Feature Selection via a Fourier Framework

Anonymous Author(s)
Affiliation
Address
email

Abstract

The Fourier analysis over the Boolean cube has been an essential tool in a wide range of problems in computer science. However, such analysis is restricted to mutually independent random variables making its practical usage limited. We demonstrate that this limitation is overcome, developing a novel Fourier expansion on the Boolean cube with correlated variables. We further apply our Fourier analysis to stochastic mappings, extending its capabilities beyond classical usage. As one application of this analysis, we investigate the feature selection problem and reformulate it in the Fourier domain. We propose two feature selection algorithms, one for supervised and the other for an unsupervised variant of the problem. The computational complexity of the algorithms can be as low as $O(nd)$ with n being the number of samples and d the number of features. Based on our Fourier analysis, we show that for binary features the proposed algorithms find provably *asymptotically optimal* feature subsets. Through exhaustive numerical experiments, we demonstrate that our methods outperform state-of-the-art feature selection algorithms (e.g., mRMR, ReliefF, MCFC, Laplacian Score) on various data sets.

1 Introduction

Feature selection contributes not only to reducing computational complexity and running time, but also to interpretability of the learning model. The objective is to remove as many features as possible without significantly increasing the classification loss. A natural solution is the *wrapper* method in which the feature subsets are evaluated directly by an induction algorithm [13]. However, this approach is computationally expensive and, hence, prohibitive in large data sets. An alternative solution is the *filter* approach in which an intermediate measure, independent of the induction learning algorithm, is used to evaluate the feature subsets. Filter methods are preferred as they are computationally more efficient and relatively robust against overfitting. Several measures has been introduced in the literature [2, 8, 12, 21, 27, 29, 32]. The challenge in this area, that remains open, is to design a computationally efficient measure which is provably related to the generalization loss.

To address this challenge, in this work, we take a different approach. We develop a novel Fourier expansion for functions of *correlated* binary random variables. The Fourier expansion provides a powerful tool to characterize *nonlinear* redundancies in features and *nonlinear* dependencies in features-label relation. Using this framework, we study supervised and unsupervised feature selection, and propose our algorithms. Aiming to address the above challenges, we provide a theoretical analysis and derive conditions under which our algorithms find the optimal feature subset. Further, through numerical experiments, we show that our algorithms outperform several well-known feature selection techniques. That said, the contributions of this paper are three-fold as summarized in the following.

Fourier expansion for correlated random variables: The standard Fourier expansion on the Boolean cube has been central in a wide range of applications such as computational learning theory [3, 15, 23, 24, 25], noise sensitivity [18, 26], and other information-theoretic problems [9]. In this expansion, any real-valued function on the Boolean cube can be written as a linear combination of

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

parities [26, 30]. The Fourier coefficients quantify the levels of "nonlinearities" in a function. Highly nonlinear functions have Fourier expansion with large coefficients for high-degree parities. Thus, the Fourier expansion is potentially a powerful tool in the feature selection problem. However, there are limitations making it impractical for this purpose. First, it is assumed that the features are mutually independent. Secondly, this expansion is defined only for deterministic functions of the features. These assumptions are too strong, as real-world learning problems, often, involve correlated features with stochastic labeling.

In this work, we address both limitations. First by developing a generalized Fourier expansion for functions of *correlated binary* random variables (Proposition 1). For this purpose, we adopt a Gram-Schmidt-type orthogonalization and construct a set of orthogonal basis functions. We address the second limitation by adapting our Fourier expansion to the more general space of stochastic mappings (e.g., mappings from one probability space to another). To the best of our knowledge, this is the first generalization of the Fourier expansion for correlated binary random variables. Although this Fourier expansion is defined on the Boolean cube, our algorithms are applicable to non-binary features too. We view the Binary Fourier as a framework that captures a special class of nonlinearities — those characterized via the *parities*. Alternatively, we could generalize our Fourier expansion to discrete features and, based on it, design feature selection algorithms. However, such a generalization requires character theory, which is beyond the scope of this paper. We note that there are other forms of orthogonal decomposition including the Hoeffding-Sobel decomposition [7, 16, 28] and its generalization [7]. However, such decompositions are basis-free. Our Fourier expansion is defined by constructing a set of *orthonormal* basis functions which makes it suitable for feature selection.

Unsupervised feature selection: In practice, collecting unlabeled data is usually a less expensive task that motivates to develop unsupervised feature selection algorithms. Popular unsupervised feature selection approaches rank the features based on local geometrical structures (e.g. Laplacian Score [14]), manifold structures (e.g. MCFC [5]), or their discriminate power (e.g. UDFS [31]). In this paper, we aim to capture nonlinear redundancies in the statistics of the data. We take an information-theoretic perspective, and group the features into *redundant* and *sufficiently informative*. All the accessible *information* about the data can be captured from the later group. The former is statistically a nonlinear function of the later, hence can be removed without affecting the learning's performance. This approach extends the notion of "redundant" features to the unsupervised setting [4, 20, 32]. Built upon the Fourier framework, we develop an Unsupervised Fourier Feature Selection (UFFS) algorithm, which captures the redundant features. Instead of ranking the features, the UFFS finds redundant features and declares the rest of the features as informative. We prove that, when the features are binary, all nonlinear redundancies are detected. In that case, the algorithm finds the smallest sufficiently informative feature subset (Theorem 1). Although such guarantees are established for binary features, we empirically show that the algorithm performs well on non-binary real-valued features too. Through comprehensive numerical experiments, we show that the UFFS is applicable in a wide range of applications and significantly outperforms popular methods such as MCFC, Laplacian Scre (LS), and UDFS (see Table 2).

