Course Plane:
Machine Learning in Conflict prediction

Fall 2020

Contents

1 Introduction

This course plan is meant to serve as your guide throughout the course and as an outline of my expectations to you. However, it might be subject to slight change.

1.1 About the course

The goal of this course is to give you the skills and knowledge necessary to create actual conflict forecasting using machine learning and modern data sources. Such forecasts will include information regarding probabilities of conflict at sub national level, assessment of uncertainty regarding the results and evaluation of model performance compared to various baseline models. You will also learn to presented this information in a intuitive and honest manner.

Notably this course is centred around machine learning in conflict prediction and *not* machine learning in general. Thus, we will only look at some very specific approaches and algorithms. There are tons more to look at, but we simply do not have time to cover everything. Indeed, this course should be seen as an "advanced introductory" course. It is advanced because you need a lot of prerequisite knowledge and/or skills to do machine learning in conflict prediction. However, it is also introductory in the sense that we only have time to go through the fundamentals. Hopefully, your future studies will allow you to dive deeper into the subject.

Naturally, if you want to go into more advanced stuff (e.g. Gaussian Processes and/or Artificial Neural Networks) please let me know and I will provide you with some resources.

1.2 How to approach the curriculum

You all come from different backgrounds: some know a lot about coding, some know a lot about conflict, some know a bit of both and some are quite new to all of it. As such, you should prioritise your weakness when going through the curriculum. If you have not programmed in python before focus extra on VanderPlas and Friedman. If you are used to python then pay less attention to VanderPlas and if you know your machine learning then pay less attention to Friedman. If you know your python and machine learning focus on the articles. You get the point.

Given the goal presented above, the core of the course is the practical exercises. Admittedly, some of you might not have time to read the whole curriculum thoroughly while also solving the exercises (assuming you have other courses and meaningful lives). Anticipating this, I will point to at least one "must-read" text for each week which you should at least familiarise yourselves with. The broader curriculum should be treated as a collection of useful references for this course as well as your future endeavours.

To people not overly familiar with Python I highly recommend going through VanderPlas (2016) chapter 1-5 as soon as possible. It will make a substantial difference regarding your understanding

of the code in the assignments and probably also your learning outcome.

1.3 Exercises

There will be 5 exercises, each divided into two parts. E.g. 1A, 1B, 2A, 2B, and each part will be the subject of a separate week. The deadline for submitting the exercises to feedback is every second week (Monday, 23.59). Here you are expected to hand in both part A and B of a given exercise. E.g 1A and 1B (see plan below).

There will be no (official) group work. Everybody will be coding. You will evaluate each others work (two-way blind) through Peergrade. This will ensure you'll get a lot of direct feedback but also that you get exposed to other peoples solutions and code. Thus, if you did not get it right the first time, you will be able to revisit your code and correct it. If you do this throughout the course, the exam will be a breeze.

Weeks with no exercises will give you the chance to focus on incorporating feedback (if you have not done so yet), optimise code, create better solutions, catch up if you fallen behind and of course work on the portfolio exam (See below). Doing the lectures these weeks, I will follow up on past exercise, comment on common mistakes and have discord open for help, guidance and whatever question you might have regarding past exercises or the exams.

1.4 Exams (Updated with regards to re-submission)

This course uses **portfolio exams**. This entails that there will be two exams to hand in. The two exams will mirror the exercises pretty closely. They will include both coding tasks, questions regarding understanding, interpretation, reflection, pros and cons etc. The main difference between the exams and the weekly exercises is that the exams will be longer and that I will supply no code for the exams. However, fear not, if you have simply finished your exercises and incorporated (and maybe even reflected upon) the relevant feedback you will have everything you need to get a passing – even potentially good – grade.

