Information Theoretic Deductions Using Machine Learning with an Application in Sociology

Arunselvan Ramaswamy[1][a], Yunpeng Ma[1], Stefan Alfredsson[1][b], Fran Collyer[2][c] and Anna Brunström[1][d]

[1] *Dept. of Mathematics and Computer Science, Karlstad University, Sweden*

[2] *School of Humanities and Social Inquiry, University of Wollongong, Australia*

Keywords: Conditional Entropy, Binary Classification, Information Theory, Supervised Machine Learning, Automated Data Mining, Machine Learning in Sociology.

Abstract: Conditional entropy is an important concept that naturally arises in fields such as finance, sociology, and intelligent decision making when solving problems involving statistical inferences. Formally speaking, given two random variables X and Y, one is interested in the amount and direction of information flow between X and Y. It helps to draw conclusions about Y while only observing X. Conditional entropy $H(Y|X)$ quantifies the amount of information flow from X to Y. In practice, calculating $H(Y|X)$ exactly is infeasible. Current estimation methods are complex and suffer from estimation bias issues. In this paper, we present a simple Machine Learning based estimation method. Our method can be used to estimate $H(Y|X)$ for discrete X and bi-valued Y. Given X and Y observations, we first construct a natural binary classification training dataset. We then train a supervised learning algorithm on this dataset, and use its prediction accuracy to estimate $H(Y|X)$. We also present a simple condition on the prediction accuracy to determine if there is information flow from X to Y. We support our ideas using formal arguments and through an experiment involving a gender-bias study using a part of the employee database of Karlstad University, Sweden.

1 INTRODUCTION

Conditional entropy is an information theoretic term that quantifies the amount of information required to describe one random variable, given that the value of another random variable is known. It naturally arises in many scenarios and calculating its value is often very useful. Consider the following – given the global financial trend, e.g., Dow Jones Industrial Average or S&P 500 index, it is interesting to predict trends in a specific Company, say company $\mathcal{C}$. Such an analysis would indicate if $\mathcal{C}$ is a market sensitive or a market leading company. Also, the degree and direction of *information transfer* between various stock market indices is important (Dimpfl and Peter, 2013) (Darbellay and Wuertz, 2000). This is often studied using a quantity related to conditional entropy called the *transfer entropy*.

Let us suppose that random variable X represents the evolution of the global stock market, and that random variable Y represents the evolution of company $\mathcal{C}$. The conditional entropy under question is

$$H(Y|X) = -\sum_{x \in \mathcal{X}} p_x \sum_{y \in \mathcal{Y}} p_{y|x} \log p_{y|x}, \tag{1}$$

where $p_x = P(X = x)$, $p_{y|x} = P(Y = y|X = x)$, $\mathcal{X}$ and $\mathcal{Y}$ are the supports of random variables X and Y, respectively. For the sake of simplicity we have assumed that X and Y are discrete valued (and even countable). Suppose, they are continuous then the summations in (1) are replaced by integrals. $H(Y|X)$ quantifies the amount of information needed to describe Y, given that X is known. It can be shown that $0 \leq H(Y|X) \leq H(Y)$, where $H(Y) = -\sum_{y \in \mathcal{Y}} p_y \log p_y$ is the entropy of Y, the amount of uncertainty in Y. If $H(Y|X) = 0$, then Y can be fully determined by X, e.g., when $Y = X^2$. X and Y are independent *iff* $H(Y|X) = H(Y)$ (MacKay et al., 2003). Finally, note that conditional entropy is an asymmetric quantity –

[a] https://orcid.org/0000-0001-7547-8111
[b] https://orcid.org/0000-0003-0368-9221
[c] https://orcid.org/0000-0002-9877-6645
[d] https://orcid.org/0000-0001-7311-9334

Ramaswamy, A., Ma, Y., Alfredsson, S., Collyer, F. and Brunström, A.
Information Theoretic Deductions Using Machine Learning with an Application in Sociology.
DOI: 10.5220/0012371600003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 320-328
ISBN: 978-989-758-684-2; ISSN: 2184-4313

$H(Y|X)$ only considers the information flow from X to Y. Let us circle back to the stock market scenario. If $H(Y|X) = 0$, then company $\mathcal{C}$ follows the global trend to the highest degree. The value of $H(Y|X)$ is inversely proportional to the degree of dependency of $\mathcal{C}$ on the global trend.

In practice, calculating the conditional entropy is infeasible as it requires complete knowledge of the joint and marginal probability distributions of X and Y. Hence, it is often estimated using X and Y observations (Pham, 2004). Such data-driven estimators are often sophisticated, while suffering from significant estimation biases (Beirlant et al., 1997).

In the field of computer security, a side channel attack is an attack that is based on extra information that can be gathered owing to the fundamental way in which a computer algorithm is implemented (Golder et al., 2019). In order to prevent side-channel attacks, security analysts must ascertain the amount of data, which when collected, can be used to compromise a computer. Let us define X to be the random variable associated with the gathered data – the "side-channel information". Let us also define the two valued random variable Y as equalling 1 if the security is compromised, and 0 otherwise. Then $H(Y|X)$ quantifies how much of X must be observed in order to accurately determine Y. Again, calculating $H(Y|X)$ exactly is infeasible. Machine Learning was recently used to skirt around this (Drees et al., 2021) (Gupta et al., 2022). However, these works are preliminary and empirical without formal backing. In any case they do not attempt to estimate conditional entropy or draw information theoretic conclusions.

