ECO1681 – Machine Learning para Economistas*

Gilberto Boaretto – E-mail: gilbertoboaretto@hotmail.com

Departamento de Economia – PUC-Rio

2021.2

1 Introdução

Quando o número de regressores é muito grande ou mesmo maior que o número de observações, os métodos estatísticos e econométricos tradicionais – como o modelo linear estimado por mínimos quadrados – podem não apresentar boas propriedades ou mesmo não serem implementáveis. Tal fenômeno é cada vez mais frequente dado o volume cada vez maior de dados e informações existentes, o que ficou conhecido por *big data*. Em séries temporais, por exemplo, facilmente podemos ver casos nos quais o número de regressores é maior que a dimensão temporal. Assim, modelos e métodos que consigam lidar com grande volume de informação podem ser bastante úteis em problemas empíricos de Macroeconomia, Finanças, Organização Industrial, entre outras áreas.

2 Objetivos

Este curso visará apresentar ferramental estatístico/econométrico que permita lidar com vários desses desafios práticos. Tal ferramental é amplamente conhecido sob o nome *machine learning*, mas também designado por *statistical learning* ou mesmo econometria de alta dimensão. Fazendo uso de diferentes meios para superar a chamada "maldição da dimensionalidade", as técnicas lineares e não lineares precisam ser estudadas a fundo de modo ao usuário ter garantia de que as está empregando corretamente – e não apenas apertando botões ou rodando códigos alheios.

É esperado que ao final do curso os alunos tenham visto um vasto conjunto de modelos e técnicas que potencialmente possuem várias aplicações práticas. Além da apresentação com alguma profundidade teórica, o curso visará a implementação (programação) dos modelos/métodos estudados. Para tal, faremos uso principalmente do software estatístico **R**. Ao longo do curso serão dados exemplos práticos e os alunos deverão resolver listas que combinarão questões conceituais e práticas visando a fixação do conteúdo – além da atribuição de nota.

*Versão de 05/07/2021.

3 Conteúdo

O curso está estruturado em duas partes.

Parte I

1. Introdução: machine/statistical learning, econometria de alta dimensão; big data, dados estruturados *versus* não-estruturados, aprendizado supervisionado e não-supervisionado; previsão, inferência, ajuste *versus* interpretabilidade; programação em **R** e organização de dados.

   Ref. **JWHT** – Caps. 1 e 2.

2. Modelos lineares e métodos de penalização: MQO, Ridge, LASSO e família LASSO.

   Ref. **JWHT** – Caps. 3 e 6.

3. Escolha da penalização, algoritmos e avaliação de modelos

   Ref. **JWHT** – Cap. 5.

4. Aplicações: previsão de retorno de ações e previsão dos casos de Covid.

   Ref. **GKX**, 2018; **KNS**, 2020; **MSVVZ**, 2020.

Parte II

5. Redução de dimensionalidade: PCA, modelo de fatores e PCR.

   Ref. **JWHT** – Cap. 10.

6. Árvore de regressão e não linearidade: bagging, random forest e boosting.

   Ref. **JWHT** – Cap. 8.

7. Aplicações: previsão de atividade industrial e previsão de inflação.

   Ref. **SW**, 2002; **GMV**, 2017; **MVVZ**, 2021.

8. Se houver tempo hábil, poderemos falar um pouco sobre Redes Neurais, por exemplo.

4 Calendário/Cronograma

Data	Horário	Aula / Conteúdo
		Parte I
1. Introdução		
09/08/2021	9h-11h	Aula 1 – Introdução - O que é machine learning?
16/08/2021	9h-11h	Aula 2 – Organização de dados no **R**
2. Modelos lineares e métodos de penalização		
23/08/2021	9h-11h	Aula 3 – MQO e Ridge
30/08/2021	9h-11h	Aula 4 – LASSO
06/09/2021	9h-11h	recesso
13/09/2021	9h-11h	Aula 5 – Variações de LASSO
3. Escolha da penalização, algoritmos e avaliação de ajuste		
20/09/2021	9h-11h	Aula 6 – Escolha da penalização via critério de informação e via validação cruzada
27/09/2021	9h-11h	Aula 7 – Métricas de ajuste e avaliação de modelos
4. Aplicações		
04/10/2021	9h-11h	Aula 8 – Previsão de retorno de ações e previsão dos casos de Covid
11/10/2021	9h-11h	recesso
		Parte II
5. Redução de dimensionalidade: PCA, modelo de fatores e PCR		
18/10/2021	9h-11h	Aula 9 – Redução de dimensionalidade e análise de componentes principais (PCA)
25/10/2021	9h-11h	Aula 10 – Modelo de fatores e regressão com componentes principais (PCR)
01/11/2021	9h-11h	recesso
6. Árvore de regressão e não-linearidade		
08/11/2021	9h-11h	Aula 11 – Introdução à não-linearidade: os modelos de árvore de decisão
15/11/2021	9h-11h	Aula 12 – Bagging e Boosting
22/11/2021	9h-11h	Aula 13 – Random Forest
7. Aplicações		
29/11/2021	9h-11h	Aula 14 – Previsão de atividade industrial e previsão de inflação
8. Introdução a Redes Neurais		
06/12/2021	9h-11h	Aula 15 – Introdução a Redes Neurais, deep learning e redes rasas
Finalização dos trabalhos		
13/12/2021	9h-11h	Conversa livre: dúvidas e revisão

5 Avaliação

Serão dadas quatro listas conceituais e aplicadas ao longo do semestre. Presença e participação nas aulas também serão consideradas.

6 Referências Bibliográficas

Referência base

James; Witten; Hastie; Tibshirani (**JWHT**, 2017) – *"An Introduction to Statistical Learning"*.

Artigos

Gu; Kelly; Xiu (**GKX**, 2018). "Empirical Asset Pricing via Machine Learning". *The Review of Financial Studies*, 33(5), 2223-2273.

Kozak; Nagel; Santosh (**KNS**, 2020). "Shrinking the cross-section". *Journal of Financial Economics*, 135(2), 271-292.

Medeiros; Street; Valladão; Vasconcelos; Zilberman (**MSVVZ**, 2020). "*Short-Term Covid-19 Forecast for Latecomers*". arXiv preprint arXiv:2004.07977.

Stock; Watson (**SW**, 2002). "Forecasting Using Principal Components From a Large Number of Predictors". *Journal of the American Statistical Association*, 97(460), 1167-1179.

Garcia; Medeiros; Vasconcelos (**GMV**, 2017). "Real-time inflation forecasting with high-dimensional models: The case of Brazil". *International Journal of Forecasting*, 33(3), 679-693.

Medeiros; Vasconcelos; Veiga; Zilberman (**MVVZ**, 2021). "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods". *Journal of Business & Economic Statistics*, 39(1), 98-119.

Outras referências (nível intermediário)

Friedman; Hastie; Tibshirani (2009) – *"The Elements of Statistical Learning"*.

Hastie; Tibshirani; Wainwright (2016) – *"Statistical Learning with Sparsity"*.

Referências avançadas (nível de pós-graduação)

Bühlmann; van de Geer (2011) - *"Statistics for High-Dimensional Data"*.

Efron; Hastie (2017) – *"Computer Age Statistical Inference"*.

Fan; Li; Zhang; Zou (2020) - *"Statistical Foundations of Data Science"*.

Jolliffe (2010) – *"Principal Component Analysis"*.