- The first exam will be public in week 44. **Deadline: Wednesday 9. December at 23.59**
- The second exam will be public in week 51. **Deadline: Wednesday 6 of January at 23.59**

The following concerns re-submission of the exams in the case of a student not passing. It comes directly from the "studieordning", and as such it is not something I am able to change: The exams should be handed in on Absalon and both exams must be passed in order to pass the course. There will be given one grade based on the combined quality of both exam assignments. Students who does not pass an assignment at first will have **one** more chance to hand in a revised assignment at a later (yet unspecified) point. **Note that in order to hand in a second time, you must have handed in something at the first deadline – blanking is not a possibility.**

- Re-submission for the first exam will be possible from week 51. **Deadline: Wednesday 6 of January at 23.59**
- Re-submission for the second exam will be public in week 1. **Deadline: Wednesday 15 of January at 23.59**

1.5 Note on maths and Friedman et al. 2001

There are a lot of maths in Friedman et al. 2001. For some of you this will be fine. For some of you this will be inconvenient, and for some of you it'll bring back substantial traumas. However, I implore you all to read through it despite whatever disposition you might have towards maths.

The book is perfectly readable without understanding all (or even most) of the maths. And at some point the maths might even start to make sense. Getting comfortable reading technical text will benefit you hugely down the line – even if the maths still doesn't make a 100% sense.

If you want to brush up you maths skills (now or in the future) here are three book titles to get you started:

- **Easy**: Moore, Will H., and David A. Siegel. A mathematics course for political and social research. Princeton University Press, 2013.
- **Less easy**: Gill, Jeff. Essential mathematics for political and social research. Cambridge: Cambridge University Press, 2006
- **Still manageable**: Deisenroth, Marc Peter, A. Aldo Faisal, and Cheng Soon Ong. Mathematics for machine learning. Cambridge University Press, 2020.

If nothing else, Moore and Siegel have a nice introduction to mathematical notation and a list of what differetn symbols (usually) mean.

2 The Plan

2.1 Week 36: Getting started and getting the data

Agenda:

- How the course is structured
- Getting the data

Exercise: 1A. Deadline Monday 14.09.2020

Must reads: Everything on the curriculum for this first lecture is pretty central. Thats why it's on the curriculum for the first lecture. That being said, if you have not done so yet I highly recommend

that you read: **Hegre, Håvard, et al. "ViEWS: a political violence early-warning system." Journal of peace research 56.2 (2019)**.

Readings:

- Hegre, Håvard, et al. "Introduction: Forecasting in peace research." (2017): 113-124.
- Sundberg, R. and Melander, E. (2013). Introducing the ucdp georeferenced event dataset. Journal of Peace Research, 50(4):523 - 532.
- Croicu, M. and Sundberg, R. (2017). Ucdp ged codebook version 18.1. Department of Peace and Conflict Research, Uppsala University
- VanderPlas, Jake. Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.", 2016. Capter 1-4

2.2 Week 37: Filling in the gaps – Interpolation of missing data

Agenda:

- Getting PRIO dynamic data merged with the rest of the data.
- Seeing some examples of predictions to get a grasp of the field and for inspirations (see curriculum).

Exercise: 1B. Deadline Monday 14.09.2020

Must reads: The curriculum for this lecture does not have a direct link to the exercises (1B). Instead it is meant to give you and idea of regarding state-of-the-art (SOTA) in the field. As such, if you only read one thing for this lecture it should be **Hegre, Håvard, et al. "ViEWS: a political violence early-warning system." Journal of peace research 56.2 (2019): 155-174**.

Readings:

- Weidmann, N. B. and Ward, M. D. (2010). Predicting conflict in space and time. Journal of Conflict Resolution, 54(6):883 - 901
- Hegre, H., Nygaard, H. M., Karlsen, J., Strand, H., and Urdal, H. (2013). Predicting Armed Conflict, 2010-2050. International Studies Quarterly, 57(2):250 - 270.
- Perry, C. (2013). Machine learning and conflict prediction: a use case. Stability: International Journal of Security & Development, 56(2(3)). (18 pages)
- Hegre, Håvard, et al. "ViEWS: a political violence early-warning system." Journal of peace research 56.2 (2019): 155-174.