Supervised feature selection: Well-known criteria for supervised feature selection can be grouped into correlation measures (e.g., Pearson correlation, Fisher Score), information-theoretic measures [2, 21, 27, 29, 32], and Kernel-based measures [8, 12]. Although correlation criteria are computationally more efficient, they usually are not able to detect *nonlinear* dependencies in features-label relations. Methods based on kernels can detect the nonlinear dependencies. However, the computational complexity of computing a kernel grows super linearly, if not quadratic, with the number of the samples [6]. Mutual Information (MI) criteria, on the other hand, can detect nonlinear dependencies with lower computational complexity [2]. In addition, mutual information can be used to bound the Bayes misclassification rate. However, estimating multi-variate mutual information is known to be a difficult task with high sample complexity.

In this work, we propose a computationally efficient measure that captures nonlinear dependencies and has provable relation to the Bayes misclassification rate. For that, we first formulate the feature selection in an ideal setting as follows: given a parameter k, the objective is to find k features such that the misclassification rate of the Bayes classifier, restricted to them, is minimized. We reformulate this problem in the Fourier domain and characterize the optimal feature subset. Build upon this formulation, we develop a measure to evaluate feature subsets. We prove that when the features are binary, an exhaustive search based on this measure finds an asymptotically optimal feature subset. That is a feature subset whose Bayes misclassification rate is at most $O(n^{-1/2})$ larger than that of the optimal feature subset (Theorem 2). Since the exhaustive search is an NP-hard problem, we propose a search algorithm called fixed-depth search. Given a depth parameter t, the idea is to evaluate only the feature subsets of size at most t. With this approach, we propose a Supervised Fourier Feature Selection (SFFS) algorithm with computational complexity $O(nd^t)$, where n is the number of the samples and d is the number of the features. Through our numerical experiments, we show in Figure 1 that SFFS, even with $t = 1$ or 2, outperforms well-known feature selection algorithm (e.g., mRMR, and ReliefF, MI) on various data sets (See Section 4).

Notations: As a shorthand, in this paper, for any natural number m, the set $\{1, 2, \cdots, m\}$ is denoted by $[m]$. Also, for any subset $\mathcal{J} \subseteq [d]$ with ordered elements $\{j_1, j_2, \cdots, j_k\}$, the vectors $(X_{j_1}, X_{j_2}, \cdots, X_{j_k})$, and $(x_{j_1}, x_{j_2}, \cdots, x_{j_k})$ are denoted, respectively, by $\mathbf{X}^{\mathcal{J}}$ and $\mathbf{x}^{\mathcal{J}}$.

2 Fourier Expansion for Correlated Random Variables

In this section, we propose a novel Fourier expansion for functions of *correlated* binary features. For convenience in presenting our results, we restrict ourselves to binary features. One can extend our approach to discrete features via *group characters*.

We start with a brief overview of the well-known Fourier expansion on Boolean cube [26]. Let $\mathbf{X} = (X_1, X_2, ..., X_d)$ be a vector of mutually independent random variables taking values from a subset $\mathcal{X} \subset \mathbb{R}^d$. Let μ_j and σ_j be the mean and standard-deviation of $X_j, j \in [d]$. Suppose that these random variables are non-trivial, that is $\sigma_j > 0$ for all $j \in [d]$. The Fourier expansion is defined via a set of basis functions called *parities*. The *parity* for a subset $\mathcal{S} \subseteq [d]$ is defined as $\phi_{\mathcal{S}}(\mathbf{x}) \triangleq \prod_{i \in \mathcal{S}} \frac{x_i - \mu_i}{\sigma_i}$ for all $\mathbf{x} \in \mathbb{R}^d$. Since X_i's are mutually independent, the parities are orthonormal, that is $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})^2] = 1$ for any subset $\mathcal{S}$, and $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})\ \phi_{\mathcal{T}}(\mathbf{X})] = 0$ when $\mathcal{T} \neq \mathcal{S}$ (that is $\exists x \in \mathcal{T} \bigcup \mathcal{S}$ such that $x \notin \mathcal{T} \bigcap \mathcal{S}$). Under the assumption that $\mathcal{X} = \{-1, 1\}^d$, the parities form an orthonormal basis for the space of bounded function $f : \{-1, 1\}^d \mapsto \mathbb{R}$ [26]. That is, any bounded function $f : \{-1, 1\}^d \mapsto \mathbb{R}$ can be written as a linear combination of the form $f(\mathbf{x}) = \sum_{\mathcal{S} \subseteq [d]} \mathtt{f}_{\mathcal{S}}\ \phi_{\mathcal{S}}(\mathbf{x})$, for all $\mathbf{x} \in \{-1, 1\}^d$, where $\mathtt{f}_{\mathcal{S}} \in \mathbb{R}$ are called the *Fourier coefficients* of f with respect to $P_{\mathbf{X}}$, the distribution of $\mathbf{X}$. Further, the Fourier coefficients can be computed as $\mathtt{f}_{\mathcal{S}} = \mathbb{E}[f(\mathbf{X})\phi_{\mathcal{S}}(\mathbf{X})]$, for all subsets $\mathcal{S} \subseteq [d]$.