Our Contributions. We present a Machine Learning based easy-to-implement method to estimate the conditional entropy $H(Y|X)$, where Y is a bi-variate random variable and X is discrete valued. This estimate is used to measure the degree of dependency of Y on X. Given observations of X and Y, we discuss how to transform the problem of estimating the conditional entropy to a supervised learning problem. We then use the prediction accuracy of the supervised learning algorithm to estimate the conditional entropy. We present sufficient conditions on the accuracy of the learning algorithm for guaranteed information flow from X to Y. We support our ideas through formal arguments and experiments on real datasets.

1.1 Conditional Entropy in Sociology: A Data-Driven Approach

Companies and organizations aim to ensure their male and female employees are equally represented, that their policies are not skewed towards one gender. Conditional entropy plays an essential role when analyzing the current state of affairs with respect to gender equality in these organizations. Here, we present an ML-based approach for such an analysis. First, the employee database is used to create a supervised learning (classification) dataset. There is one input instance corresponding to each employee. It only contains gender neutral information such as date of birth, salary, position, etc. Gender revealing information such as names, gender, etc., are excluded. The input instances are labeled using their gender – 0 for female and 1 for male. We are therefore in the setting of binary classification. We use X to represent the random variable associated with the input, and Y to represent the class random variable.

If the gender does not play a role in hiring and subsequent career development, then one cannot reliably predict the gender Y merely using the gender neutral information X, such as the title and salaray. In the parlance of information theory, $H(Y|X) = H(Y)$. On the other hand, suppose there is *gender bias*, then $H(Y|X) < H(Y)$. *In this paper, we call a dataset gender biased when the genders are unequally represented (gender plays a role in hiring, career development, etc.) when* $H(Y|X) < H(Y)$. As explained earlier, we present a supervised learning approach to estimating $H(Y|X)$, and checking if $H(Y|X) < H(Y)$.

The gender inequality problem is very well studied in literature, see, e.g., (Heiberger, 2022). Also, ML tools have been been previously used to empirically study problems in sociology, e.g., (Zajko, 2022) (Molina and Garip, 2019). To the best of our knowledge, this is the first time in literature, wherein ML is used within the framework of statistical inference to answer a sociological question, in particular a gender-bias question. Further, the framework is backed by formal theory.

Here is another scenario where our methodology is useful. Let us suppose that a vast region is flooded. After emergency relief operations, the responsible government committee must formulate a plan to allocate resources and funds for the long-term rebuilding process. For the sake of simplicity, let us suppose that the region can be divided into neighborhoods – each consisting of groups of houses, one or more community centers, commercial buildings, schools and hospitals. The committee must allocate resources at the "neighborhood level". Resource allocation must be done in a fair and equitable manner, proportional to the losses incurred by the neighborhood. Further, the associated timelines, e.g., to release funds, must facilitate immediate and equitable relief. The plan must not be influenced by the political orientation, racial identity or affluence of any given neighborhood.

Given a proposal, how does one go about checking if the plan satisfies the above mentioned equity constraints? Traditionally, a human expert evaluates the plan, the draft is then amended based on the feedback and sent for reevaluation, this process is repeated a few times. The framework presented in this paper can hasten the feedback loop through the introduction of an automated bias checking routine. For plan evaluation, we first create a dataset containing datapoints with the following information on the neighborhoods: general information about a neighborhood (e.g., number of houses, schools, average household income, etc.), information on flood damage (e.g., fraction of the neighborhood affected, loss to critical services, etc.), and information on flood relief (e.g., allocated resources and funds, timeline, etc.). The data should *not* include information that must be disregarded during planning – neighborhood-racial-identity, affluence, etc. *There is one datapoint corresponding to every neighborhood in the flooded region.*

A Machine Learning (ML) algorithm is then trained on the dataset to predict, e.g., the majority political orientation of neighborhoods. If the prediction accuracy is high enough, then our analysis shows that the draft is biased. *Put simply, the predictor should not perform better than guessing, when predicting the neighborhood-political-orientation for a draft to be deemed unbiased.* From an information theoretic perspective, a high accuracy indicates that there is some information regarding political orientation present in the data. Note that the bias may be positive or negative. On the other hand, suppose the prediction accuracy is low (accuracy corresponding to guessing), then the draft is probably fair. Although the ML based bias predictor can provide as assessment within a matter of hours, if not minutes, *it does not have the capacity to assess the cause for the bias*. The draft may be passed to an expert for feedback only if the ML predictor detects a bias, thus saving valuable time and effort.

2 FORMAL PROBLEM SETUP AND OUR APPROACH

Today, enormous data on individuals, anonymized and otherwise, is readily available in the public domain. Decision making bodies utilize this data, analyze them, and base important decisions on conclusions based on these analyses. Social biases, e.g., gender and race, intrinsic to the data, affect the decision making process. Information theoretic quantities like conditional entropy play a pertinent role in quantifying the biases in the aforementioned databases. However, these quantities are hard to estimate. In this paper, we propose a supervised learning based framework to solve the problem of estimating the conditional entropy of information extracted from databases. We propose a novel framework based on ML and Statistics. We stick to the parlance of gender bias to illustrate our ideas. It must however be noted that they can be readily extended to study other types of biases and information flows.

