Effectiveness of propensity score methods with density estimation in identifying overlap for causal inference

Jonathan Tjong

Supervisor(s): Jesse Krijthe, Rickard Karlsson

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2023

Name of the student: Jonathan Tjong
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Rickard Karlsson, Frans Oliehoek

Abstract

For causal inference, sufficient overlap is needed. It is possible to use propensity scores with the positivity assumption to ensure overlap is present. However, positivity is not enough to properly identify the region of overlap. For this, propensity scores need to be used in combination with density estimation. This project aims to evaluate this method, discovering in which scenarios it performs well or fails in identifying the region of overlap. More specifically, how it scales with more features or outliers, and how using different classifiers affects the performance. The method was tested with samples from a simulated dataset. The predicted overlap was compared with the true overlap of the known distributions. Following the experiments, the method seems to perform best when the treatment and control groups share one region of overlap. In this case, logistic regression works best out of the classifiers that were tested. The overall performance drops when the two groups have multiple regions of overlap. For this, the random forest classifier performs best instead. Throughout all scenarios, the performance of the model drops with increasing dimensionality. Furthermore, having a small percentage of outliers only slightly affects the model. With more outliers, logistic regression is the only classifier further affected.

1 Introduction

When researchers want to see the effects of a newly developed vaccine, the participants in the treatment group and control group that will be compared need to be similar. When there is a subset that contains participants from both the treatment and control groups that are similar to each other, it is called overlap. With a sufficient amount of overlap, factors irrelevant to the research or experiment that may negatively influence the results are minimized [2]. However, sufficient overlap is not a given when performing observational studies. Here the researchers have no influence on the allocation of treatments to the subject. As such this allocation is not randomized and will likely not have sufficient overlap. Thus, it is crucial to find out if there is sufficient overlap, and if there is not, which samples do make for sufficient overlap.

One can account for sufficient overlap with the positivity assumption that comes with propensity scores. The propensity score is the conditional probability that a person or sample belongs to the treatment group [12]. The positivity assumption says that the propensity score should be bounded away from 0 and 1 [7]. This assumption ensures that all samples have comparable samples in the other group, keeping the groups similar to each other. However, the typical methods used to verify positivity are not fit to identify the exact region of overlap. Features describe the participants (e.g. age, height, etc.) and each participant has their own values for each feature. Propensity scores only represent the conditional probability of a participant's feature values given which group the participant belongs to. Overlap on the other hand, relies on the opposite conditional probability: which group the participant belongs to given the participant's values. By using both propensity scoring and density estimation, it is possible to get the latter conditional probability and make predictions that more accurately resemble the actual region of overlap.

In this project, instead of using the typical propensity score methods, propensity scoring in combination with density estimation will be used to identify the region of overlap. The aim is to answer the following research question:

> *When do propensity score methods with density estimation work well and when do they fail to identify overlap for different types of datasets?*

Thus, the aim of the project is to evaluate this method and discover when they perform well or when they can be less reliable. For instance, how do they scale with more features or outliers? It is common to use logistic regression for propensity score methods [15], but how do other classifiers perform in comparison?

While evaluating the method in different contexts, it will be implemented to be part of a shared open-source package in Python, containing several other similar methods that also aim to evaluate overlap. This package will make these methods more accessible for other researchers who may need to evaluate overlap for their own research.

The structure of the paper is as follows. Section 2 covers the problem description, providing the background for the project. The methodology will be discussed in section 3. Section 4 is about the experiments and the results. Section 5 looks into responsible research and how it is handled throughout the project. Section 6 is the discussion and finally, section 7 contains the conclusions and possible future work that could be done.

2 Problem Description

This section will give more context about the problem and explains how this project aims to answer the research question. It will also look at other works related to this project.

2.1 Background

For a given threshold $0 < \epsilon < 1$, overlap is defined as the region of X where:

$$P(X = x|T = t) > \epsilon \text{ for both T = 1 and T = 0} \quad (1)$$

Where T = 1 indicates the sample belonging to the treatment group and T = 0 indicates the control group. Confounding factors are factors that were not accounted for in the experiments, but which may still negatively influence the results. To minimize confounding factors, samples in the treatment group and control group should be similar to each other [2].

One way of accounting for sufficient overlap is with the positivity assumption that comes with propensity scores. The

propensity score is $P(T = 1|X = x)$, the conditional probability that a person or sample belongs to the treatment group [12]. The positivity assumption states that the propensity score should be bounded away from 0 and 1 [7]. For observational studies, if a sample would violate this, it would mean that there would be no comparable samples in the other group. Thus, the results may be biased if the positivity assumption is neglected.