With this overview, we are ready to construct our Fourier expansion. Note that, in a general probability space with correlated features, the standard Fourier expansion is no longer well-defined. Because, the parities $\phi_{\mathcal{S}}$ are not necessarily orthogonal. That said, we construct our Fourier expansion by adopting a Gram-Schmidt-type procedure to make the parities *orthogonal*. Then, we use this basis to develop our Fourier expansion for function of correlated random variables. The orthogonalization process is explained in the following.

Orthogonalization process: Fix the following ordering for subsets of $[d]$:

$$\varnothing, \{1\}, \{2\}, \{1, 2\}, \{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}, \cdots, \{1, 2, ..., d\}. \tag{1}$$

For any pair of functions g_1, g_2 denote $\langle g_1, g_2 \rangle = \mathbb{E}[g_1(\mathbf{X})g_2(\mathbf{X})]$. We apply the Gram-Schmidt process on the parities $\phi_{\mathcal{S}_i}$ with the above ordering and $\langle g_1, g_2 \rangle$ as the inner product. With this method, the orthogonalized parity corresponding to the ith subset is obtained from the following operation:

$$\tilde{\psi}_{\mathcal{S}_i} = \phi_{\mathcal{S}_i} - \sum_{j=1}^{i-1} \langle \psi_{\mathcal{S}_j}, \phi_{\mathcal{S}_i} \rangle\, \psi_{\mathcal{S}_j}, \qquad \psi_{\mathcal{S}_i} = \begin{cases} \frac{\tilde{\psi}_{\mathcal{S}_i}}{\|\tilde{\psi}_{\mathcal{S}_i}\|_2} & \text{if } \|\tilde{\psi}_{\mathcal{S}_i}\|_2 > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where $\|\tilde{\psi}_{\mathcal{S}_i}\|_2 = \sqrt{\langle \tilde{\psi}_{\mathcal{S}_i}, \tilde{\psi}_{\mathcal{S}_i} \rangle}$. Note that the first orthogonalized parity is given by $\psi_{\varnothing}(\mathbf{x}) \triangleq 1$ for all $\mathbf{x} \in \mathbb{R}^d$. It is not difficult to check that the resulted nontrivial parities $\psi_{\mathcal{S}_i}$'s are orthonormal, that is $\langle \psi_{\mathcal{S}_i}, \psi_{\mathcal{S}_j} \rangle = 0$ for $i \neq j$ and $\langle \psi_{\mathcal{S}_i}, \psi_{\mathcal{S}_i} \rangle = 1$ if $\psi_{\mathcal{S}_i}$ is not trivial. Note also that different orderings for the subsets of $[d]$ result in different orthogonalized parities. Hence, unless otherwise stated, we use the ordering in (1). Next, in the proceeding proposition, we establish our Fourier expansion for functions of correlated binary random variables. The proof is given in Appendix B.

Proposition 1 (**Correlated Fourier Expansion**). *Let $P_{\mathbf{X}}$ be a probability distribution on $\{-1, 1\}^d$ and $f : \{-1, 1\}^d \mapsto \mathbb{R}$ be a bounded function. Let $\psi_{\mathcal{S}}$'s be the orthogonalized parities as defined in (2). Then, for all $\mathbf{x} \in \{-1, 1\}^d$ except a measure-zero subset, f is decomposed as*

$f(\mathbf{x}) = \sum_{\mathcal{S}\subseteq[d]} f_{\mathcal{S}}\psi_{\mathcal{S}}(\mathbf{x})$*, where the summation is taken over all* $\mathcal{S}$ *for which* $\psi_{\mathcal{S}}$ *is not trivial. Further, the coefficients* $f_{\mathcal{S}}$ *are unique and obtained from* $f_{\mathcal{S}} = \mathbb{E}[f(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})]$.

Remark 1. In contrary to our Fourier expansion, which is established only for binary features, the orthogonalization process is not restricted to such an assumption. Because, by construction, the orthogonalized parities are orthonormal for any value domain $\mathcal{X} \subset \mathbb{R}^d$. If $\mathcal{X} = \{-1, 1\}^d$, then the parities span the space of all function on $\mathcal{X}$; otherwise they span a *subspace* of such functions. We clarify this in the following example.

Example 1. Set $d = 3$ and let X_1 and X_2 be independent random variables with Gaussian distribution $N(0, 1)$. Suppose $X_3 = X_1X_2$ with probability one. There are eight standard parities, one for each subsets, as $(1, x_1, x_2, x_1x_2, x_3, x_1x_3, x_2x_3, x_1x_2x_3)$. By performing the orthogonalization process, as in (2), there are only four non-trivial orthogonalized parities as $\psi_{\varnothing} = 1, \psi_{\{1\}} = x_1, \psi_{\{2\}} = x_2$, and $\psi_{\{1,2\}} = x_1x_2$. The rest of the parities are zero, because $\|\tilde{\psi}_{\mathcal{S}}\|_2 = 0$ for any of the subsets $\{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$. Now, suppose we change the relation of X_3 to $X_3 = \mathsf{sign}[X_1X_2]$. In this case, after the orthogonalization process, there are eight non-trivial parities. For instance, it is not difficult to check that $\tilde{\psi}_{\{3\}} = x_3 - \frac{2}{\pi}x_1x_2$. Hence, $\|\tilde{\psi}_{\{3\}}\|_2 > 0$, implying that $\psi_{\{3\}}$ is not redundant.