Machine Learning (ML) is the study of algorithms that self-improve at performing tasks only through data (Bishop and Nasrabadi, 2006) (Goodfellow et al., 2016) (Hastie et al., 2009). Supervised learning is an important ML paradigm where the algorithm learns to perform tasks by emulating examples. Mathematically speaking, a supervised learning algorithm tries to learn an unknown map $Y : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the input and output spaces respectively, using example data called training data, represented by the set $\mathcal{D} = \{(x_n, Y(x_n)) \mid x_n \in \mathcal{X}\}_{1 \le n \le N}$ for some $1 \le N < \infty$. A supervised learning algorithm $\mathcal{A}$ learns a map $Y^{\mathcal{A}}$, using $\mathcal{D}$, that approximates the unknown map Y. Specifically, it finds $Y^{\mathcal{A}}$ that minimizes prediction errors on $\mathcal{D}$.

We are given X and Y observations $\{(x_n, y_n)\}_{1 \le n \le N < \infty}$, where X is discrete-valued and Y is bi-valued. This observation set naturally translates to the training dataset for a binary classification algorithm $\mathcal{A}$. Training $\mathcal{A}$ yields a proxy map $Y^{\mathcal{A}}(X)$ and an associated prediction accuracy. It is then used to (a) estimate the required conditional entropy $H(Y|X)$, and (b) assert whether $H(Y|X) < H(Y)$ in order to comment on the dependency of Y on X.

In the context of a gender bias study, X represents all perceived gender neutral information and Y the gender information, 0 for female and 1 for male. The observation data/training data is typically obtained from a database - the starting point for our analysis. The prediction accuracy is used to decide if there is gender information contained in X, if the database is *gender biased.* Intuitively speaking, we draw this conclusion based on whether the prediction accuracy is significantly better than guessing. If so, then $\mathcal{A}$ is able to exploit a possibly hidden pattern in the data with respect to gender. A human expert may now be summoned to carefully check the data in order to find the source of the perceived gender bias. Finally, we also comment on the degree of bias by estimating the required conditional entropy. It must be noted that the exact classification algorithm used in the analysis depends on the nature of the data X. Typically, one works with an ensemble of algorithms.

Our Approach. Let us suppose that we are given a database containing datapoints on people. Additionally, we are given the gender information for each datapoint. From this we construct a classification dataset $\mathcal{D} = \{(x_n, y_n) \mid 1 \leq n < \infty\}$, where x_n is the supposed gender neutral information regarding the n^{th} individual, extracted from the database, and y_n is the corresponding gender. We let $y_n = 0$ for female and $y_n = 1$ for male. Let X be the random variable associated with the perceived gender neutral information, and Y be the gender random variable. Define,

$$p := \frac{\text{\# of males in the database} \vee \text{\# of females in the database}}{|\mathcal{D}|}, \quad (2)$$

where $\vee$ denotes the max operator. We train an ensemble $\mathcal{E}$ of binary classification algorithms using $\mathcal{D}$. Let us suppose that there is at least one algorithm $\mathcal{A} \in \mathcal{E}$ with an accuracy that is strictly greater than p. Then, we argue that $\mathcal{D}$ is gender biased. In particular, we show that the conditional entropy $H(Y \mid X) < H(Y)$. Recall that $H(Y \mid X)$ represents the amount of information required to describe the outcome of Y given that the value of X is known. It is known that $H(Y \mid X)$ is at most $H(Y)$, i.e., $H(Y \mid X) \leq H(Y)$. Suppose X does not contain any information about Y, X and Y are independent random variables, then $H(Y \mid X) = H(Y)$. Hence, $H(Y \mid X) < H(X)$ indicates that X contains gender information, it can describe the outcome of Y.

On the other hand, suppose that none of the classifiers in the ensemble have an accuracy better than p, then it is very likely that $H(Y \mid X) = H(Y)$. From a theoretical standpoint, the Baye's classifier is the optimal classification algorithm, in that it has the highest accuracy. We may replace our ensemble with the Baye's classifier in order to be sure. The only issue is that the Baye's classifier is computationally infeasible for most classification problems. To summarize our approach, we create a classification dataset from the database entries and solve the gender classification problem. We show that high accuracy, specifically better than guessing, is indicative of a gender bias in the dataset.

3 INFORMATION THEORETIC BACKING FOR GENDER BIAS DETECTION USING CLASSIFICATION

In the parlance of machine learning, X is called the feature random vector and Y is called the class random variable. In order to detect gender bias in $\mathcal{D}$, one typically estimates the mutual information $I(X,Y)$, which is related to conditional entropy through the following formula:

$$I(X,Y) = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y) \quad (3)$$

where $H(X)$ and $H(Y)$ are the entropies of X and Y respectively, $H(X \mid Y)$ and $H(Y \mid X)$ are the conditional entropies of X given Y and Y given X, respectively. The mutual information $I(X,Y)$ quantifies the inference that can be drawn regarding one random variable by observing the other. Mutual information is symmetric – $I(X,Y) = I(Y,X)$ – and always positive – $I(X,Y) \geq 0$.