However, the positivity assumption cannot precisely identify the region of overlap. For the actual region, by our definition, $P(X = x|T = t)$ is needed. The positivity assumption is only concerned with the conditional probability $P(T = 1|X = x)$. To more accurately predict this region of overlap, density estimation can be used in combination with propensity scoring.

By Bayes' theorem:

$$P(X = x|T = t) = \frac{P(T = t|X = x)P(X = x)}{P(T = t)} \quad (2)$$

Where $P(T = t|X = x)$ is the propensity score and $P(T = t)$ can be estimated by the treatment and control group size. $P(X = x)$ can be estimated with density estimation. With this equation $P(X = x|T = t)$ can be obtained, from which the region of overlap can be estimated.

2.2 Related works

Several other methods have been proposed, with the aim to identify if a sufficient amount of overlap is present.

Crump et al. [4] proposed a method in which the samples outside the interval $[\alpha, 1 - \alpha]$ are discarded, more specifically the interval $[0.1, 0.9]$. This is in line with the positivity assumption and follows the common practice of researchers discarding sets of samples that do not have a similar counterpart in the other group. They showed this method reduced the asymptotic variance of the average treatment effect (ATE) compared to using the whole dataset.

Traskin and Small [16] criticized this method for not being interpretable, as it is hard to understand the propensity scores that the method uses. They aimed to improve upon this method by using a classification tree. This method would result in a tree diagram, which described the study population. However, while it was more interpretable, it performed worse than the method of Crump et al.

Oberst et al. [8] also had the same criticism and came up with an even more interpretable method called OverRule. This rule-based classification method uses the predicted overlap labels from another overlap estimator, a propensity score estimator for example. From this, it creates intuitive rules and a region of overlap. These rules are human-readable and can even be understood by people without a background in machine learning.

3 Methodology

This section will go over how the data was generated, the metrics used to measure the performance of the model, and the estimators used for the model.

3.1 Simulated Data

For these experiments, simulated data is used as opposed to real data. This is because it is possible to set up and tune the data, such that it represents a specific scenario of choice. Not only does this apply to the initial distributions of the treatment and control group, but also adding outliers can be done according to what is needed. This makes it very flexible and consistent in use. The simulated data is generated using the NumPy [6] library. The same experiments were performed 15 times with different random seeds, after which the results were averaged out.

To measure the accuracy of the model, the true overlap needs to be known. To find the true overlap, SciPy [17] was used, specifically the module `scipy.stats.multivariate_normal`. Using this, it is possible to get the probability density function and thus the density of the distribution for a specific location. It is then possible to get the region of overlap where the densities of both classes are higher than the overlap threshold ϵ.

3.2 Metrics

The primary accuracy metric used in this project is Intersection over Union (IoU). IoU is the most used metric when it comes to comparing two overlapping shapes [10], for instance when performing object detection. The main concept is intuitive, as the intersection (the area that is shared) is compared to the union (the total area the shapes cover). For classification, the IoU can be calculated as follows:

$$IoU = \frac{TP}{TP + FP + FN} \quad (3)$$

Where TP, FP, FN are True Positive, False Positive, False Negative respectively, as can be found in the confusion matrix. In this scenario, a positive sample means a sample located in an overlap region, whereas a negative sample is located outside any overlap region.

3.3 Estimators

The propensity score method relies on using a machine learning classifier to compute the propensity scores. Logistic regression is most commonly used for this [19]. However, it is a linear classifier, so it will have more difficulty when a problem is non-linear. As such, decision tree classifiers and random forest classifiers, which are non-linear, will also be used in this project. Their performance will be compared with logistic regression in specific scenarios. Scikit-learn [9], a machine learning package in Python, was used for the classifiers in the code.

Additionally, to more accurately predict the region of overlap, we want to use the probability distribution, which can be found with density estimation. In this project, kernel density estimation will be used for this. It is a non-parametric estimator, which makes it very flexible. In the code of this project, Scikit-learn was used to implement this method.

In order to perform well, the hyperparameters of the classifiers and the kernel density estimator need to be tuned. Since Scikit-learn was used for these, it is possible to use `sklearn.model_selection.GridSearchCV`. This module exhaustively generates all combinations of the parameters it

is given. By using cross-validation, it tries to find the best combination of hyperparameters.

