negative group work experiences in the past. That being said, all of the discussion post comments and responses were very helpful and supportive, so perhaps I was wrong to fly solo. I did have some friends who had taken the course in past years with whom I discussed my final project and the course content, which was affirming because they agreed that the course was difficult and time-consuming, but well worth my effort! I grew the most in my verbal descriptions of machine learning tasks during the final project when I intentionally broke down each step into interpretable chunks, which helped me identify where my knowledge was lacking and help me answer my own questions for the benefit of my listener.

Communicating Results:

When I started this course, I had used ggplot before but had never created so many R notebooks (I had focused on R Shiny apps), but there was actually a lot of overlap between their stylizing capabilities and I was able to apply what I knew and build on my knowledge of R in the notebook format. I also developed my communication skills and learned the verbiage of data scientists when they talk about machine learning and identified new ways to communicate data through graphs (heatmaps, box plots) and tables (kable).

5. Future Applications and Goals

Real-World Applications:

I plan to apply what I've learned in this course to my degree in bioinformatics as a tool to further research on genetic therapeutics and diseases. At my last two co-ops, I worked parallel to data science teams that were using machine learning to further therapeutic development of oligonucleotide-based therapies, and wanted to get engaged with their research, but had no foundation in machine learning to jump from. This interest was the impetus for my enrollment in DA 5030, and I hope that what I have learned will take me further in my field.

Continuous Learning:

Similar to many industry professionals, I started learning computer science using courses online, and have found this to be an excellent source of knowledge and skill. As well as honing my machine learning skills using tutorials and projects, I also intend on staying up to date with published articles regarding machine learning and data mining in the gene therapy area, as well as staying in contact with other academics and researchers in the data science field.

Personal Goals:

I am graduating from Northeastern this semester, which means that I am hoping to get a job in industry or academia relating to my degree in bioinformatics. I think that this experience in machine learning makes me a more versatile and marketable candidate, and I hope that pursuing a career in data science will prove an interesting and motivational path. One of the things I love most about working in biotech and scientific research is that it is highly collaborative and relies on constant learning and developing of theory and skills, which I hope will allow me to have a life of learning.

Conclusion:

I will be taking all of the skills and knowledge that I have acquired during this course out into the world with me, and I am thrilled that this was how I chose to spend my last semester at school. I appreciate all of the energy and discussion that went into the creation and execution of this course, and can't wait to keep building on the framework that was set up so wonderfully for me by this class.