Abstract

Randomised and controlled experimentation is not always possible when evaluating AI agents over performance tests. Intelligence or performance testing environments are usually designed and/or modified to the specifications of the evaluated agents. In this paper, we discuss the use of the propensity score matching statistical technique for eliminating or reducing bias when analysing and comparing agent performance-scores across different environmental (test) settings. We show how, once matching has been achieved, the contrast in performance scores between two evaluated agents can vary significantly. Therefore, matching on propensity scores can enable a fair comparative assessment of performance between agents that cannot be evaluated over identical environments and experimental settings.

1 Introduction and Motivation

The evaluation of artificial intelligence has become very popular in the last decade with new environments designed for assessing various sorts of AI [Chmait, 2017; Chmait et al., 2017; Hernández-Orallo, 2017; Chmait et al., 2016a; Chmait et al., 2016b; Insa-Cabrera et al., 2012]. In fact, the evaluation of general-purpose AI has claimed its own success with state-of-art models and evaluation techniques targeting the measurement of general intelligence in machines presented year after year [Hernández-Orallo et al., 2017]. Nevertheless, many barriers are yet to be overcome in this field of research. For instance, although agents are being designed to solve more general problems in AI, (universal) intelligence tests that can be administered to different sorts of artificial agents, under identical experimental settings, are still far-off from being attained. Even when evaluating model-free reinforcement learning agents, test environments need to be tuned appropriately. Moreover, at the time being, there is no feasible way to accurately measure complexity [Hernández-Orallo, 2015] across different types of assessment tasks in order to make sure that agents are evaluated over tasks of similar complexities/difficulties. Consequently, in many scenarios, artificial agents are still evaluated over different environments and performance tasks, and under different environmental or test settings, which inhibits our ability to precisely compare and contrast their performances to one another.

The motivation behind this work stems from the latter problem. To that end, with AI agents evaluated across a range of different environments that are usually tuned to the parameters and specifications of these agents, how can we precisely compare the performances of such agents in an unbiased way? We propose the use of Propensity Score Matching (PSM) for reducing assignment-bias when allocating agents to performance-evaluation tasks and canonical test problems. As a result, the comparative assessment of performance between agents that cannot always be evaluated over identical settings is made feasible.

This paper is organised as follows. The next section gives an overview of the PSM statistical technique. We then discuss how to apply PSM to balance the factors/covariates affecting the measurement of performance between two agents evaluated over a series of tasks. Finally, we give an illustrative example of how such bias can be reduced (or, in the best case scenario, eliminated) by comparing covariate balance before and after the application of PSM to sample test scores.

2 Propensity Score Matching

Rosenbaum and Rubin have first introduced the concept of PSM in 1983 [Rosenbaum and Rubin, 1983]. The overall idea behind PSM is straightforward. In statistical terms, with the absence of randomised controlled trials, the assignment of treatments to subjects is usually non-random. Therefore, subjects receiving or excluded from treatment will not only differ in their treatment condition, but also in other properties or characteristics [Heinrich et al., 2010; Thavaneswaran, 2008]. To eliminate selection bias, the PSM technique matches treated and untreated observations on the estimated probability of being treated which is calculated as their propensity score. More technically, the propensity score [d’Agostino, 1998] for subject $i$ s.t. $(i = 1, 2, \ldots, N)$ is the conditional probability of being assigned to a particular treatment $Z_i = 1$ versus a control $Z_i = 0$ given a list of some observed attributes $x_i$ (called covariates or pre-treatment variables) where:

$$ps(x_i) \equiv pr(Z_i = 1|X_i = x_i) \in \{0, 1\}$$
Usually, \( ps \) is estimated via discriminant analysis or using a logistic regression. It is assumed that the \( Z_i \)s are independent given the \( X \)’s as follows:

\[
pr(Z_1 = z_1, \ldots, Z_N = z_N | X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^N ps(x_i)^{z_i}(1 - ps(x_i))^{1-z_i}
\]

PSM ensures balanced covariates (corresponding to the \( X \)s), where a balancing score, \( \text{balance}(X) \) is a “function of the observed covariates \( X \)” s.t. the conditional distribution of \( X \) given \( \text{balance}(X) \) is the same for the treated \( (Z = 1) \) and control \( (Z = 0) \) units” [Rosenbaum and Rubin, 1983]. Therefore, matching or regression (covariance) adjustment on \( ps \) produces unbiased estimates of the treatment effects, when the treatment assignment is un-confounded [Rosenbaum and Rubin, 1983] [d’Agostino, 1998]. The un-confounding property (of the treatment assignment) is satisfied if: \( Z \), the treatment assignment, and \( Y \), the response or potential outcome of the experiment, are known to be conditionally independent given the covariates, \( X \) (for example if \( Y_0, Y_1 \perp \perp Z | X \) where \( Y_0 \) and \( Y_1 \) are respectively the potential outcomes under control and treatment).

In summary, PSM can allow us to estimate the causal effect of a treatment by eliminating assignment bias of treatments to subjects. This is achieved by making sure that subjects in treatment and control groups that have equal (or similar) propensity scores have similar distributions on their pre-treatment variables (background covariates).