Instead of the default "accuracy" scoring method, the "log-loss" measure was used for tuning the classifiers with GridSearchCV. The log-loss measure is most suited for scoring the predicted probabilities [18]. Log-loss penalizes the model more heavily if the model is confidently wrong about its predictions. Since the predicted probabilities need to correctly correspond to the confidence level for propensity scoring, it makes sense to use the log-loss measure here.

3.4 Method

As such, the method works as follows:

- Firstly, the model is initialized with the machine learning classifier of choice.
- As input it needs to know for each sample which class it belongs to and its values for each feature. Additionally, the overlap threshold ϵ needs to be chosen.
- The optimal hyperparameters of both the classifier and density estimator need to be chosen.
- The model will be fitted to the training data.
- Finally, it will predict for each sample if they are in the overlap region or not. The output of the model will be calibrated, returning the calibrated probabilities instead.

4 Experiments and Results

This section will go over the experiments that were conducted and will discuss the results that were produced. These experiments aim to answer how scalable the method (using different classifiers) is with number of features, with a linear and non-linear problem, and how well it performs with an increasing amount of outliers.

4.1 Scalability with Number of Features and Samples

In the following experiments, the method will be tested with an increasing number of features, increasing the dimensionality of the sample data. The overlap threshold ϵ was kept at 0.05 across all experiments that were performed.

When generating the same distributions, but increasing the dimensionality, the region of overlap changes significantly. To make fair comparisons between the different dimensions, it was decided to keep the size of the overlap region (the percentage of samples in the overlap region) approximately the same. This was done by changing the means and the standard deviations of the distributions. In these experiments, the overlap region was kept around 30%.

One Region of Overlap

Firstly, the two classes will be normal distributions next to each other. This means there is only one region of overlap between the two classes.

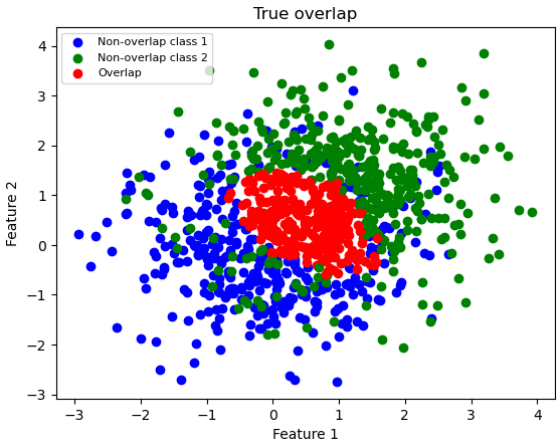

Figure 1: True overlap of two Gaussians (red points) in 2D, 500 samples each. Non-overlapping points of class 1 in blue and class 2 in green.

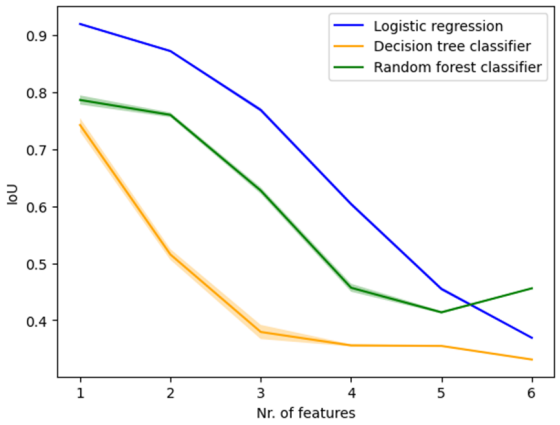

Figure 2: Performance using different classifiers with an increasing number of features. One region of overlap (30% of the data) between two classes. The uncertainty in the Figure is the variance of the model.

An example can be seen in Figure 1 above. The example has 2 features, so 2 dimensions. The green and blue points are the non-overlapping samples from the two classes. The red points are the samples in the overlap region.

