FLS 6497 | DCP/USP

Ciência de Dados e Aprendizado de Máquina em Ciência Política

Fernando Meireles

Prédio Fil. e C. Sociais, sala 122
Quartas-feiras, das 14 às 18h
meirelesff@hotmail.com
fmeireles.com/materiais

Apresentação

Este curso introduz um conjunto de ferramentas que nos permite usar dados de diferentes formatos para responder questões substantivas sobre política. Com ênfase em aprendizado de máquina – i.e., modelos que aprendem a fazer generalizações a partir do reconhecimento de padrões em amostras –, seu objetivo principal é capacitar alunos(as) a aplicar noções de Ciência de Dados e de programação em problemas concretos de classificação, predição e descoberta, o que lhes permitirá construir aplicações como classificadores de texto e de imagem, detectores de *outliers* ou modelos flexíveis de MrP.

A abordagem do curso será principalmente prática. Na maior parte do tempo, estudaremos tópicos por meio da resolução de exercícios, dentro e fora de sala de aula. De início, após cobrirmos noções úteis de programação de revisarmos a aplicação de modelos de regressão, estudaremos os diferentes tipos de problemas em Ciência de Dados; tipos de aprendizagem e seus principais algoritmos; estratégias de validação e de *tuning*; e, finalmente, realizaremos projetos que servirão para testar conhecimentos adquiridos. Concluído este percurso, a expectativa é que alunos e alunas obtenham a experiência necessária para incorporar *skills* de Ciência de Dados em suas rotinas de pesquisa ou de trabalho.

Objetivos

São estes os principais objetivos de ensino do curso:

1) *Desenvolver habilidades de programação*. Embora este não seja um curso que ensinará programação diretamente – a como uma forma de aplicar aprendizado de máquina –, alunos e alunas terão a oportunidade de praticar a escrita de código para resolver problemas de pesquisa.

2) *Aprender a conduzir projetos básicos de Ciência de Dados de ponta a ponta*. Entre outros, alunos e alunas apreenderão a estruturar perguntas em Ciência de Dados, organizar dados necessários e definir estratégias para respondê-las – o que incluirá criar *pipelines*, estabelecer métricas de avaliação, validar e ajustar modelos e algoritmos, entre outros.

3) *Estimular o trabalho colaborativo em pesquisa científica*. Por conta da dinâmica do curso, que envolverá trabalhos em duplas e desenvolvimento de *papers*, alunos e alunas serão desafiados a identificar produções recentes na literatura internacional; e a redigir textos que os(as) ajudem a preparar teses, dissertações ou artigos para publicação.

Pré-requisitos

O curso pressupõe conhecimentos de estatística, modelos de regressão e análise de dados. Formalmente, o pré-requisito é já ter cursado a disciplina **FLS 6183 Métodos Quantitativos II**.

Também é esperado que alunos(as) tenham familiaridade com **R** ou **Python**. Como escolher entre as duas linguagens? Se você já trabalha com **R** e seus interesses são acadêmicos, seguir com essa escolha é o melhor; por ser mais demandado no mercado e ser mais usado em áreas conexas, como a engenharia de dados, **Python** pode ser interessante para quem deseja se qualificar profissionalmente, mas é necessário já ter um nível de programação para além do básico para conseguir acompanhar o curso. Em qualquer caso, recursos didáticos serão disponibilizados em ambas as linguagens, ainda que a minha capacidade de fornecer ajuda seja consideravelmente maior em **R**.

Leituras

Embora tenhamos poucas leituras analíticas, manuais serão usados para cobrir a implementação de modelos e estudo de conceitos. São eles:

- [**ITSL**] Introduction to statistical learning
- [**HML**] Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow
- [**MLRTM**] Machine Learning with R, the tidyverse, and mlr
- [**MLR3**] MLR3 Book

Logística

Nossas aulas terão dois blocos. No inicial e menor deles, teremos uma exposição dos temas abordados e discussão dos trabalhos de leitura indicada. Encerrada esta parte, teremos sessões nas quais alunas e alunos aplicarão conhecimentos vistos. Para facilitar essa parte, usaremos *pair programming*, prática que consiste no trabalho em duplas no qual há uma parte que principalmente escreve código e, a outra, o revisa e comenta em tempo real.

Avaliação

O aproveitamento no curso de cada estudante será avaliado de três formas: exercícios, que serão realizados dentro e fora de sala; dois pequenos projetos; e um *working paper* final. A nota final no curso será dada pela soma aritmética das notas de cada tarefa avaliativa.

Exercícios (*15%*)

Em cada aula teremos uma lista de exercícios para praticar o conteúdo visto. Estes serão apresentados em sala e deverão ser realizados em duplas seguindo o sistema de *pair programming* para serem entregues até o encontro seguinte. A entrega pontual e regular dos exercícios e o esforço aplicado para resolvê-los serão os critérios de avaliação.

