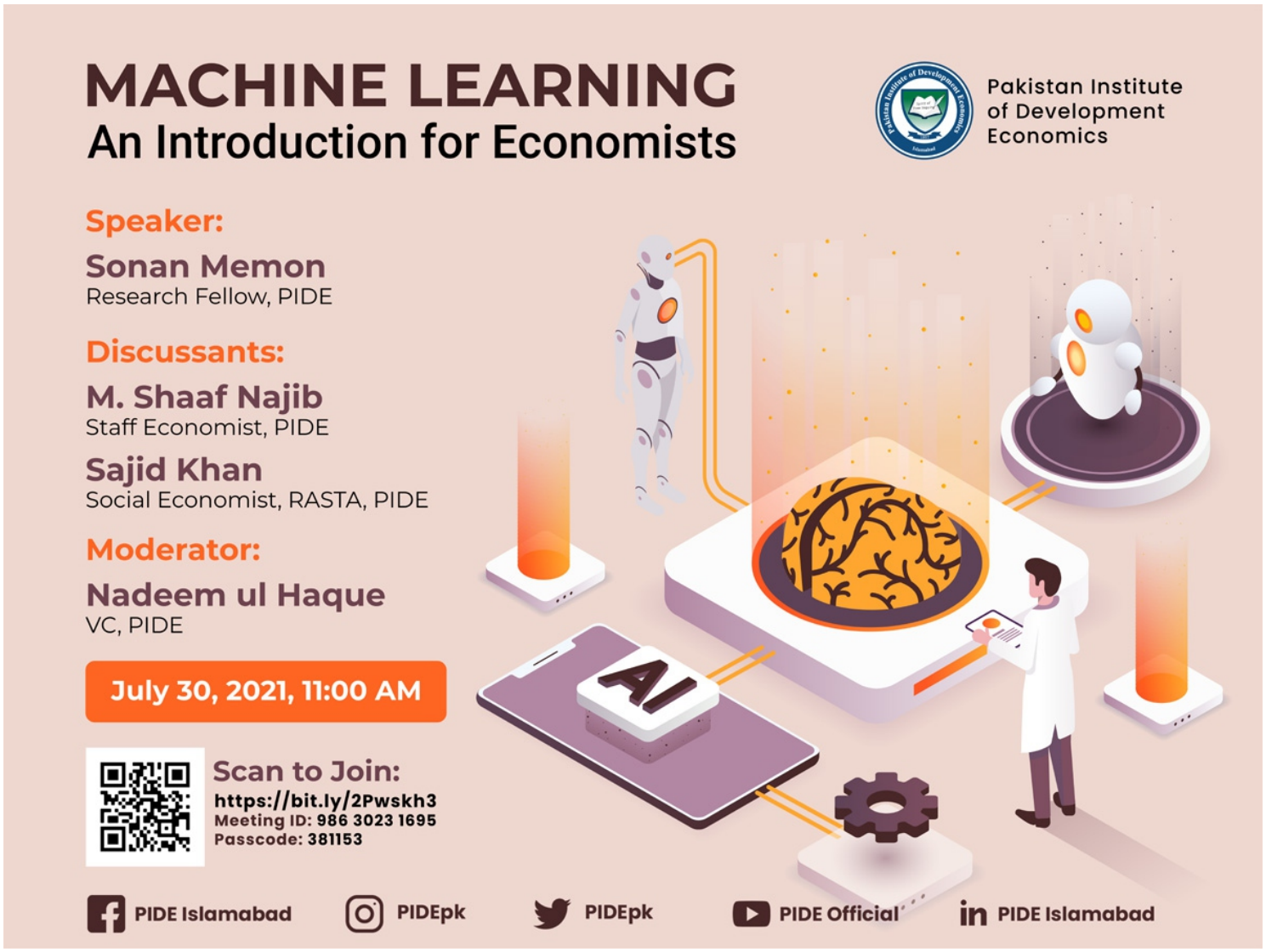
Webinar

Brief 62:2021

Machine Learning:

An Introduction for Economists

Pakistan Institute of Development Economics



Preamble

The objective of this webinar is to provide a brief and non-technical overview of; What Machine learning is and its recent applications in economic literature. This webinar deals with an important aspect of the usage of machine learning and discusses why machine learning tools needed to be incorporated in academic and policy-relevant research in Pakistan.

The major takeaway from the presentation

- Machine learning is a set of algorithms and computational methods which enable the computer to learn the complicated and non-linear pattern from the training data without being explicitly programmed to do so. Machine learning algorithms are complex function approximation techniques that can find highly complex, flexible, and generalizable structures in data.
- Machine Learning not only fits the highly complicated functional data but also finds the generalizable structures within the data., which makes Machine Learning techniques outperform traditional approximation techniques and methods.
- Explaining the major difference between Machine Learning and econometrics, Mr. Sonan pointed out that usually, an economist from the point of researching policy-relevant issues is interested in the causal inference and identification of the parameters. My prediction is the prime goal and the identification of the parameters is usually not valid; which could be interpreted as Machine Learning can't establish causal inference but can only interpret.