After performing the experiment, the results were plotted in Figure 2. Figure 2 shows that logistic regression generally performs better than the other two classifiers, with the random forest classifier being better than the decision tree classifier. This problem, two Gaussians overlapping, is perfect for a linear classifier because it is linearly separable. As such, it is expected that logistic regression works well in this scenario

The performance for all classifiers generally seems to drop with higher dimensionality. This may be caused by the "curse of dimensionality" that the classifiers suffer from [3, 5, 14]. With higher dimensions, a bigger increase in sample size is needed to keep the same accuracy. Since the sample size is kept the same in this experiment, the accuracy of the model is

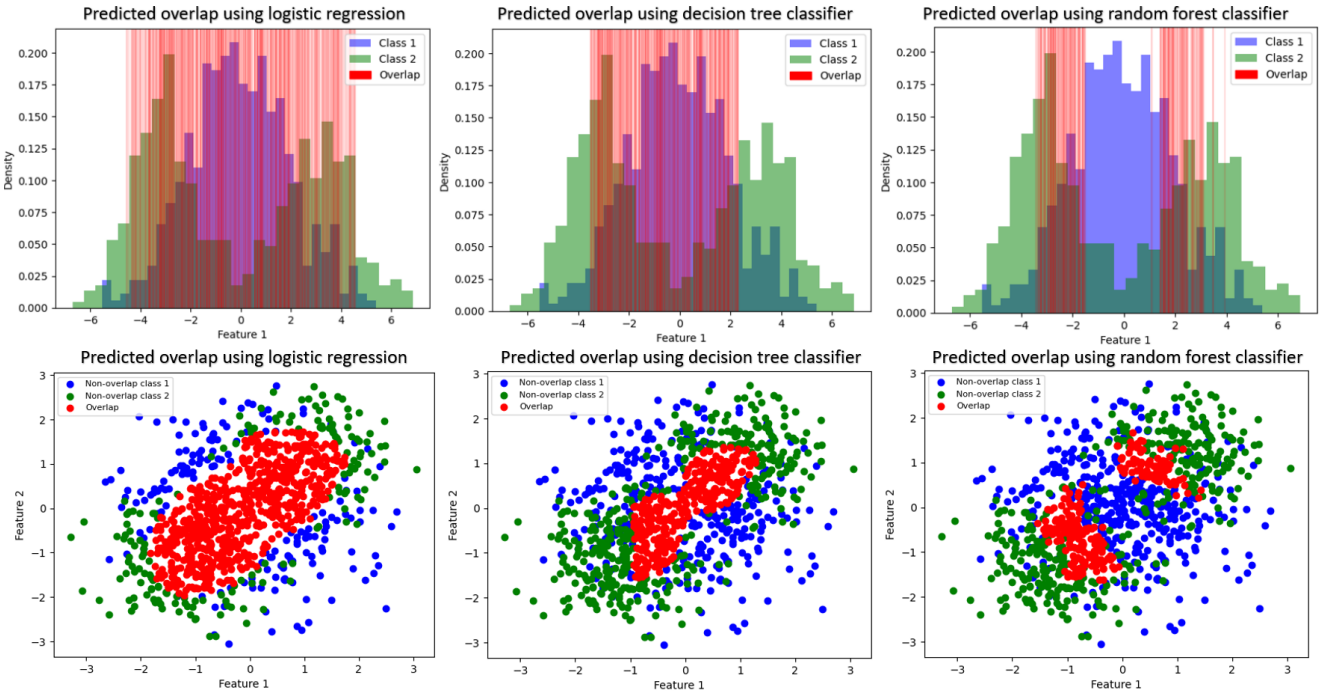

Figure 3: A closer look at the model identifying the overlap in 1D and 2D with two regions of overlap (30% of the data) between two classes. Different classifiers were used: logistic regression (left), decision tree classifier (middle), and random forest classifier (right).

expected to decrease. Not only the classifiers, but also the kernel density estimator that is used alongside the classifiers is affected by the "curse of dimensionality" [13]. Another cause for the decrease in performance could be that the propensity score represents all dimensions with a single value, which may start to falter with higher dimensionality.

It should also be noted that the performance of logistic regression drops quite a lot with higher dimensions, to the point where it crosses with the random forest classifier going to the sixth dimension. As such, one should be careful when using data with more than five features, since logistic regression might not be the optimal choice anymore.

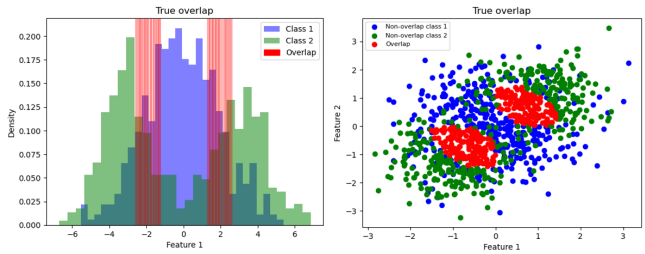

Figure 4: A closer look at the true overlap of a nonlinear problem when the data has 1 and 2 features (1D and 2D). One class is split into two normal distributions on either side of the other class's distribution. For the 500 samples used, the overlap region was kept at around 30%.