When $I(X,Y) = 0$, no information can be drawn regarding Y by observing X. Due to the symmetric nature of mutual information a similar statement can be made by swapping X and Y in the previous sentence. We are, however, only interested in the former and we will stick to it. From (3) it therefore follows that $H(Y) = H(Y \mid X)$. Simply put, observing X does not reduce the uncertainty in Y. On the other hand, when $I(X,Y) > 0$, information regarding Y can be inferred by observing X. The higher the value, the better the inference. Again from (3), we get the equivalent condition that $H(Y) > H(Y \mid X)$. Colloquially speaking, observing X has reduced the uncertainty in Y. Suppose $H(Y \mid X) = 0$, then $I(X,Y)$ takes the maximum possible value of $H(Y)$, and Y is fully determined by observing X.

Let us circle back to the problem at hand. We are interested in quantifying the gender inference (describing Y) through only observing X (the perceived gender neutral information regarding an individual extracted from the database). If we show that $I(X,Y) = 0$, or equivalently that $H(Y \mid X) = H(Y)$, then gender inference is not possible, and the *database is not gender biased*. If, on the other hand, we show that $I(X,Y) > 0$, or equivalently that $H(Y \mid X) < H(Y)$, then the gender can be fully or partially inferred, and the *database is gender biased*. Hence, the first order of business is to estimate these entropy quantifiers.

Instead of using complex biased entropy estimators from literature, we propose solving an associated classification problem. We propose predicting the gender of individuals Y solely using X. This yields a natural binary classification problem, and a natural training dataset $\mathcal{D}$. We show that using $\mathcal{D}$ to effectively train a classifier (Machine Learning model) to predict the gender with high accuracy for any individual from the database is an indicator of gender bias.

Our Contribution. *We show $H(Y \mid X) < H(Y)$ when a binary classifier, trained on $\mathcal{D}$, has an accuracy that is strictly greater than p (recall the definition of p*

from (2)*). In effect, showing that an accuracy strictly greater than p implies that the database is gender biased. We also analyze the scenario when the classifier used is optimal – Bayes classifier. In the course of this analysis, we estimate $H(Y \mid X)$ and $H(Y)$.*

Baye's classifier is the theoretical optimal, in that, it has the highest accuracy among all classifiers trained on $\mathcal{D}$. However, it requires full knowledge of the underlying population distribution that generated the data in the database, and consequently $\mathcal{D}$ itself. As there is no access to the population distribution, we resort to other feasible albeit suboptimal classifiers that only require $\mathcal{D}$. The dataset $\mathcal{D}$ is assumed to be generated from the joint unknown distribution on (X,Y) - also known as the population distribution - through repeated and independent sampling.

3.1 Entropy of a Gender Guessing Algorithm

First, we define the following notations: $p_x := P(X = x)$, $p_y := P(Y = y)$ and $p_{y|x} := P(Y = y \mid X = x)$. These together, define the required population distribution. Without loss of generality, we assume that the database contains at least as many male representatives as female ones. Since we assumed that $\mathcal{D}$ is generated by the underlying population distribution, provided it is also large, we have $P(Y = 1) \approx p$ and $P(Y = 0) \approx 1-p$, where p is defined in (2). Now, consider a randomized classifier $\mathcal{G}$ (gender guessing algorithm) that predicts with probability p that a given query instance x belongs to class-1, and predicts class-0 with probability $1-p$. In effect, $\mathcal{G}$ does not truly consider x when predicting gender. If we associate the random variable $Y^{\mathcal{G}}(x)$ with the prediction of $\mathcal{G}$, then $Y^{\mathcal{G}}(x) = 1$ with probability p and $Y^{\mathcal{G}}(x) = 0$ with probability $1-p$. Note that x is an instance/realization of X.

Let $p^{\mathcal{G}}_{y|x} := P(Y^{\mathcal{G}}(x) = y \mid X = x)$. Then, the conditional entropy,

$$H(Y^{\mathcal{G}}(X)|X) = -\sum_{x\in\mathcal{X}} p_x \sum_{y\in\mathcal{Y}} p^{\mathcal{G}}_{y|x} \log p^{\mathcal{G}}_{y|x},$$

where $\mathcal{X}$ is the support of X – the set of all possible values of X – and $\mathcal{Y} = \{0,1\}$. Since $\mathcal{G}$ does not truly consider X during prediction, we have that $p^{\mathcal{G}}_{y|x} = p_y$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and the above equation becomes

$$H(Y^{\mathcal{G}}(X)|X) = -\sum_{x\in\mathcal{X}} p_x \sum_{y\in\mathcal{Y}} p_y \log p_y = H(Y).$$

This is not surprising since $\mathcal{G}$ has not used X for prediction. Suppose $\mathcal{D}$ is *not* gender biased, then $\mathcal{G}$ is the optimal predictor. In other words, where one cannot do better than guessing, and X is useless in predicting the gender.