Projetos (*35%*)

Também teremos dois pequenos projetos que deverão ser entregues individualmente no formato de *notebooks* (feitos com Rmarkdown ou Jupyter). A ideia é que ambos os desafios não apenas testem conhecimentos, mas, também, ofereçam ideias de uso de aprendizado de máquina para pesquisas em Ciência Política. Nestes, a avaliação levará em conta a capacidade de aplicar o conhecimento visto ao longo do curso e a capacidade de cumprir os objetivos propostos.

Projeto 1 – Modelo *data-driven* de MrP (*15%*) No primeiro projeto, o objetivo será treinar um modelo a partir de dados de *survey* para predizer votos em candidaturas às eleições presidenciais e, com ele, projetar as estimativas de votos em dados da PNADc. O produto final deverá ser um modelo de MrP que use aprendizado de máquina.

Projeto 2 – Classificador de imagens de satélite (*20%*) No segundo projeto, que corresponderá a *20%* da nota final do curso, extraíremos imagens de satélite de locais de votação georreferenciados no Brasil para, usando redes neurais convolucionais, os classificarmos em determinadas categorias. O resultado final deverá ser uma *pipeline* que permita gerar diferentes esquemas de classificações de imagens de satélite de locais de votação (ou de outras localidades georreferenciadas) no Brasil.

Working paper (*50%*)

Finalmente, os(as) alunos(as) deverão entregar um *working paper* a título de avaliação final. Este deverá aplicar algum dos métodos que veremos no curso e ter entre *10 e 15 páginas*. Idealmente, será possível aproveitar essa oportunidade para rascunhar um capítulo de tese ou dissertação, ou um artigo para publicação futura. Para estimular o trabalho colaborativo, serão aceitos trabalhos finais realizados em dupla. Criatividade, aplicação correta de noções vistas no curso e estrutura dos textos (i.e., boas motivações, seções adequadas de metodologia e de resultados) serão avaliados.

Na última aula, alunos e alunas apresentarão suas ideias e resultados parciais para obterem *feedback* e tirar dúvidas. A datas de entrega da avaliação será combinada no decorrer do curso.

Política de Gênero

Em aulas de metodologia, homens frequentemente monopolizam a participação. Para evitar isso, seguiremos três protocolos neste curso: no uso de computadores nas atividades de *pair programming*, mulheres serão priorizadas; para intervir, é necessário estender a mão; quando mulheres falam, colegas não as interrompem.

Atendimento a Necessidades Especiais

Alunas(os) com quaisquer necessidades ou solicitações individuais não devem exitar em procurar auxílio, tanto por e-mail quanto pessoalmente.

Ferramentas

Para resolver tarefas e praticar em casa, certifique-se de ter as ferramentas que usaremos devidamente instaladas em seu computador ou *notebook*. Para quem usará **R**, isso inclui tê-lo instalado e, também, o Rstudio. É possível encontrar tutoriais na internet cobrindo os passos necessários. Já para quem pretende usar **Python**, minha recomendação é usar Python 3 e alguma IDE como VScode ou spyder para escrever e gerenciar scripts e repositórios.

Para além destes *softwares*, será necessário instalar algumas das *libraries*. Em **Python**, usaremos principalmente a biblioteca Scikit-learn, que oferece um conjunto de ferramentas de pré-processamento, construção de *pipelines*, seleção e validação de modelos, para além uma ampla gama de algoritmos supervisionados e não-supervisionados; e a biblioteca Keras, que é uma suíte para a construção de modelos de *deep learning* via Tensorflow. A depender do seu sistema operacional e da disponibilidade de pré-requisitos em seu computador, ambas as bibliotecas podem retornar erros durante a instalação, caso no qual eu posso tentar ajudar em sala ou por e-mail. Também é recomendado utilizar um ambiente virtual antes de fazer qualquer coisa (aqui uma explicação).

Para instalar os pacotes que precisaremos, basta executar do terminal:

```
pip install --upgrade pip
pip install tensorflow scikit
```

Em **R**, o equivalente mais próximo do Scikit-learn é a biblioteca mlr3, que também oferece um conjunto de ferramentas e adota princípios de programação orientada a objetos (discutiremos isso em aula). Keras e Tensorflow, por sua vez, já têm versões em **R**. Para instalar todos os pacotes que usaremos, basta rodar o seguinte código no **R**:

```
install.packages(c("mrl3", "tensorflow", "keras"))
```

Isso feito, é preciso instalar o `Tensorflow` com:

```
library(tensorflow)
install_tensorflow()
```

Plano das Aulas

Aula 1 – Apresentação do curso

Leituras sugeridas:

– M. J. Salganik., *Bit by bit: Social research in the digital age*., Princeton University Press, 2019., https://www.bitbybitbook.com/en/1st-ed/.

Aula 2 – Introdução à Ciência de Dados & Revisão de Programação

Leituras:

– ITSL, Cap. 2.1 & 2.2

– HML, Cap. 1

Leituras sugeridas:

– S. Athey and G. W. Imbens., "Machine learning methods that economists should know about"., In: *Annual Review of Economics* 11 (2019), pp. 685–725.