Two Regions of Overlap

When the problem is nonlinear, however, linear classifiers like logistic regression are usually less accurate than nonlinear classifiers. Though this model does use logistic regression weighted by a kernel density estimator, which makes the model not completely linear. As such, to find out how sensitive this model is when applied to nonlinear problems, for the second experiment there will be two regions of overlap in two different locations. One class is split into two normal distributions on either side of the other class's distribution. This means that the two classes are not linearly separable, making this a nonlinear problem. Figure 4 shows two examples of this. For this experiment, the region of overlap was also kept at approximately 30% of the full data across different dimensions by adjusting the means and standard deviations.

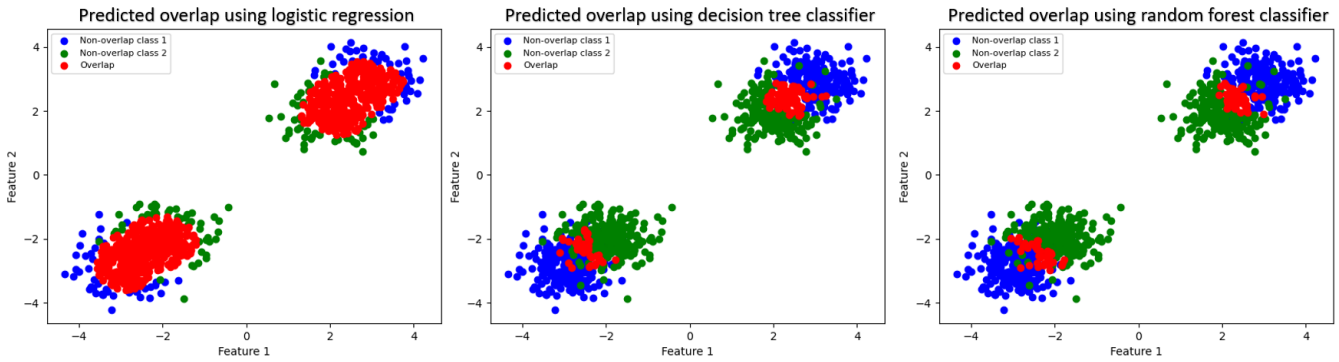

Figure 5: Different classifiers identifying the overlap in 2D when the classes have overlapping distributions on two separate locations.

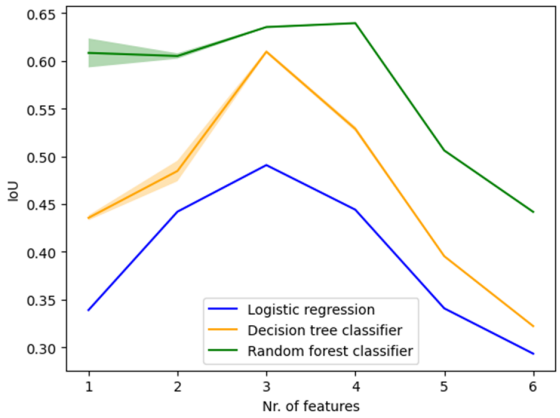

Figure 6: Performance using different classifiers with an increasing number of features. Two regions of overlap (30% of the data) between two classes. The uncertainty in the Figure is the variance of the model.

In Figure 6, as expected, logistic regression performs far worse than in the previous scenario. The nonlinear classifiers also perform worse than in the first experiment though. The random forest classifier seems to perform better than the other two in this scenario. Similarly to the first experiment, all classifiers perform worse with higher dimensionality.

Unlike the first experiment however, the performance of the model starts low and goes up with higher dimensions, before going back down again. Because of this, one set of predictions was plotted in Figure 3 (on the previous page) to take a closer look at this behaviour. The Figure contains the predictions for both one feature and two features (1D and 2D), since the performance jumped the most between these two dimensions. Plots of the true overlap regions can be seen in Figure 4, at the start of this section covering two regions of overlap.

Using logistic regression, the model seems to classify the two overlap regions, the area in between, and a bit around it as overlap for both 1D and 2D. However, the IoU is higher in 2D, with the False Positive (FP) rate lowering from 62% to 38%. One explanation could be that with more dimensions, there are fewer non-overlapping samples located between the two overlap regions.