3.2 Baye's Classifier $\mathcal{B}$ when $\mathcal{D}$ Is not Gender Biased

One sufficient condition for $\mathcal{D}$ to be gender unbiased is when X and Y are uncorrelated, when $p_y = p_{y|x}$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The Baye's classifier is given by the following deterministic mapping from $\mathcal{X}$ onto $\mathcal{Y}$:

$$x \mapsto \operatorname*{argmin}_{y\in\mathcal{Y}} \ell(0,y)p_{0|x} + \ell(1,y)p_{1|x}, \quad (4)$$

where $\ell(y_1,y_2) = 1$ when $y_1 \neq y_2$ and 0 otherwise, $y_1, y_2 \in \mathcal{Y}$. ℓ is called the 0-1 loss function, it penalizes a wrong prediction by 1. Baye's predictor is a mapping that minimizes the average loss, where the average is taken with respect to the population distribution. When using the 0-1 loss function, minimizing the average loss is equivalent to maximizing the average accuracy, again the average is taken with respect to the population distribution. One can use (4) to show that the Baye's classifier $\mathcal{B}$ predicts 0 at x when $p_{1|x} < p_{0|x}$, and 1 otherwise. Hence, (4) is equivalent to

$$x \mapsto \operatorname*{argmax}_{y\in\mathcal{Y}} p_{y|x}.$$

Since we assumed that X and Y are uncorrelated, $\mathcal{B}$ reduces to the constant map

$$x \mapsto \operatorname*{argmax}_{y\in\mathcal{Y}} p_y.$$

We assumed without loss of generality that $p_1 \geq p_0$, hence $\mathcal{B}$ predicts that every query instance x belongs to class-1. The Bayes classifier $\mathcal{B}$ becomes the majority classifier, it classifies every query instance as the majority class – class-1 in our case. $\mathcal{B}$ has an accuracy of p $(= p_1)$, since it is correct only on query instances that are male. The associated prediction random variable $Y^{\mathcal{B}}(x) = 1$ with probability 1 for all $x \in \mathcal{X}$. Further, $P(Y^{\mathcal{B}}(x)$ is the correct label for $x) = p$. Define a random variable $Z(X) = \mathbb{1}(Y^{\mathcal{B}}(X)$ is the correct label for $X)$, then $\mathbb{E}Z(X) = p$. Note that $\mathbb{1}$ is the indicator random variable whose outcome is always 0 or 1. It is 1 only when $Y^{\mathcal{B}}(x)$ is the correct label for x and 0 otherwise. $Z(X)$ is fully determined by X, $Z(X) = 1$ when X belongs to class-1 (male) and $Z(X) = 0$ otherwise. Therefore, $Z(X)$ is 1 with probability p $(= p_1)$ and 0 with probability $1-p$ $(=p_0)$. Its entropy is given by

$$\begin{aligned} H(Z(X)) &= -\sum_{z\in\{0,1\}} P(Z(X)=z)\log P(Z(X)=z) \\ &= -(p\log p+(1-p)\log(1-p)) \\ &= -(p_1\log p_1+p_0\log p_0)=H(Y). \end{aligned} \tag{5}$$

The uncertainty associated with predicting the gender correctly is equal to the uncertainty in the gender itself. Since $Z(X)$ is fully determined by X, it is clear that the knowledge of X has not reduced the uncertainty in gender prediction. Therefore, we can conclude that $\mathcal{B}$ has not extracted any gender information from X, *spurious or otherwise*.

3.3 Entropy of a Classification Algorithm $\mathcal{A}$

Intuitively speaking, the database is gender biased if the gender of an individual can be determined with high accuracy, using only the perceived gender neutral information that can be extracted from the database. The higher the bias, the better the gender determination, the easier the gender prediction. Let us suppose that we have a classification algorithm $\mathcal{A}$ that is trained on $\mathcal{D}$ to predict the gender variable Y. Let $Y^{\mathcal{A}}(X)$ be the prediction random variable obtained by training $\mathcal{A}$ on $\mathcal{D}$. It is the best approximation of Y found by $\mathcal{A}$ using $\mathcal{D}$. Define $p(\mathcal{A},x)) := P(Y^{\mathcal{A}}(X)$ predicts the class of X correctly$|X = x)$. Suppose the accuracy is $> p$, then $p(\mathcal{A},x)) > p$. Note that the accuracy of a classifier is approximately the probability that a given query instance is classified correctly.

The binary entropy function $H_b(p) = -[p\log p + (1-p)\log(1-p)]$ for $0 \le p \le 1$. It monotonically increases with p for $0 \le p \le 1/2$, and it monotonically decreases as p increases from 1/2 to 1. Since $p(\mathcal{A},x)) > p \ge 1/2$, $H_b(p) > H_b(p(\mathcal{A},x)))$. Let $Y^t(X)$ represent the true label of X and $Y^f(X)$ its false label. We are now ready to calculate the conditional entropy:

$$\begin{aligned} &H(Y^{\mathcal{A}}(X)|X) \\ &= -\sum_{x\in X} p_x \sum_{y\in\mathcal{Y}} P(Y^{\mathcal{A}}(X)=y|X=x) \\ &\qquad\qquad \log P(Y^{\mathcal{A}}(X)=y|X=x), \\ &= -\sum_{x\in X} p_x \sum_{j\in\{t,f\}} P(Y^{\mathcal{A}}(X)=Y^j(X)|X=x) \\ &\qquad\qquad \log P(Y^{\mathcal{A}}(X)=Y^j(X)|X=x), \\ &= \sum_{x\in X} p_x H_b(p(\mathcal{A},x))). \end{aligned} \tag{6}$$