This example shows that the orthogonalization removes nonlinear redundancies. However, as discussed in Remark , it captures only a class of non-linearities for non-binary features. This is because the orthogonalization process is based on the binary Fourier expansion. We view our binary Fourier as a framework that captures a special class of nonlinearities — those characterized via orthogonalized parities. Our numerical experiments confirm that such an approximation is sufficient to outperform state-of-the-art methods for many data sets (see Table 2). It is also noted that dimension reduction methods such as PCA do not necessarily capture the nonlinear redundancies. For instance, the features in the above example are pairwise uncorrelated and, hence, the covariance matrix is the identity matrix. In Appendix H, we show that our approach outperforms PCA as well.

3 Feature Selection: a Fourier Perspective

In this section, we build upon our Fourier expansion to study unsupervised and supervised feature selection problems. The Fourier expansion provides a powerful tool to characterize non-linear relations among the features and the labels.

3.1 Unsupervised Feature Selection: Informative vs. Redundant

We build upon our orthogonalization process in (2) and develop our UFFS algorithm (see Algorithm 1) to capture non-linear redundancies in the features. For this purpose, we first define a measure to identify the features as “sufficiently informative” and “redundant”. Intuitively, the former group contains all the *information* accessible from the features. The later consists of the features that are a function of the “informative” features, and hence, can be removed from the data set.

Suppose that there are d features denoted by the random vector $\mathbf{X} = (X_1, X_2, ..., X_d)$ taking values from a subset $\mathcal{X} \subset \mathbb{R}^d$. We say $\mathcal{J} \subseteq [d]$ is a “sufficiently informative” feature subset, if $H(\mathbf{X}) = H(X^{\mathcal{J}})$, where H is the Shannon entropy. This definition is related to the notion of Markov Blanket [21], as $\mathcal{J}$ is a Markov blanket for any feature in $\mathcal{J}^c$. Also, $\mathcal{J}$ being sufficiently informative immediately leads to $\mathcal{J}^c$ being redundant. Because, the condition $H(\mathbf{X}) = H(X^{\mathcal{J}})$ implies that there exists a mapping T, such that $X^{\mathcal{J}^c} = T(X^{\mathcal{J}})$, with probability one [10]. Hence, all the features not included in $\mathcal{J}$ can be removed. With this elimination, the dimension is reduced from d to $|\mathcal{J}|$. As there are multiple such $\mathcal{J}$’s, the objective is to find the smallest one[1]. Tolerating small amounts of imperfections, we formalize the above notion in the following.

Definition 1 (Sufficiently Informative). For discrete features and $0 \leqslant \varepsilon \leqslant 1$, a feature subset $\mathcal{J}$ is said to be ε*-sufficiently informative*, if $H(\mathbf{X}|X^{\mathcal{J}}) \leqslant \varepsilon$. The feature subset $\mathcal{J}$ is sufficiently informative, if $H(\mathbf{X}|X^{\mathcal{J}}) = 0$. Such $\mathcal{J}$ is called *minimal*, if it has the minimum cardinality among all sufficiently informative feature subsets.

Next, we make a connection between the above definition and the orthogonalization process in (2). We employ this process to extract a sufficiently informative feature subset. Fix the standard ordering as in

[1] The set of all features is a trivial example of a sufficiently informative feature subset.

(1), and generate the orthogonalized parities $\psi_{\mathcal{S}_i}$. We start by deriving an upper-bound on $H(\mathbf{X}|X^{\mathcal{J}})$ in terms of the orthogonalized parities in (2). Note that $H(\mathbf{X}|X^{\mathcal{J}}) = H(X^{\mathcal{J}^c}|X^{\mathcal{J}})$, where $\mathcal{J}^c$ is the complement of $\mathcal{J}$. Thus, from the chain rule [10], this quantity equals to $\sum_{i\in\mathcal{J}^c} H(X_i|X^{\mathcal{J}}, X^{i-1})$. As X_i is a discrete random variable, then $H(X_i|X^{\mathcal{J}}, X^{i-1}) = H(\phi_{\{i\}}|X^{\mathcal{J}}, X^{i-1})$, where $\phi_{\{i\}}$ is the standard parity as in Section 2. From the orthogonalization process in (2), we can write $\phi_{\{i\}} = \tilde{\psi}_{\{i\}} + \sum_{\mathcal{S}\subseteq[i-1]} \alpha_{\mathcal{S}}\psi_{\mathcal{S}}$, where $\alpha_{\mathcal{S}} = \langle \phi_{\{i\}}, \psi_{\mathcal{S}}\rangle$. Therefore, as the terms in the summation are functions of X^{i-1}, we get the following upper-bound

$$H(\phi_{\{i\}}|X^{\mathcal{J}}, X^{i-1}) = H(\tilde{\psi}_{\{i\}}|X^{\mathcal{J}}, X^{i-1}) \leqslant H(\tilde{\psi}_{\{i\}}), \tag{3}$$

where the last inequality follows by removing the conditioning in the entropy. Lastly, by adapting this bound for all $i \in \mathcal{J}^c$, we get our designed bound: $H(\mathbf{X}|X^{\mathcal{J}}) \leqslant \sum_{i\in\mathcal{J}^c} H(\tilde{\psi}_{\{i\}})$. Using this upper bound, we prove the following theorem in Appendix C.