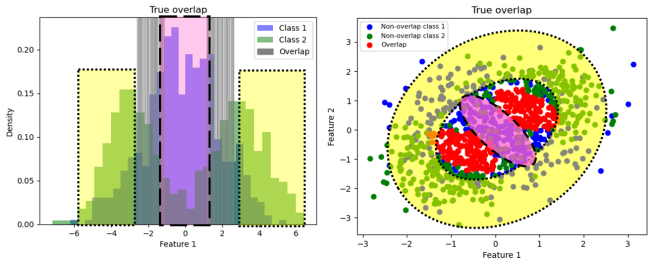

Figure 7: For both 1D to 2D, the general regions with non-overlapping samples are highlighted. The region between the two overlap regions is in dashed purple. The region outside the two overlap regions is in dotted yellow.

As Figure 7 shows, the region between the two overlap regions (dashed purple) gets smaller from 1D to 2D. Meanwhile, the region outside the two overlap regions (dotted yellow) gets larger. The model is most likely to get the samples wrong in the dashed purple area, as it tends to interpret the two overlap regions as one big region of overlap (the plots in the leftmost column of Figure 3 showcase this). As such, the region the model most likely gets wrong shrinks when going to a higher dimension, which makes the model perform better.

The middle column in Figure 3 shows that the decision tree classifier also primarily made mistakes regarding the samples between the two overlap regions. Perhaps the reason for the decision tree classifier gaining performance from 1D to 2D is similar to the reason for logistic regression, albeit less severe.

Unlike the other two classifiers, the random forest classifier did not seem to start low in performance. This may be due to the fact that random forests do not overfit [11]. The rightmost column in Figure 3 shows that it indeed did not falsely predict any overlap between the two overlap regions.

After the fourth dimension the performance does not seem to rise anymore, but instead drops again. This corresponds with the drop in performance seen in the first experiment, in Figure 2. The gain in performance becomes less and is likely negated by the loss in performance with higher dimensions.

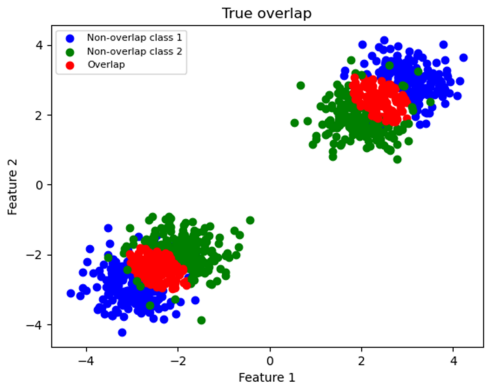

Figure 8: True overlap in 2D when the two classes (500 samples each) have overlapping distributions on two separate locations. Non-overlapping points of class 1 in blue and class 2 in green

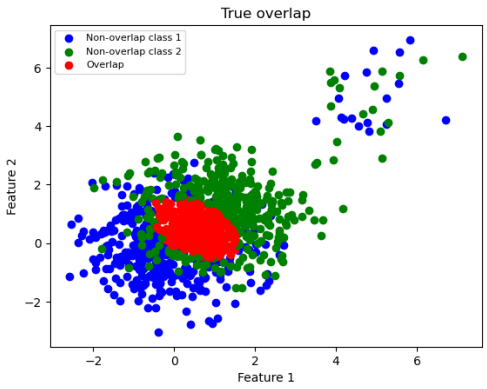

Figure 9: An example of how outliers are applied to the scenario depicted in Figure 1. Here 2.5% of the samples consist of outliers, located around the mean (5,5). The main distribution has one region of overlap (30% of the data) between two classes with 2 features (2 dimensions).

Separate Clumps with Overlap

After the second experiment, another specific scenario was tried out and plotted to observe the behaviour of the different classifiers when there are separate clumps of samples containing overlap. Here the two classes are each split into two Gaussians, resulting in two regions of overlap, as seen in Figure 8. For this example there are two features and the total amount of overlap is 30%. Figure 5 (on the previous page) shows that logistic regression severely overestimates the overlap region, classifying 82% of the samples to be in the overlap region. The nonlinear classifiers underestimate the overlap, as the decision tree and random forest classifier classify 15% and 11% of the samples as overlap.

Interestingly, even though each clump looks exactly the same as the distributions in the first experiment (Figure 1), logistic regression performs poorly when there are multiple cases of this scenario at the same time. This is due to it no longer being a linear problem, unlike the first experiment.

