“Primum Non Nocere” for Personalized Education

Loizos Michael
Open University of Cyprus
loizos@ouc.ac.cy

Abstract

In offering personalized education, one should be wary of the effects that an educational intervention may have on a student. Much like in medical ethics, we posit that the principle of non-maleficence should be a cornerstone in the design of educational systems. We provide a high-level and rather non-technical overview of ongoing work, in which a learning-theoretic framework that can be used to investigate and address the issue of interest herein, is proposed.

The dream of offering personalized education is steadily becoming a concrete reality. Beyond the technological issues that need to be addressed before this goal is fully met, there are also certain conceptual or ethical issues that merit careful examination. One such issue relates to a principal and widely-familiar maxim from medical ethics — “*primum non nocere*” — positing that the physicians’ primary concern should be to do no harm through their intervention. The same stance towards this non-maleficence principle should be adopted by any system offering personalized education.

Beyond the ethical considerations surrounding this principle, its adoption is especially pertinent on account of the very tangible ramifications it may have: Presenting a student with personalized educational material raises the distinct possibility of discouraging the student from studying, leading, in turn, the student to achieve performance that is worse than what the student would have achieved had the personalization been absent. Phenomena like this one are closely related to what scholars have called “self-defeating prophecies”. In the particular context of personalized education, the “prophecy” amounts to the prediction that a student will perform in a certain manner (according to a pre-specified standard) at the end of the course. The personalization that is, thus, triggered risks making the predicted outcome false.

A solution to the problem of avoiding such self-defeating prophecies should be sought in the design of the system that makes the aforementioned predictions. A typical architecture for such a system features a central learning module, so that the system makes predictions on the future performance of a student by induction from the monitoring of the performance of numerous other students in courses offered in the past. The problem becomes, then, one of adapting the learning module so that it takes into account the effects of its own predictions (through the triggered personalization).

To our knowledge this question has not been formally addressed in a learning-theoretic context. Typical approaches that seek to predict a future state of affairs (see, e.g., [Box and Jenkins, 1990; Murphy, 2002; Michael, 2011]) proceed on the basis that the predictions remain outside the state of affairs of the object-level world. We suggest that a solution to the issue at hand may come by making explicit the process of recording these predictions in the current state of affairs, before the latter evolves to a future state of affairs. We provide, in the sequel, the basic ideas of ongoing work that seeks to address this issue [Michael, 2012], appropriately presented for the domain of personalized education.

Denote by $\mathcal{S}$ the set of all possible student profiles, which we take to incorporate any course information available to students, and the state of their knowledge as measured by predetermined tests. Denote by $\mathcal{P}$ the set of all outcomes that the educational system wishes to predict. Denote by $\mathcal{T}$ a set of functions from $\mathcal{S}$ to $\mathcal{P}$, so that each function $t \in \mathcal{T}$ represents a way in which student profiles may be mapped to outcomes, with an unknown such function $t^* \in \mathcal{T}$ being the actual one. We shall assume that class $\mathcal{T}$ is learnable by some algorithm $\mathbb{A}$ (e.g., under the PAC semantics [Valiant, 1984] and extensions that account for a causal setting [Michael, 2011]), in the following sense: If a system passively observes student profiles and actual outcomes $\langle s, t^*(s) \rangle$ at two check points during the course (say, the student profile after the first homework grading, and the outcome being the final course grade of that student), it eventually identifies a hypothesis h that given the profile $s \in \mathcal{S}$ of a student at the first check point predicts with high probability the outcome $t^*(s)$ for that student at the second check point. The question, then, is this: How can algorithm $\mathbb{A}$ be transformed to algorithm $\mathbb{B}$ so that its predictions are still reliable despite their announcement (through the triggered personalization)?