Theorem 1. *Let $\mathcal{J}_\varepsilon \subseteq [d]$ be the set of all i's such that $\|\tilde{\psi}_{\{i\}}\|_2 > \varepsilon$. Then, for sufficiently small $\varepsilon > 0$, $H(\mathbf{X}|X^{\mathcal{J}_\varepsilon}) = O(\varepsilon)$. Further, if the features take values from $\{-1, 1\}^d$, then $\mathcal{J}_\varepsilon$ with $\varepsilon = 0$ is a sufficiently informative subset with minimum cardinality.*

Algorithm 1 Unsupervised Fourier Feature Selection

```
Input: $n$ training samples $\mathbf{x}_i \in \mathbb{R}^d$, depth parameter $t \leqslant d$, and redundancy threshold $\varepsilon \in (0,1)$
procedure ORTHOGONALIZATION
    Compute the empirical mean $\hat{\mu}_j$ and standard deviation $\hat{\sigma}_j$ of each feature.
    Generate all subsets $\mathcal{S}_i \subseteq [d]$ with size at most $t$ and with the standard ordering as in (1). Compute the matrix $\hat{\mathbf{B}}$ with elements:
        $\hat{b}_{j,i} \leftarrow \frac{1}{n}\sum_{l=1}^{n}\Big[\prod_{u\in\mathcal{S}_j}\frac{x_{lu}-\hat{\mu}_u}{\hat{\sigma}_u}\Big]\Big[\prod_{v\in\mathcal{S}_i}\frac{x_{lv}-\hat{\mu}_v}{\hat{\sigma}_v}\Big]$
    Set $\hat{\mathbf{A}} \leftarrow \hat{\mathbf{B}}$
    for row $j$ of $\hat{\mathbf{A}}$ do
        update the $j$th row: $\hat{\mathbf{A}}_{j,*} \leftarrow \hat{\mathbf{A}}_{j,*} - \sum_{\ell<j} \hat{a}_{\ell,j}\hat{\mathbf{A}}_{\ell,*}$
        Compute norm$(\mathcal{S}_j) \leftarrow \sqrt{[\hat{b}_{j,j} - \sum_{r<j}\hat{a}_{r,j}^2]^+}$
        if norm$(\mathcal{S}_j) \leqslant \varepsilon$ then
            Set the $j$th row of $\hat{\mathbf{A}}$ zero: $\hat{\mathbf{A}}_{j,*} \leftarrow \mathbf{0}$
        else
            Normalize the $j$th row: $\hat{\mathbf{A}}_{j,*} \leftarrow \frac{\hat{\mathbf{A}}_{j,*}}{\text{norm}(\mathcal{S}_j)}$
    return All $j \in [d]$ with norm$(j) \geqslant \varepsilon$ as non-redundant.
```

Unsupervised Feature Selection: As a result of Theorem 1, $\|\tilde{\psi}_{\{i\}}\|_2$ can be viewed as a measure of the *redundancy* of each feature and that the orthogonalization procedure can remove them. We use this measure for unsupervised feature selection, where n independent and identically distributed (i.i.d.) instances $\{\mathbf{x}(i), i \in [n]\}$ are available. The idea is to perform the orthogonalization process as in (2) and find the features j for which $\|\tilde{\psi}_{\{j\}}\|_2$ is smaller than a threshold ε. These features are declared as redundant. As for the algorithm, two issues need to be addressed: 1) the orthogonalization is an NP hard process, as there are 2^d feature subsets, and 2) estimation of $\|\tilde{\psi}_{\{j\}}\|_2$ from the training instances. In what follows, we address these issues.

1) Fixed-depth search: We propose to address the first issue using a *fixed-depth* search method. Given a parameter $t \leqslant d$, the orthogonalization is performed only on feature subsets of size at most t. For that we use the standard ordering as in (1), but restricted to subsets of size at most t.