It is important to note that the IoU of the examples in Figure 5 ranged from 0.34 to 0.40. Even with drastically different predictions, the IoU was close between the three different classifiers. This indicates that it is hard to gather from the IoU alone how the model behaves precisely. One could use accuracy instead, which looks at the number of correct classifications. However, when using accuracy, a model that classifies all samples as non-overlapping would get a score of 0.70 in the case of 30% overlap, thus favouring underestimating models (in the case of 30% overlap or less). Aside from measuring the performance of the model, one could take a closer look at the FP and FN (False Positive and False Negative) rates to get a better indication of whether the model is overestimating or underestimating. As expected, in this example logistic regression had a very high FP rate of 54% and just an FN of 1%, while the nonlinear classifiers had somewhat high FN rates around 17% with their FP around 2%.

4.2 Outliers

In these experiments, the method will be tested with an increasing number of outliers.

For this experiment, the data will have two features, so it will be two-dimensional. The same distributions from Figure 1 will be used. For both classes, a percentage of the training samples will be in a smaller, identical distribution, but in a different location. While the main portion of the samples lies around mean (0,0) and (1,1), the outliers (samples of both classes) will be around mean (5,5). This may cause the model to identify overlap there as well. This is acceptable, as long as the main region of the predicted overlap is not affected. The expectation is that logistic regression will be affected the most, because with enough outlier points a smaller second region of overlap would appear for the classifier. And as seen before, logistic regression has difficulty with a nonlinear problem like that, since it is a linear classifier.

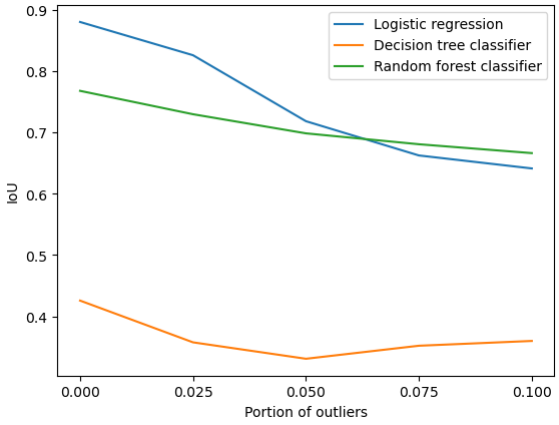

Figure 10: Performance using different classifiers with an increasing amount of the data consisting of outliers. Two regions of overlap (30% of the data) between two classes with 2 features (2 dimensions).

As Figure 10 shows, the performance drops with an increase in outliers up to 5% for all classifiers. As expected, logistic regression is affected the most, decreasing in performance by more than the other two. However, logistic regression still performs best up to 5%, but this is because the scenario of two Gaussians is favourable for logistic regression, as seen in the first experiment.

With more than 5% consisting of outliers, the performance of logistic regression seems to drop below that of the random forest classifier. Though, it should be noted that 10% of the samples being outliers might be a lot. With 2.5%, the decrease in performance is more comparable to the decrease in performance of the nonlinear classifiers. And up to 5% it still outperforms the other classifiers in this scenario.

5 Responsible Research

When conducting research, it is important to do so properly and responsibly. As a student, it is expected of me to internalize the Netherlands Code of Conduct with the guidance of my supervisors [1]. This section will focus on certain aspects of responsible research that are especially relevant to this project. As such, the possible future use cases, accessibility, and reproducibility of the code will be discussed in more detail.

This project is about overlap in causal inference. Causal inference is often used in the medical field, where the effects of certain treatments need to be assessed. This is a high-stakes field, where biased or incorrect results of the research conducted can have major consequences. As such, the programs and code used for this need to be reliable. We should be sure about the capabilities of our model and not overpromise on its performance.

The code and the data used in this project need to be accessible and reusable. To ensure this, the code used for this project will be available as a GitHub repository[1]. This enables others to revisit or replicate this project in the future. For the simulated data that is used in this project, data is randomly generated in Python. To make sure the experiments can be repeated, the seeds used for the experiments can be found in the repository as well.

Additionally, the open-source package implemented for this project is written in Python, a very interoperable language. It is flexible, widely used, and can interface with many other languages and tools. Importantly, Python programs can run on most personal computers using free software, making it very accessible.