2.3 Week 38: Probing prediction: out-of sample predictions and evaluation metrics

Agenda:

- Introducing the concept of out of sample prediction
- Create train and test
- Generate some baseline models to test

Exercise: 2A Deadline Monday 28.09.2020

Must reads: The curriculum for this lecture goes a bit more into the reason we use predictions to evaluate our models. If you only read one, let it be **Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York, NY, USA. Chapter 7**'. Don't worry if you do not get everything, just try to understand the intuition. Alternatively see **Ward, M. D., Greenhill, B. D., and Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. Journal of Peace Research, 47(4):363 - 375** for a more 'soft' introduction.

Readings:

- Colaresi, Michael, and Zuhaib Mahmood. "Do the robot: Lessons from machine learning to improve conflict forecasting." Journal of Peace Research 54.2 (2017): 193-214.
- Ward, M. D., Greenhill, B. D., and Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. Journal of Peace Research, 47(4):363 - 375.
- Schrodt, P. A. (2014). Seven deadly sins of contemporary quantitative political analysis. Journal of Peace Research, 51(2):287-300.
- Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York, NY, USA. Chapter 1, 2 and 7

2.4 Week 39: Few and far between – Imbalanced data and evaluation metrics

Agenda:

- Evaluating (imbalanced) data

Exercise: 2B Deadline Monday 28.09.2020

Must reads: If you only read one paper for this lecture it should be **He, H. and Garcia, E. A. (2008). Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, 9:1263 - 1284**.

Readings:

- He, H. and Garcia, E. A. (2008). Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, 9:1263 - 1284.
- King, G. and Zeng, L. (2001a). Explaining rare events in international relations. International Organization, 55(3):693 - 715
- King, G. and Zeng, L. (2001b). Improving forecasts of state failure. World Politics, 53(4):623 - 658.
- Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G.,Ulfelder, J., and Woodward, M. (2010). A global model for forecasting political instability. American Journal of Political Science, 54(1):190 - 208.

2.5 Week 40: Drawing the line – Machine Learning Algorithms 1

Agenda:

- Linear Regression (regression)
- Logistic Regression (Classification)
- Decision Trees (regression and Classification)
- (Optional Naive Bayes)
- (Optional SVM)

Exercise: 3A Deadline Monday 19.10.2020

Must reads: If you only read one chapter for this lecture it should be chapter 9 (since I assume you are all familiar with Linear and Logistic regression). Understanding Decision Trees will be central for understanding Random Forest and XGboost later on.

Readings:

- Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York, NY, USA. Chapter 3, 4, 9 (not on curriculum but 6.6.3 for Naive Bayes and 12 for SVM)

2.6 Week 41: Stronger together – Ensemble model

Agenda:

- Introducing the concept of ensembles
- Tying together with imbalanced data and undersampling
- Bayesian correction
- Quantifying (kinds of) Uncertainty

Exercise 3B Deadline Monday 19.10.2020

Must reads: To be honest you should read everything here. I think both articles are very worthwhile, but the concept of ensembles is better presented in the book. Maybe skim the book if your are in a hurry, but do try to read the articles.

Readings:

- Hegre, Håvard, Håvard Mokleiv Nygård, and Ranveig Flaten Ræder. "Evaluating the scope and intensity of the conflict trap: A dynamic simulation approach." Journal of Peace Research 54.2 (2017): 243-26
- Ward, Michael D., and Andreas Beger. "Lessons from near real-time forecasting of irregular leadership changes." Journal of Peace Research 54.2 (2017): 141-156.
- Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York, NY, USA. Chapter 16

2.7 Week 42: FALL BREAK

2.8 Week 43: Thinking about theory – Feature engineering

Agenda:

- Connection to theory
- Feature engineering

Exercise: 4A Deadline Monday 02.11.2020

Must reads: Well, the only 'real' text on the curriculum for this lecture is simply an example of a proxy feature. I will upload "a more general note on actual feature engineering". Read that.

