２０２１年度

大学院文学研究科博士課程前期２年の課程入学試験

（　秋期・社会人特別選抜　）問題

筆記試験　　計算人文社会学　専攻分野

試験開始の合図があるまで、この問題冊子を開いてはいけない。

受験記号番号	

成 績	

２０２１年度

大学院文学研究科博士課程前期２年の課程入学試験

（秋期・社会人特別選抜）問題

筆記試験（　　計算人文社会学　　　　専攻分野）

注意　問題用紙は2枚、解答用紙は3枚である。解答の順序は自由であるが、どの問題の解答であるかが分かるように、 問題番号を間違いなく記入すること。

問題 1.　インターネット上で行われる閲覧、検索、売買、コミュニケーション、その他の様々な行動の履歴を記録したデータをデジタルトレースデータという。このようなデジタルトレースデータを、人文・社会科学の研究で応用することの長所と短所はどこにあるか。サーベイデータなどの従来のデータの使用と比較しながら、できるだけ具体的に説明しなさい。

問題は次の頁に続く

問題2.　以下の英文を読み，問いに答えなさい．

(1) 下線部 (A)を日本語に訳しなさい．

(2) 下線部 (B)を日本語に訳しなさい．

(3) 下線部 (C) のように述べる理由はなにか．本文に即して説明しなさい。

Causal Inference

As a primer, consider the fundamental problem of causal inference: We observe an individual (or any unit of analysis) in one condition alone (treatment or control) and cannot measure individual-level variation in the effect of the treatment (for authoritative reviews of causal inference in the social sciences, see Morgan &Winship 2007, 2014). We instead focus on an aggregate average effect that we treat as homogeneous across the population (Xie 2013). (A) In experimental design, we randomly assign individuals to treatment and control groups and directly estimate the average causal effect by comparing the mean output between the groups (Imbens & Rubin 2015).

Social scientists now use SML* to identify heterogeneous treatment effects in subpopulations in existing experimental data. For example, Imai & Ratkovic (2013) discover groups of workers differentially affected by a job training program. They interact the treatment (i.e., being in the program) with different inputs and use a lasso model to select the inputs that are most important in predicting increases in worker earnings. Similarly, Athey & Imbens (2016) develop causal trees to estimate treatment effects for subgroups. Different from standard regression trees in ML** (where one seeks to minimize the error in predictions, $\hat{Y}$), causal trees focus on minimizing the error in treatment effects. One can then obtain valid inference for each leaf (subgroup) with honest estimation, that is, by using half the sample to build the tree (select the optimal partition of inputs), and the other half to estimate the treatment effects within the leaves. Wager & Athey (2018) extend the method to random forests that average across many causal trees and allow for personalized treatment effects (where each individual observation gets a distinct estimate). Similarly, Grimmer et al. (2017) propose ensemble methods that weight several ML models and discover heterogeneous treatment effects in data from two existing political science experiments.

(B) Most empirical work in sociology relies on observational data where we do not control assignment to treatment. One way to estimate the causal effect in this case is to assume the potential output to be independent of assignment to treatment, conditional on other observed inputs. Under this so-called selection-on-observables assumption, we can estimate a causal effect by matching treatment and control groups on their propensity score (that is, the likelihood of being in the treatment group conditional on inputs). (C) Estimation of this score is well suited to SML as it involves a prediction task (where the effects of inputs are not of interest). Recent work uses boosting (McCaffrey et al. 2004), neural networks (Setoguchi et al. 2008, Westreich et al. 2010), and regression trees for this task (Diamond & Sekhon 2013, Hill 2011, Lee et al. 2010, Wyss et al. 2014) as alternatives to traditional logistic regression.

*SML: Supervised Machine Learning,　**ML: Machine Learning,

Molina Mario and Filiz Garip, 2019,Machine Learning for Sociology, *Annual Review of Sociology,* 45, 27-45（より p.34 の一部を抜粋）

受験記号番号	

受験記号番号	

受験記号番号	