Since $Y^{\mathcal{A}}(X)$ is really a proxy for the gender variable Y, (6) *is an estimate for $H(Y)$.* We have previously discussed that $H_b(p(\mathcal{A},x))) < H(Y)$ for all $x \in \mathcal{X}$. Using this in (6), we get that

$$H(Y^{\mathcal{A}}(X)|X) < H(Y). \tag{7}$$

Suppose the trained accuracy of classifier $\mathcal{A}$ is strictly greater than p, then we have shown that the conditional entropy of the associated gender prediction random variable $Y^{\mathcal{A}}(X)$ is strictly less than the intrinsic entropy of the gender random variable Y. This indicates that $\mathcal{A}$ was able to exploit some gender information that is present inside X. Hence, the database is gender biased and *we were wrong in expecting X to be gender neutral.*

Being optimal, the Baye's classifier $\mathcal{B}$ has a better accuracy than $\mathcal{A}$. As stated earlier, the only issue is that it is incomputable in practical scenarios. We may nevertheless talk about the conditional entropy $H(Y^{\mathcal{B}}(X)|X)$, where $Y^{\mathcal{B}}(X)$ is the prediction random variable associated with $\mathcal{B}$. Similar to $\mathcal{A}$, define $p(\mathcal{B},x)) := P(Y^{\mathcal{B}}(X)$ predicts the class of X correctly$|X = x)$. Then, $p(\mathcal{B},x)) \ge p(\mathcal{A},x)) > p$ for $x \in \mathcal{X}$. As in the case of $\mathcal{A}$, we can show that $H(Y^{\mathcal{B}}(X)|X) < H(Y)$. Further, since $H(Y) = H(Y^{\mathcal{G}}(X)|X)$, we have that

$$H(Y^{\mathcal{B}}(X)|X) < H(Y^{\mathcal{G}}(X)|X). \tag{8}$$

When the accuracy of the Baye's predictor is $> p$, the associated conditional entropy is strictly less than the "guessing entropy". The Baye's predictor is able to perform better than guessing, as it is able to exploit, possibly hidden, gender information that is present in X.

3.4 Bayes Predictor Has Accuracy of p

We previously discussed that the majority classifier – which classifies every query instance as belonging to class-1 – has an accuracy of p, see (2) for the definition of p. This is because a majority classifier is correct only on male query instances and the fraction of males is p, see Section 3.2 for details. Since we have the freedom to choose $\mathcal{A}$ its accuracy can never be $< p$. If it has a lower accuracy then we can choose the majority classifier as $\mathcal{A}$. In the previous section, we analyzed the case where the accuracy of $\mathcal{A}$ is strictly greater than p. In this section, we analyze the only non-trivial case left – when the accuracy of $\mathcal{A}$ equals p.

Since the accuracy of $\mathcal{A}$ depends on the chosen algorithm, for the sake of analysis, let us suppose that $\mathcal{A}$ is chosen to be the Baye's classifier $\mathcal{B}$. Therefore, we are in the scenario where $\mathcal{B}$ has an accuracy of p.

Define $\overline{Y}^{\mathcal{B}}(X) := \mathbb{1}(Y^{\mathcal{B}}(X)$ is the correct class of $X)$, the random variable associated with the correct prediction of the Baye's classifier.

$$H(\overline{Y}^{\mathcal{B}}(X)|X) = -\sum_{x\in X} p_x \sum_{z\in\{0,1\}} P\left(\overline{Y}^{\mathcal{B}}(X) = z|X = x\right) \log P\left(\overline{Y}^{\mathcal{B}}(X) = z|X = x\right). \quad (9)$$

Since $P\left(\overline{Y}^{\mathcal{B}}(X) = z|X = x\right) = p$ for $x \in \mathcal{X}$, the RHS of (9) equals $\sum_{x\in X} p_x H(Y)$, which in turn equals $H(Y)$.

Now that we have shown $H(\overline{Y}^{\mathcal{B}}(X)|X) = H(Y)$, we see that the uncertainty associated with a correct prediction by $\mathcal{B}$ is equal to the intrinsic uncertainty in gender. The Bayes classifier $\mathcal{B}$, that tries to find gender patterns in X, is only as good as the majority classifier that does not consider X when predicting gender. Since $\mathcal{B}$ is the optimal classifier, we can conclude that that there is no gender information in X, at least none that is useful.

3.5 Degree of Bias

In the supervised learning paradigm of Machine Learning, one attempts to learn an unknown map from the given input space $\mathcal{X}$ to the output space $\mathcal{Y}$. To do this, training data – $\{(x, Y(x)) \mid x \in \mathcal{X},\ Y(x)$ is the label of $x\}$ – is used. Y is the said unknown map or the unknown output random variable. $\mathcal{A}$ uses training data to learn $Y^{\mathcal{A}}$ – an approximation of Y. Let us suppose that Y is a random map that is independent of X, Y does not truly consider X when labelling it. From an information theoretic perspective this means that $I(X,Y) = 0$. Further, since $Y^{\mathcal{A}}$ is a proxy for Y, we expect $I(X, Y^{\mathcal{A}}) \approx 0$.