– J. Grimmer, M. E. Roberts, and B. M. Stewart., "Machine learning for social science: An agnostic approach"., In: *Annual Review of Political Science* 24 (2021), pp. 395–419.

Aula 3 – Aprendizado Supervisionado: Regressão

Leituras:

– ITSL, Cap. 3

– HML, Cap. 2 (para **Python**)

– MLR3, Cap. 2 (para **R**)

Leituras sugeridas:

– Y. Li and M. S. Shugart., "The seat product model of the effective number of parties: A case for applied political science"., In: *Electoral Studies* 41 (2016), pp. 23–34.

– R. S. Erikson and C. Wlezien., "Forecasting the 2020 presidential election: Leading economic indicators, polls, and the vote"., In: *PS: Political Science & Politics* 54.1 (2021), pp. 55–58.

Aula 4 – Aprendizado Supervisionado: Classificação

Leituras:

– ITSL, Cap. 4

Leituras sugeridas:

– S. Streeter., "Lethal force in black and white: Assessing racial disparities in the circumstances of police killings"., In: *The Journal of Politics* 81.3 (2019), pp. 1124–1132.

– S. Müller., "The temporal focus of campaign communication"., In: *The Journal of Politics* 84.1 (2022), pp. 000–000.

Aula 5 – Aprendizado Supervisionado: Modelos Lineares

Leituras:

– ITSL, Cap. 6

Aula 6 – Aprendizado Supervisionado: Modelos Não-Lineares

Leituras:

– ITSL, Cap. 7

Aula 7 – Ensemble: Stacking, Bagging, Boosting

Leituras:

– ITSL, Cap. 8

Leituras sugeridas:

– A. R. Kaufman, P. Kraft, and M. Sen., "Improving supreme court forecasting using boosted decision trees"., In: *Political Analysis* 27.3 (2019), pp. 381–387.

– J. M. Montgomery and S. Olivella., "Tree-Based Models for Political Science Data"., In: *American Journal of Political Science* 62.3 (2018), pp. 729–744.

– L. Chen and H. Zhang., "Strategic Authoritarianism: The Political Cycles and Selectivity of China's Tax-Break Policy"., In: *American Journal of Political Science* 65.4 (2021), pp. 845–861.

– P. Broniecki, L. Leemann, and R. Wüest., "Improved Multilevel Regression with Poststratification through Machine Learning (autoMrP)"., In: *The Journal of Politics* 84.1 (2022), pp. 000–000.

Aula 8 – Aprendizado Não-Supervisionado

Leituras:

– ITSL, Cap. 12

Leituras sugeridas:

– Z. B. Magyar., "What makes party systems different? A principal component analysis of 17 advanced democracies 1970–2013"., In: *Political Analysis* 30.2 (2022), pp. 250–268.

– H. Mueller and C. Rauh., "Reading between the lines: Prediction of political violence using newspaper text"., In: *American Political Science Review* 112.2 (2018), pp. 358–375.

– C. Zucco Jr and T. J. Power., "Fragmentation without cleavages? Endogenous fractionalization in the Brazilian party system"., In: *Comparative Politics* 53.3 (2021), pp. 477–500.

Aula 9 – Resampling & Validação

Leituras:

– ITSL, Cap. 5

Leituras sugeridas:

– M. Neunhoeffer and S. Sternberg., "How cross-validation can go wrong and what to do about it"., In: *Political Analysis* 27.1 (2019), pp. 101–106.

– S. Raschka., "Model evaluation, model selection, and algorithm selection in machine learning"., In: *arXiv preprint arXiv:1811.12808* (2018).

Aula 10 – Tuning & Feature Selection

Leituras:

– MLR3, Cap. 4

Leituras sugeridas:

– M. J. Denny and A. Spirling., "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it"., In: *Political Analysis* 26.2 (2018), pp. 168–189.

– S. Jordan, H. L. Paul, and A. Q. Philips., "How to Cautiously Uncover the "Black Box" of Machine Learning Models for Legislative Scholars"., In: *Legislative Studies Quarterly* (2022).

Aula 11 – Deep learning

– ITSL, Cap. 10

– HML, Cap. 10.1

Leituras sugeridas:

– C. Chang and M. Masterson., "Using word order in political text classification with long short-term memory models"., In: *Political Analysis* 28.3 (2020), pp. 395–411.

– F. Cantú., "The fingerprints of fraud: Evidence from Mexico's 1988 presidential election"., In: *American Political Science Review* 113.3 (2019), pp. 710–726.

– D. Muchlinski, X. Yang, S. Birch, et al., "We need to go deeper: Measuring electoral violence using convolutional neural networks and social media"., In: *Political Science Research and Methods* 9.1 (2021), pp. 122–139.

Aula 12 – Revisão de Trabalhos Finais & Encerramento