2) Empirical orthogonalization: We propose a recursive formula to perform the orthogonalization and estimate $\|\tilde{\psi}_{\{j\}}\|_2$. Let $b_{j,i} = \langle \phi_{\mathcal{S}_j}, \phi_{\mathcal{S}_i}\rangle$, and define $a_{j,i} = \langle \psi_{\mathcal{S}_j}, \phi_{\mathcal{S}_i}\rangle$. Therefore, (2) can be written as that $\tilde{\psi}_{\mathcal{S}_i} = \phi_{\mathcal{S}_i} - \sum_{j<i} a_{j,i}\psi_{\mathcal{S}_j}$. Due to the orthonormality of $\psi_{\mathcal{S}_i}$'s, we obtain that $\|\tilde{\psi}_{\mathcal{S}_i}\|_2^2 = b_{i,i} - \sum_{j<i} a_{j,i}^2$. Further, the coefficients $a_{j,i}$ can be calculated recursively as

$$a_{j,i} = \frac{1}{\sqrt{b_{j,j} - \sum_{r<j} a_{r,j}^2}}\Big(b_{j,i} - \sum_{\ell<j} a_{\ell,j}a_{\ell,i}\Big) \tag{4}$$

With this formulas, we first compute an empirical estimate of $b_{j,i}$'s, denoted by $\hat{b}_{j,i}$. Then, we compute an estimation of $a_{j,i}$'s (denoted by $\hat{a}_{j,i}$) by calculating (4) with $b_{j,i}$ and $a_{j,i}$ replaced by $\hat{b}_{j,i}$ and $\hat{a}_{j,i}$, receptively. Lastly, we obtain an empirical estimate of $\|\tilde{\psi}_{\mathcal{S}_i}\|_2$ by computing $\sqrt{\hat{b}_{i,i} - \sum_{j<i}\hat{a}_{j,i}^2}$.

3) Clustering the features: The above two processes are implemented in Algorithm 1. For large dimensional data sets, we can group the features into multiple clusters of approximately equal size (say m features). Then, we perform Algorithm 1 on each cluster, and remove the redundant features within it. With this approach, the computational complexity of UFFS algorithm with depth parameter t and cluster size m is $O(n\frac{d}{m}m^{2t})$. The parameters m and t are chosen independently of (n, d). For instance, we choose $t = 3$ and $m = 40$. As a result, we obtain a complexity linear in the size of the data set. We present our experimental results in Section 4.

3.2 Supervised Feature Selection

We build upon our Fourier expansion in Section 2 and propose a Supervised Fourier Feature Selection (SFFS) algorithm. We consider the learning problem with d real-valued features and with labels taking values from $\mathcal{Y}$. The features $\mathbf{X} \in \mathbb{R}^d$ and the label $Y \in \mathcal{Y}$ are generated according to an unknown distribution $P_{\mathbf{X}Y}$. Available are n i.i.d. instances $\big\{(\mathbf{x}(i), y(i)), i = 1, 2, ..., n\big\}$ generated from $P_{\mathbf{X}Y}$. For convenience in presenting the theoretical results, we restrict ourselves to binary classification with $0-1$ loss function. In this case, the expected loss is the *misclassification* probability.

We describe the feature selection problem by first defining the optimum feature subset and the minimum *misclassification* probability in the ideal setting, where $P_{\mathbf{X},Y}$ is known. Given $k \leqslant d$, the optimum feature subset $\mathcal{J}^*$ and the minimum misclassification probability $P_{\text{opt}}(k)$ are defined as

$$P_e(\mathcal{J}) = \min_{g\in\mathcal{G}_k} \mathbb{P}_{\mathbf{X}Y}\{Y \neq g(X^{\mathcal{J}})\}, \qquad \mathcal{J}^* = \underset{\mathcal{J}\subseteq[d],\,|\mathcal{J}|=k}{\arg\min}\; P_e(\mathcal{J}), \qquad P_{\text{opt}}(k) = P_e(\mathcal{J}^*), \tag{5}$$

where $\mathcal{G}_k$ is the collection of all functions on $\mathbb{R}^k$. In agnostic settings, where only a training data set is available, the above optimization is infeasible to solve. Instead, an intermediate measure M_n is defined to evaluate feature subsets using the training instances. Then the feature selection problem reduces to the optimization: $\hat{\mathcal{J}}_n = \arg\min_{\mathcal{T}\in\mathsf{T}_k} M_n(\mathcal{T})$, where T_k is a collection of feature subsets with at most k-elements.

Within this framework, we construct our SFFS algorithm by proposing a measure to evaluate different feature subsets (see (7)). For binary features, we prove in Theorem 2 that maximizing this measure over different feature subsets give $\hat{\mathcal{J}}_n$, such that $P_e(\hat{\mathcal{J}}_n)$ converges to $P_e(\mathcal{J}^*)$, as the sample size n tends to ∞. Although the theoretical guarantees are established for binary features, SFFS algorithm is not restricted to such assumptions. We empirically show, in section 4, that SFFS outperforms several state-of-the-art feature selection algorithms on many benchmark data sets.

We start with developing a representation of $P_{\text{opt}}(k)$ in the Fourier domain. Note that the Bayes predictor of Y from the observation $x^{\mathcal{J}}$ is given by $g^*(x^{\mathcal{J}}) = \text{sign}\big[\mathbb{E}[Y|x^{\mathcal{J}}]\big]$. We proceed by characterizing the above conditional expectation in the Fourier domain. As a key ingredient in our characterization, we need to define the notion of *projection onto a feature subset*. Let $\mathcal{J} \subseteq [d]$ be a feature subset with k elements. Denote the elements of $\mathcal{J}$, in the ascending order, as $j_1 < j_2 < \cdots < j_k$. Fix the following ordering of subsets of $\mathcal{J}$:

$$\varnothing, \{j_1\}, \{j_2\}, \{j_1, j_2\}, \{j_3\}, \{j_1, j_3\}, \{j_2, j_3\}, \{j_1, j_2, j_3\}, \cdots, \{j_1, j_2, ..., j_k\}.$$

Apply the orthogonalization process with respect this ordering and to all the parities $\psi_\mathcal{S}$ with $\mathcal{S} \subseteq \mathcal{J}$. Let $\psi_{\mathcal{S}_i}, i = 1, 2, ..., 2^k$ be the resulted orthogonalized parities. This process is called orthogonalization with respect to the feature subset $\mathcal{J}$. with this process, we are ready to define the projection onto $\mathcal{J}$.