$I(X,Y) = H(Y) - H(Y \mid X)$ and $\approx H(Y) - H(Y^{\mathcal{A}} \mid X)$ (as $Y^{\mathcal{A}}$ is a proxy for Y). When $I(X,Y) = 0$ we may conclude that X and Y are independent. They are dependent when $I(X,Y) > 0$, with the mutual information value quantifying the degree of dependency. In the parlance of information theory, there is information about Y in X, and vice versa, knowing X allows one to predict Y to a degree that is proportional to the aforementioned value. In Section 3.3, we discussed conditions under which $H(Y) > H(Y^{\mathcal{A}} \mid X)$, and since $I(X,Y) \approx H(Y) - H(Y^{\mathcal{A}} \mid X)$, we get that $I(X,Y) > 0$. Further, $H(Y) - H(Y^{\mathcal{A}} \mid X)$ is a good approximation of how well one can predict the gender Y using only X.

Suppose that we are given two different learning problems with training data $\mathcal{D}_1$ and $\mathcal{D}_2$. (X_1, Y_1) is the input-output random variables pair for the first problem and (X_2, Y_2) is the one for the second learning problem. Say that $\mathcal{A}$ is trained on $\mathcal{D}_1$ and $\mathcal{D}_2$, yielding approximations $Y^{\mathcal{A}}(X_1)$ and $Y^{\mathcal{A}}(X_2)$, respectively. For the sake of argument, say $H(Y_1) - H(Y^{\mathcal{A}}(X_1) \mid X_1) < H(Y_2) - H(Y^{\mathcal{A}}(X_2) \mid X_2)$. We can conclude that $\mathcal{A}$ has discovered a stronger dependency between X_2 and Y_2, as compared to the dependency between X_1 and Y_1.

4 EXPERIMENTAL STUDIES

In this section, we use our framework to conduct a "gender bias analysis" on an employee database of Karlstad University The database contains employee records of the teaching and research staff at the University, from a few departments. There are 294 entries, each containing 9 attributes (features) including name, gender, teaching and research activities, department, and salary. Out of the 294 entries, 211 are males and 83 are females. Using the definition of p in (2), for the University database $p = \frac{211}{294} \approx 0.72$.

4.1 Dataset Preparation

We need to first prepare the dataset corresponding to the associated classification task, then train a classification algorithm using it. Each data entry is divided into two components – X and Y. Y corresponds to the gender attribute, $Y = 0$ when the entry is female and $Y = 1$ otherwise. The remaining attributes constitute the X component. *The names are replaced by unique random strings in order to remove gender information.* All categorical features, e.g., department code, are converted into vectors using one-hot-encoding. All real-valued features, e.g., salary are standardized. We thus need to solve a binary classification problem and we have the following dataset for training: $\mathcal{D} := \{(X,Y) \mid Y \in \{0,1\},\ \ X$ is the attributes vector, excluding gender$\}$. Given a feature vector X, our ML algorithm needs to classify it as belonging to one of two classes – class-0 or class-1.

4.2 Choosing a Classifier and Empirical Results

As X is a combination of real and categorial features, we choose to use the *Random Forest Classifier* (RF) for our task. Another reason for choosing RF is due to the small size of the dataset – 294 datapoints. For this configuration of training data, RF is considered empirically superior to most other classification algorithms. We used the following set of hyper-parameters for our experiments:

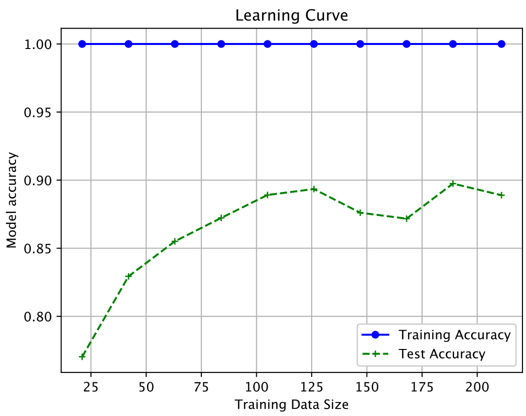

Figure 1: Performance of Random Forest Classifier in terms of accuracy.

1. 100 trees in the forest.
2. The *gain impurity measure* to divide the split from the root node and for subsequent splits.
3. 2 samples to split an internal node.

The dataset $\mathcal{D}$ is divided into training and test data – 80% of $\mathcal{D}$ is used to train RF and the remaining 20% is used as test data, also referred to as the "hold out test data". Let us represent the training data using $\mathcal{D}_{train}$ and the hold out test data using $\mathcal{D}_{test}$. The training progress of RF is illustrated in Fig. 1. The x-axis represents the number of training datapoints and the model accuracy is plotted along y-axis. The blue bold line represents the accuracy of RF within the training data $\mathcal{D}_{train}$. The green dashed line represents the accuracy of RF on the hold out test data $\mathcal{D}_{test}$. We observed that the training accuracy is very close to 1 after training with just 25 datapoints from $\mathcal{D}_{train}$. The accuracy on the test data $\mathcal{D}_{test}$, however, is low at the beginning. As RF is trained on more data, the test accuracy increases to 0.87. Hence, the accuracy of RF is strictly better than p $(= 0.72)$. Let $Y^{RF}(X)$ represent the map found by RF after training using $\mathcal{D}_{train}$. It follows from the discussion in Section 3.3 that $H\left(Y^{RF}(X)|X\right) < H(Y)$, that there is information about Y in X, and that the knowledge of X reduces the uncertainty in Y. RF is able to exploit some "gender pattern" in X in order to predict the gender Y at a rate that is better than guessing (completely disregard X when predicting Y). Put another way, suppose there is no gender information in X, then the RF predictions can only be as good as guessing.