Readings:

- Weidmann, Nils B., and Sebastian Schutte. "Using night light emissions for the prediction of local wealth." Journal of Peace Research 54.2 (2017): 125-140.
- a more general note on feature engineering

2.9 Week 44: Thinning the herd: Feature importance and selection

Agenda: - Feature importance - Feature selection

Exercise: 4B Deadline Monday 02.11.2020

Must reads: That text that you know you should have read, but you haven't come around to yet – yes, that text. You know the text i'm talking about... You should read that text.

Readings:

- That text we both know you haven't read yet..

2.10 Week 45: A forest full of tress – Machine Learning algorithms 2

Agenda:

- Random Forest
- General questions regarding curriculum, exercises and exam assignments

Exercise: None

Must reads: That will be **Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York, NY, USA. Chapter 15**. Understanding Random Forest will be central for understanding XGboost later on - and it is a canonical and seminal algorithm.

Readings:

- Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York, NY, USA. Chapter 15

2.11 Week 46: Boosting

Agenda:

- Presenting boosting/Adaboost
- General questions regarding curriculum, exercises and exam assignments

Exercise: None

Must reads: Read the one chapter presented here **Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York, NY, USA. Chapter 10**. It will be central for understanding XGboost later on.

Readings:

- Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning. Springer series in statistics New York, NY, USA. Chapter 10

2.12 Week 47: Connecting the dots – Machine Learning algorithms 3

Agenda:

- introducing the XGboost algorithm

Exercise: 6A Deadline Monday 07.12.2020

Must reads: This test is a must read: **Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785-794. ACM**. Perhaps one of the best and most robust machine learning algorithms for structured data and general problems out there at the moment. This should be one of your go-to models.

Readings:

- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785-794. ACM.

2.13 Week 48: Tying the knots – more XGboost

Agenda:

- More XGboost
- The problem with models such as Xgboost and Random Forest
- grid search

Exercise: 6B Deadline Monday 07.12.2020

2.14 Week 49: Where to go from here – better models and novel data

Agenda:

- Probabilistic generative models as the ideal
- Other algorithms: Gaussian Processes and Neural Networks
- Other data: text and images as data

Must reads: Read at least **Mueller, H. F. and Rauh, C. (2016). Reading between the lines: Prediction of political violence using newspaper text. American Political Science Review, 2(112):358-375**. Its is a nice text and the idea has huge potential. The methods, however, are already a bit outdated but that is the name-of-the game in the field of machine learning right now. Everything moves really fast.

Grimmer and Stewart is not actual on the curriculum and it is also a bit dated by now (since most people use Neaural Networks; RNNs, LSTMs etc. for text operations now). However it is still a very good – and seminal – read.

Exercise None. Well, fix you old code! Implement feedback; steal that cool code you evaluated; do all the neat things you should have done from the beginning but didn't know about. You will thank yourself at the exam.

Readings:

- Chadefaux, T. (2014). Early warning signals for war in the news. Journal of Peace Research, 51(1):5 - 18.
- Mueller, H. F. and Rauh, C. (2016). Reading between the lines: Prediction of political violence using newspaper text. American Political Science Review, 2(112):358-375.
- Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." Political analysis 21.3 (2013): 267-297. (Not strictly on the curriculum)

2.15 Week 50: Predictions about predictions – summing up the course

Agenda:

- Potentials and pitfalls of predictions
- General questions

Exercise: None. Though, if you have not finished VanderPlas chapter 1-5 go do that. Really.

Must reads: Two very short texts. You can manage to read both.

Readings:

- Cederman, L.-E. and Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations?
- Chadefaux, T. (2017). Conflict forecasting and its limits. Data Science, 1(1-2):7 - 17.