### 3 A Demonstration Using a Hypothetical Example

We present a hypothetical scenario in which two agents \( A \) and \( B \) are evaluated over a series of tasks from an intelligence/performance test. For our purposes, we assume that the scores from these experiments reflect the ability of the testee in solving the tasks, but do not factor in other important parameters that can have an impact on the performance of the evaluated agents. Examples of such confounding parameters are the following:

1. the number of test iterations before an agent returns an answer to the test,
2. the processing time per iteration,
3. the agent’s memory requirements,
4. the number of bits received from the environment as an observation (in the context of an agent-environment framework [Hutter, 2004]).

An artificial dataset of was created corresponding to a table of five columns holding a list of scores for agents \( A \) and \( B \)

\[\text{Score estimates} \quad \text{Std. Error} \quad pr(> |t|)\]

|                | Score estimates | Std. Error | \( pr(> |t|) \) |
|----------------|-----------------|------------|----------------|
| (Intercept)    | 90.520          | 4.058      | \(< 2 \times 10^{-16} \) |
| Agent B        | -6.621          | 7.444      | 0.374          |

Since it was assumed that the tests scores do not account for confounding parameters that could have influenced the agents’ performances, the OLS estimates in Table 1 can be biased.

In order to guarantee an unbiased comparative analysis of performances, we can balance the experiments on all the characteristics (the covariates or pre-treatment variables that could be confounding) that could influence the outcome (the scores) from these experiments. Therefore, we encode the confounding parameters 1 to 4 as covariates in a PSM model. In other words, we balance the (propensity scores of) agents \( A \) and \( B \) according to the values of the 4 confounding covariates. In the context of PSM, agents \( A \) and \( B \) in this example would correspond to the treatment and control units respectively.

The propensity score distribution from the PSM is shown in Figure 1. The circles in Figure 1 represent the propensity scores from each experiment. We observe a close match between the treatment units and control units (no unmatched

\[\text{Distribution of Propensity Scores}\]

Figure 1: Propensity score distribution.
treatments). The unmatched control units are discarded from the comparison (as they correspond to biased observations).

It is important to note that discarding test results (unbalanced experiments) would return less accurate estimates of the overall performance of the agents since their scores are extracted from a narrower set of tests. Nevertheless, the contrast in performances between the evaluated agents over the balanced tests is feasible and accurate. Of course, the overall performance of the agents can always be extracted as their average scores over the complete set of experiments. However, as described before, contrasting the agents’ performances using the average scores is arguably unfair since each agent was operating under different environmental/test settings, and had distinct assessment requirements. We note in passing that there are (propensity score weighting) techniques available to reduce the number of discarded/unmatched units which might result from performing an exact (e.g., one-to-one) matching on propensity scores.

The balance before and after matching is illustrated in Figure 2. The histograms clearly differ before matching (Figure 2, left) but turn out to be identical after matching (Figure 2, right). This indicates that matching was successful.

![Figure 2: Histograms of the propensity scores before and after matching.](image)

After matching was successfully completed, we repeated the OLS regression contrasting the score estimates of agents A and B using the matched/balanced dataset of test scores.

| Agent | Score estimates | Std. Error | Pr(>|t|) |
|-------|-----------------|------------|---------|
| (Intercept) | 72.781 | 6.085 | < 2e-16 |
| Agent B | 11.118 | 8.606 | 0.197 |

Results from the second OLS regression are listed in Table 2. These results show that, after matching, the performance estimate for agent B appears to be (11.11 units) higher compared to the default treatment class, agent A.

This is significantly different from our previous conclusions as, after bias (due to confounding covariates) has been eliminated, B clearly outperforms A. The same results can also be drawn from a two sample t-test performed on the score vectors of agents A and B using the balanced dataset. Given that PSM has eliminated bias from our experiments, the difference in performance estimate from the OLS regression on the balanced data can be interpreted as the causal effect of introducing agent B as a test subject as opposed to (the control unit) A being the testee.

It is noteworthy to mention that, even if we control for the confounding covariates in the OLS regression (by introducing them as dummy variables), the contrast in performance between the two evaluated agents would still be remarkable after matching.

One limitation of this study is that we make the assumption that the test scoring methodology does not capture all the characteristics that can have an impact on the performance (or score) of the evaluated agent. While this is the case for many testing environments, there are others which penalise agents according to time, memory requirements, etc. Therefore, in such case, analysis of performance before and after matching should be identical and a PSM is not required.

Finally, we point out that the use of the PSM technique can be extended to multi-agent scenarios in which the focus becomes on understanding the (change in) group performance after the introduction of a new agent (the treatment) into a group of co-operative or competing agents.

### 4 Conclusion

The motivation behind this paper was to demonstrate the potential advantage of using the propensity score matching (or weighting) statistical technique to reduce bias when analysing and comparing performance test scores between artificial agents. Bias can arise as a result of various parameters, referred to as covariates, that might directly or indirectly impact the score of an agent over a performance task. Such parameters are not always factored-in as part of the test scoring methodology. The matching by propensity scores balances the observations according the values of the covariates. Unmatched or biased units are discarded from the analysis. As a result, the PSM is shown to return more accurate or realistic interpretation of the difference in performance between two or more evaluated agents.

While this paper discusses the potential advantage of using PSM for comparing artificial agent performances, we believe that PSM could even prove to be useful for comparing performance between human and AI agents [Insa-Cabrera et al., 2011]. In such experimentation, it is extremely difficult to implement identical test settings that apply to both humans and artificial agents over a range environments and tasks. For instance, factors like speed and memory tend to be vastly different. It is possible that PSM can help alleviate this problem by reducing experimental bias, opening new doors for comparing humans to artificial systems.
References