- Machine Learning methods can be broadly classified into four categories; supervised learning, unsupervised learning, multi arms bandits and reinforcement learning, causal trees, and heterogeneous treatment effects. The supervised and unsupervised learning methods' goal is to establish association and find mapping patterns between input X and output Y. Multi arms bandit method and causal trees are used in experiments on big data.
- One of the strengths of Machine Learning is that it can fit any input data very accurately, which leads to the problem of overfitting. The overfitting problem results in low sample bias but high variance. To navigate this variance-bias trade-off, regularization techniques are used, which allows generalizing beyond the training data. One popular method of implementing regularization is 'cross validation', which splits the data into K folds of roughly equal size. The holdout method is repeated K times by using each set only once for testing and training.
- Machine learning usage in economics and development has become increasingly prominent in the last few years. Popular mentions include the usage of LASSO for macroeconomic forecasting, big data in neuro-economics and big data from supermarkets; the usage of multi arms bandit technique to allocate optimal labor market intervention to improve job-finding rate for Syrian refugees, and lastly the usage of LDA techniques in computational linguistics to analyze central bank communication and to investigate the impact of news reporting on the household inflation expectation.
- Recent studies intensively use Machine Learning techniques for the developmental purpose such as the usage of satellite and mobile data to predict poverty; improving tax compliance in India to identify the suspicious firm which was less likely to file tax returns; using night and day time satellite data for measuring extend of urbanization; Poverty maps for Bangladesh and African countries.
- The least absolute selection & shrinkage operator (LASSO) is an extension of OLS. LASSO is the augmentation of OLS with a given specific constraint. Due to constraint many of the coefficients will be exactly zero and will be dropped out from the operation, this process is called selection. The retained set of coefficients will also be shrined towards zero since LASSO favors sparsity and this process is called shrinkage. The usage of selection and shrinkage operation is for the optimal variable selection when the number of covariates is large.
- Multi arms bandits solve the exploration versus exploitation trade-off through the optimal assignment of treatments. A/B testing becomes inefficient since it allocates a fixed number of units to each treatment, some of which could be sub-optimal. MAB solves this problem through prior treatment assignment probabilities for each arm. Application of MAB for adaptive targeted experiment focused on improving job-finding rate for Syrian refugees in Jordan. MAB algorithm balanced the goal maximization of welfare and precision of treatment effect estimates.
- LDA is a hierarchical Bayesian model developed for the tropical modeling of text corpora. It estimates predetermined numbers of K topics based on high-dimensional test data on documents. LDA is quite useful to extract a sparse and meaningful representation from textual data.
- Machine learning is termed as a "Black box" for raising ethical quandaries and lack of transparency when used for policy-relevant decisions. Amazon scrapped its AI tool for being biased, it showed significant bias against the female job applicants. Machine learning creates ethical issues when it clashes with legislation, oversight, and auditing.

Discussion

During the discussion, the questions were raised which were answered in detail by Mr. Sonan. Questions included:

How can machine learning aid in finding causality as it is the prime interest of the economist and social scientist?

Machine learning is a data-driven methodology but in the case of Pakistan, limited availability of data and sources to gather data are scarce. What are the sources to gather such datasets?

Cross-validation is inconsistent in the selection of the variable; what are the ways to avoid this particular issue?

How do LASSO or other methods select the variable in machine learning?

When we analyze the textual data, how can we avoid the false understanding arising from the big data?

In response to the questions posed, Mr. Sonan pointed out that machine learning techniques have been useful in finding out helpful insights and patterns, from the policy point of view despite that it can't find the causal inference. Mr. Sonan backed his argument with the research work done in Bangladesh that used granular data, for targeting the extreme poverty areas. Another Afghanistan was discussed where the identification of the ultra-poor was made possible through the usage of machine learning and mobile phones data. Moreover, the machine learning methodology can be combined with other methodologies such as IV techniques to assist in finding the causal inference and optimization of models.

Answering the question regarding data scarcity, Mr. Sonan believed that the lack of big data and spatial data is certainly a big concern. Recent studies in the field of developmental economics have shown that machine learning allows economists and social scientists to work in a data-scarred environment which was not possible before. Usage of the satellite data or mobile phone data coupled with economics data to find solutions to policy-relevant issues where economic data was limited or unavailable.

Machine learning is referred to as a black box because we don't know what happens under the hood? In LASSO there is the factor of constraint that the sum of the value of coefficient must be less than the C, the smaller the value of C the greater the value of sparsity. Algorithms identify the best coefficients which are highly correlated with the output and drop all the others which are less or uncorrelated. Answering the last question posed, Mr. Sonan believed that systemic inherent biases because of fake news, trend manipulation on Facebook and social media is certainly a valid concern as it alters the predictions and provides false results. Systemic selection biases in the training data will result in prediction with biases too.

Closing Note

Machine learning has opened up a plethora of new opportunities for research in economics and development. It is high time for Pakistan to leverage the power of machine learning for academic and policy-relevant issues; though limited availability of big and thick data will continue to be a challenge for Pakistan.

Prepared by
Zarak Jamal Khan

Edited by
Hafsa Hina

Design by
Afzal Balti

www.pide.org.pk
policy@pide.org.pk.
92-51-924 8051
92-51-942 8065