Remark 1. *The RF worked admirably well for our purpose – to illustrate our idea. It may not be the optimal choice for another use-case, involving a different dataset. One workaround involves training an ensemble of classifiers, instead of a single one. We*

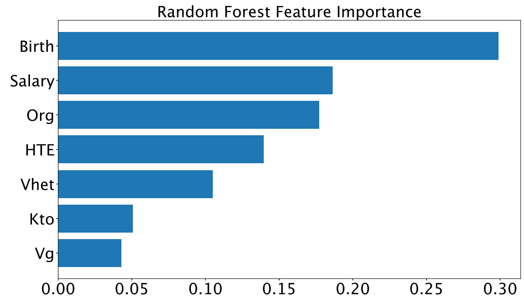

Figure 2: Feature Ranking.

then compare the accuracy of the ensemble to p for diagnostics. The accuracy of the ensemble equals the maximum of the classifier accuracies.

4.3 Feature Importance Analysis

The feature importance score measures the contribution of a particular feature to the prediction accuracy, higher scores indicate greater contributions. The top scoring features are more likely to have more Y information (gender information), as compared to low scoring features. As illustrated in Fig. 2, the *date of birth* is the highest ranked feature in our experiment. Out of all the features, it is likely to have the most gender information. This would be accounted for in cases such as Karlstad University, where the more junior individuals in these departments are women. The second highest ranked feature is salary, and again, this contains gendered information, given that women in Karlstad University, as elsewhere, are more likely to be in more junior positions and have lower salaries. Having said that, it must be noted that the highest salary is earned by a female Professor in our database. However, the corresponding datapoint is treated as an outlier by the classification algorithm.

It must be noted that there are several other studies that look at gender inequality in Technology, see, e.g., (Jaccheri, 2022). In particular, (Jaccheri, 2022) considers the under-representation of women in the field of Computer Science. Best practices are also presented for attracting, retaining, encouraging, and inspiring women in the future. The point of departure of our work, as compared to these studies, is that we present an automated way of detecting gender-bias, by estimating a suitable conditional entropy using ML. We therefore provide a simple and automated way of diagnosing possible gender biases within organizations. If a gender bias is detected by our framework, then the solutions highlighted in (Jaccheri, 2022) can be employed to improve the situation.

5 CONCLUSIONS

Given observations of random variables X and Y, we presented a ML based framework to estimate the conditional entropy $H(Y|X)$. Our framework is applicable for discrete-valued X and bi-valued Y. We trained a classification algorithm $\mathcal{A}$ to predict Y, given X. The training dataset was constructed using the X and Y observations. Algorithm $\mathcal{A}$ yielded $Y^{\mathcal{A}}$, an approximation of Y. Then, we used the prediction accuracy to calculate $H(Y^{\mathcal{A}}(X)|X)$, an approximation of $H(Y|X)$. We formally showed that $H(Y|X) < H(Y)$ when the prediction accuracy is strictly greater than p, where p is defined in equation (2). We thus obtained a simple sufficient condition on the accuracy to check the bias of the given dataset. We illustrated our methodology by analyzing the employee database from Karlstad University. We also discussed how ML tools such as *feature importance scores* can be used to obtain deeper insights.

REFERENCES

Beirlant, J., Dudewicz, E. J., Györfi, L., Van der Meulen, E. C., et al. (1997). Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Darbellay, G. A. and Wuertz, D. (2000). The entropy as a tool for analysing statistical dependences in financial time series. *Physica A: Statistical Mechanics and its Applications*, 287(3-4):429–439.

Dimpfl, T. and Peter, F. J. (2013). Using transfer entropy to measure information flows between financial markets. *Studies in Nonlinear Dynamics and Econometrics*, 17(1):85–102.

Drees, J. P., Gupta, P., Hüllermeier, E., Jager, T., Konze, A., Priesterjahn, C., Ramaswamy, A., and Somorovsky, J. (2021). Automated detection of side channels in cryptographic protocols: Drown the robots! In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, pages 169–180.

Golder, A., Das, D., Danial, J., Ghosh, S., Sen, S., and Raychowdhury, A. (2019). Practical approaches toward deep-learning-based cross-device power side-channel attack. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 27(12):2720–2733.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Gupta, P., Ramaswamy, A., Drees, J. P., Hüllermeier, E., Priesterjahn, C., and Jager, T. (2022). Automated information leakage detection: A new method combining machine learning and hypothesis testing with an application to side-channel detection in cryptographic protocols.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Heiberger, R. H. (2022). Applying machine learning in sociology: How to predict gender and reveal research preferences. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, pages 1–24.

Jaccheri, L. (2022). Gender issues in computer science research, education, and society. In *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 1*, pages 4–4.

MacKay, D. J., Mac Kay, D. J., et al. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Molina, M. and Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*, 45:27–45.

Pham, D.-T. (2004). Fast algorithms for mutual information based independent component analysis. *IEEE Transactions on Signal Processing*, 52(10):2690–2700.

Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, 16(3):e12962.