The process of announcing a prediction $p \in \mathcal{P}$ corresponds to a recording process $\triangleleft$ that maps any student profile $s \in \mathcal{S}$ to the profile $s \triangleleft p$ that is the result of the personalization for student s on the basis of prediction p. We assume only that the process $\triangleleft$ is fixed and efficiently computable. The goal, now, becomes that of identifying a hypothesis h that given the profile $s \in \mathcal{S}$ of any student at the first check point, predicts with high probability the actual outcome $t^*(s \triangleleft h(s))$ for that student at the second check point, taking into account that, due to personalization, the actual outcome is a function of the affected profile $s \triangleleft h(s)$. Thus, unlike algorithm $\mathbb{A}$, which seeks to learn a hypothesis h such that $h(s) = t^*(s)$ with high probability, algorithm $\mathbb{B}$ seeks to learn a hypothesis h such that $h(s) = t^*(s \triangleleft h(s))$ with high probability. Also unlike the former setting where the training instances comprise pairs $\langle s, t^*(s) \rangle$, now the learning algorithm observes a student profile $s \in \mathcal{S}$, chooses a prediction p, and then observes the outcome $t^*(s \triangleleft p)$.

The presence of $h(s)$ on both sides of the equation $h(s) = t^*(s \triangleleft h(s))$ makes it unclear whether such a hypothesis h even exists. Indeed, depending on $t^* \in \mathcal{T}$ and the probability distribution that determines the relative frequency with which student profiles in $\mathcal{S}$ are observed, with some unknown probability α the chosen student profile $s \in \mathcal{S}$ will be such that no predicted outcome p satisfies the equation $p = t^*(s \triangleleft p)$; these profiles could correspond, for instance, to students that actively seek to falsify the prediction p. For such cases, nothing can be done, and any returned hypothesis h will make an incorrect prediction. With probability $1 - \alpha$, however, a student profile $s \in \mathcal{S}$ will be chosen so that there exists a prediction p satisfying the equation $p = t^*(s \triangleleft p)$. For such cases, the returned hypothesis h is expected to be such that $h(s) = p$. How can such a hypothesis be efficiently and reliably learned (e.g., under the PAC semantics)?

The answer is surprisingly simple. During the training phase, algorithm $\mathbb{B}$ makes uniformly random predictions p, and triggers the corresponding personalization, so that it observes student profiles $s \in \mathcal{S}$ and actual outcomes $t^*(s \triangleleft p)$ for the chosen predictions p. Omitting the details of the proof, this strategy ensures that algorithm $\mathbb{B}$ can construct a hypothesis h that will satisfy the equation $h(s) = t^*(s \triangleleft h(s))$ with probability arbitrarily close to the optimal $1 - \alpha$. In terms of computational efficiency, algorithm $\mathbb{B}$ requires time polynomial in $\log |\mathcal{S}|$, $|\mathcal{P}|$, $1/(1 - \alpha)$, indicating that large numbers of students and long student profiles are possible without compromising the efficiency of the process.

Returning to the issue of non-maleficence, having constructed hypotheses h_1 and h_2 by executing algorithms $\mathbb{A}$ and $\mathbb{B}$ as described above, and given a student profile $s \in \mathcal{S}$ at the first check point, an educational system may predict the outcome $h_1(s)$ and $h_2(s)$ for that student at the second check point, according to the two hypotheses at hand. The first prediction corresponds to how the student will fare without intervention, while the second one to how the student will fare with — and because of — the particular intervention suggested by the prediction $h_2(s)$ itself. The educational system may then trigger the prescribed personalization if indeed its consequences $h_2(s)$ do no harm over $h_1(s)$. The case of multiple competing personalization strategies for each student profile corresponds to having multiple recording processes $\triangleleft^i$. A hypothesis h_2^i can be constructed for each $\triangleleft^i$, and its predictions checked against those of h_1 and the remaining h_2^j, so that both harm is prevented, and the best personalization strategy is applied for each student profile.

The ethical dilemma is not, however, entirely addressed, since random interventions will be required during the training phase, until the effects of the interventions can be reliably established. One could attempt to argue that these interventions should not count as maleficence on the grounds that the educational system is still not able to anticipate the effects of its interventions. A more pragmatic point of view is that, as in the early steps of every medical discipline, some harm is inescapably inflicted on certain individuals so that others may benefit from it. Fortunately, the harm to the students during the training of the educational system is simply a bit of extra, randomly assigned, homework!

References

[Box and Jenkins, 1990] George E. P. Box and Gwilym Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

[Michael, 2011] Loizos Michael. Causal Learnability. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, pages 1014–1020, July 2011.

[Michael, 2012] Loizos Michael. Introspective Forecasting. *Unpublished Manuscript*, 2012.

[Murphy, 2002] Kevin P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. Dissertation, University of California, Berkeley, Computer Science Division, July 2002.

[Valiant, 1984] Leslie G. Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.