6 Discussion

Following the experiments, the model performs best with one region of overlap between the two classes. In this scenario logistic regression works best, since it is a linear problem. However, the model drops in performance with multiple regions of overlap, even when using a nonlinear classifier. For these nonlinear problems, the random forest classifier does perform the best overall. Importantly, when there are multiple separate clumps of data, for instance in Figure 5, the model is also not accurate. This is also true even if each clump resembles the ideal scenario from the first experiment. Additionally, the experiments show that the model generally performs worse with higher dimensionality.

When it comes to outliers, logistic regression dropped the most in performance, while the nonlinear classifiers were affected slightly. Still, logistic regression dropped not as quickly in performance before 2.5% of the samples were outliers. For up to 5% of the samples being outliers, logistic regression still performed better than the nonlinear classifiers in the case of one region of overlap.

This project does have some limitations, mainly due to the limited time frame that was set for the project. The experiments that were performed aimed to cover the most important cases of interest: the scalability with the number of features and the number of outliers. However, it was not possible to test the model on many other different aspects. For example, exploring how having more than two classes affects the model's performance. Furthermore, only a few distributions and situations were tested in the experiments. This only gives an idea of how the model may behave in general, but it does not cover all cases that may occur. To test the model in a more realistic setting, real data could be used. Even though the true overlap regions would not be known, observing the model's behaviour and comparing it to other models could give more insight into how it would perform with real data.

7 Conclusions and Future Work

Identifying the region of overlap between two classes can be accomplished using propensity scoring with density estimation. This project explored how this method performs in different scenarios (one or multiple regions of overlap) and how well it scales with an increasing number of dimensions and outliers. Additionally, three classifiers were used for computing the propensity scores and compared to each other: logistic regression, decision trees, and random forests.

The results from the experiments show that the model performs best with only one region of overlap, dropping significantly in performance when it comes to multiple regions of overlap. For the former case, using logistic regression is preferred, while in the latter case, the random forest classifier seems to perform best. Furthermore, the model generally drops in performance with higher dimensionality. Regarding outliers, the model's performance drops slightly with up to 2.5% of the samples consisting of outliers. With more outliers, only logistic regression is significantly affected, while the nonlinear classifiers (decision trees and random forests) are not.

The experiments in this report only looked at certain scenarios using simulated data. Following these experiments, the behaviour of the model using real datasets could be explored further. Perhaps another avenue that could be explored is identifying overlap between more than two classes, using multiclass classification.

The results in this report give an intuition of when the method performs well and which classifiers should be used. These intuitions can be taken into account and explored fur-

[1] GitHub repository can be found at:
https://github.com/JonathanTjong/propensity_with_density

ther when using this method for identifying overlap for causal inference in the future.

References

[1] K Algra, L Bouter, A Hol, J van Kreveld, D Andriessen, C Bijleveld, R D'Alessandro, J Dankelman, and P Werkhoven. Netherlands code of conduct for research integrity 2018, 2018.

[2] Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

[3] Yoshua Bengio, Olivier Delalleau, and Clarence Simard. Decision trees do not generalize to new variations. *Computational Intelligence*, 26(4):449–467, 2010.

[4] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[5] Shifei Ding, Zhongzhi Shi, and Ahmad Taher Azar. Research and development of advanced computing technologies, 2015.

[6] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[7] Joseph Kang, Wendy Chan, Mi-Ok Kim, and Peter M Steiner. Practice of causal inference with the propensity of being zero or one: assessing the effect of arbitrary cutoffs of propensity scores. *Communications for statistical applications and methods*, 23(1):1, 2016.

[8] Michael Oberst, Fredrik Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush Varshney. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR, 2020.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[11] Steven J Rigatti. Random forest. *Journal of Insurance Medicine*, 47(1):31–39, 2017.

[12] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[13] David W Scott. Feasibility of multivariate density estimates. *Biometrika*, 78(1):197–205, 1991.

[14] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[15] Arane Thavaneswaran and Lisa Lix. Propensity score matching in observational studies. *Canada: University of Manitoba*, 2008.

[16] Mikhail Traskin and Dylan S Small. Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences*, 3:94–118, 2011.

[17] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[18] Vladimir Vovk. The fundamental nature of the log loss function. *Fields of Logic and Computation II: Essays Dedicated to Yuri Gurevich on the Occasion of His 75th Birthday*, pages 307–318, 2015.

[19] Sherry Weitzen, Kate L Lapane, Alicia Y Toledano, Anne L Hume, and Vincent Mor. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and drug safety*, 13(12):841–853, 2004.