Definition 2 (**Projection onto a subset).** Given a feature subset $\mathcal{J} \subseteq [d]$, let $\psi_\mathcal{S}$'s be the orthogonalized parities w.r.t $\mathcal{J}$. The projection of the label Y onto $\mathcal{J}$ is defined as $f^{\subseteq\mathcal{J}}(\mathbf{x}) \triangleq \sum_{\mathcal{S}\subseteq\mathcal{J}} \mathbb{E}[Y\psi_\mathcal{S}]\psi_\mathcal{S}(\mathbf{x})$, where the expectation is taken with respect to $P_{\mathbf{X},Y}$.

We show in Lemma 2, in Appendix D, that $f^{\subseteq\mathcal{J}}(\mathbf{x})$ is, in fact, equal to the conditional expectation $\mathbb{E}[Y|x^{\mathcal{J}}]$. Further, based on the above argument, we prove the following proposition in Appendix E.

Proposition 2. *Suppose* $(\mathbf{X}, Y) \sim P_{\mathbf{X}Y}$, *where* X_i*'s and* Y *take values from* $\{-1, 1\}$. *Then the minimum attainable misclassification probability equals to*

$$P_{opt}(k) = \frac{1}{2}\left[1 - \max_{\mathcal{J}\subseteq[d],\,|\mathcal{J}|=k} \|f^{\subseteq\mathcal{J}}\|_1\right]. \tag{6}$$

Further, an optimal k*-variable predictor of the labels is given by the function* $\text{sign}\big[f^{\subseteq\mathcal{J}^*}(\mathbf{x})\big]$, *where* $\mathcal{J}^*$ *is an optimal feature subset that maximizes the* 1*-norm expression above.*

A Measures for Feature Selection: Based on Proposition 2, we define $M_n^{(1)}(\mathcal{J})$ to be an empirical estimate of $\|f^{\subseteq\mathcal{J}}\|_1$. More precisely, given the training instances $(\mathbf{x}(i), y(i)), 1 \leqslant i \leqslant n$, this estimation is obtained from

$$M_n^{(1)}(\mathcal{J}) = \|\widehat{f^{\subseteq\mathcal{J}}}\|_1 \triangleq \frac{1}{n-1}\sum_{i=1}^{n} \Bigg| \sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{f}_\mathcal{S}\hat{\psi}_\mathcal{S}(\mathbf{x}(i)) - \frac{1}{n} y(i)\big(\hat{\psi}_\mathcal{S}(\mathbf{x}(i))\big)^2 \Bigg|, \tag{7}$$

where $\hat{f}_{\mathcal{S}}$ are the empirical estimation of the Fourier coefficients. A more detailed discussion on the derivation of this measure is presented in Appendix A. We construct our SFFS algorithm by adopting the fixed-depth search in Subsection 3.1 and using the above measure (see Algorithm 2). We conclude this section by proving our theoretical guarantees for SFFS algorithm. We present the following theorem which is proved in Appendix F.

Theorem 2. *Fix $k \leqslant d$ and let $\hat{\mathcal{J}}_n$ be the feature subset maximizing $M_n^{(1)}$ which is defined in (7). Let $\mathcal{J}^*$ be the optimum feature subset as in (6). Then, with probability at least $(1-\delta)$, the following bound holds*

$$P(\hat{\mathcal{J}}_n) \leqslant P(\mathcal{J}^*) + \sqrt{\frac{\lambda(k)}{n-1}\log(\frac{d}{\delta})} + \frac{2^{k/2}}{\sqrt{n-1}},$$

where $\lambda(k) = 32\ k2^{2k}c_k^2$, with $c_k \triangleq \max_{\mathcal{S}\subseteq[d],|\mathcal{S}|\leqslant k}\|\psi_{\mathcal{S}}\|_\infty^2$.

Algorithm 2 Supervised Fourier Feature Selection (SFFS)

```
Input: $n$ training samples $(\mathbf{x}_i, y_i)$, desired number of features $k$, and the depth parameter $t \leqslant k$
Output: Feature subset $\hat{\mathcal{J}}_n$
procedure FEATURE SELECTION:
    Rank all $t$-variable features subsets $\mathcal{T}$ according to $M_n^{(1)}$ as in (7) or $M_n^{(2)}$ as in (8).
    If $\mathcal{T}_i$ are the subsets in the descending order, set $\hat{\mathcal{J}}_n = \bigcup_{i=1}^{r}\mathcal{T}_i$, where $r$ chosen such that the union has $k$ different elements.
return $\hat{\mathcal{J}}_n$
```

Alternative measure: Instead of 1-norm, we can use the 2-norm measure $\|f^{\subseteq\mathcal{J}}\|_2^2$. From Parseval's identity $\|f^{\subseteq\mathcal{J}}\|_2^2 = \sum_{\mathcal{S}\subseteq\mathcal{J}} f_{\mathcal{S}}{}^2$. We prove the following statement in Appendix I.

Lemma 1. *For binary features and labels, the following bounds hold*

$$\frac{1}{2}(1 - \max_{\mathcal{J}\subseteq[d]:\ |\mathcal{J}|=k}\|f^{\subseteq\mathcal{J}}\|_2) \leqslant P_{opt}(k) \leqslant \frac{1}{2}(1 - \max_{\mathcal{J}\subseteq[d]:\ |\mathcal{J}|=k}\|f^{\subseteq\mathcal{J}}\|_2^2).$$

Note that the above bounds are close to $P_{opt}(k)$ when $\|f^{\subseteq\mathcal{J}}\|_2$ is close to 1— hence a justification for using 2-norm. That said, instead of $M_n^{(1)}$, we can use the following measure

$$M_n^{(2)}(\mathcal{J}) = \|\widehat{f^{\subseteq\mathcal{J}}}\|_2^2 \triangleq \sum_{\mathcal{S}\subseteq\mathcal{T}}(\hat{f}_{\mathcal{S}})^2. \tag{8}$$

Although we provide theoretical guarantees for $M_n^{(1)}$, we now use $M_n^{(2)}$ as an approximation that further reduces the running time of the algorithm. Further, we estimate only the standard Fourier coefficients, not the correlated ones, by first running the UFFS. With that, the computational complexity of our SFFS algorithm for a fixed k is $O(nd^t)$.

4 Numerical Experiments

We now compare the performance our UFFS and SFFS algorithms (Algorithm 1 and 2) with a number of well-known methods for unsupervised and supervised feature selection. Our numerical results are presented in two parts: one for unsupervised and one for supervised setting. The real-world data sets are the UCI repository Isolet, HAPT, Sonar, COIL20, and Wine data sets [11], and the USPS hand-written data set [17]. A summary of such data sets is given in Table 1. For the unsupervised feature selection, we additionally generate synthetic data sets that are explained below.

Synthetic data sets: We generated three data sets, denoted by S1, S2, and S3. Each data set has 30 features: 10 informative denoted by $(X_1, X_2, ..., X_{10})$, 10 nonlinear redundant $(X_{11}, X_{12}, ..., X_{20})$, and 10 linearly redundant $(X_{21}, X_{12}, ..., X_{30})$. The informative features are generated according to three distributions, one for each data set. The distribution for S1 is $N(0, \mathbf{I}_{10})$, for S2 is uniform distribution over $[-1,1]^{10}$, and for S 3 is uniform distribution over $\{-1,1\}^{10}$. Each nonlinear redundant feature is generated from $X_j = 3X_{i_1}X_{i_2}X_{i_3}$, where $j = 11, 12, ..., 20$, and i_1, i_2, i_3 are randomly and uniformly selected from $\{1, 2, ..., 10\}$. The linearly redundant features are generated from $X_j = \sum_{l=1}^{5} a_{j,l}X_{i_l}$, where i_l's are selected randomly from $\{1, 2, ..., 10\}$ and $a_{j,l} \sim \text{Unif}(0,1)$. We use the above redundancy model for each data set. For the sake of performance comparison, we add a labeling to the above data sets. However, the labels are not revealed to the algorithms. We generate a fixed but randomly generated labeling function $f(\mathbf{X})$ on $\mathbb{R}^{10}$. This function is the sign of the following random multi-polynomial in $\mathbb{R}^{10}$: $f(\mathbf{x}) = \mathsf{sign}\Big[\prod_{1\leqslant j\leqslant 3}\big(b_{0,j} + \sum_{1\leqslant i\leqslant 10} b_{i,j}x_i\big)\Big]$, where $b_{i,j} \sim \text{Unif}(0,1)$ and mutually independent.

Table 1: Properties of the tested data sets.

Data set	S1	S2	S3	USPS	Isolet	HAPT	Sonar	COIL20	Wine
Features	30	30	30	256	617	561	60	1024	13
Samples	1000	1000	1000	9298	1560	10299	208	1440	119
Classes	2	2	2	10	26	12	2	20	2

Unsupervised setting: We compare the performance of UFFS with Laplacian Score (LS) [14], MCFC [5], and UDFS [31] on the real and the synthetic data sets. The labels are not revealed to the algorithms, but used for measuring the performances. Features are randomly ordered, so that the initial ordering would not affect the experiments' outcomes. Contrary to other algorithms, UFFS does not rank the feature; instead it outputs a set of indices as the non-redundant features. We run UFFS three times: first with $t = 1$, $m = d$, second, with $t = 2, m = 50$ but on the selected features from the first run, and third, with $t = 3, m = 30$ but on the selected features from the second run. For each experiment, let k denote the number of the selected features by UFFS at the third run. For comparing the performance to the ranking algorithms, we select only the k features with the highest rank. Once the features are selected by each unsupervised algorithm, we reveal the samples of the selects features with the labels to a classifier and compute its prediction accuracy. A support vector machine (SVM) classifier with radial basis function as kernel is employed for all the studies. We perform a 5-fold cross validation using this classifier and on the entire data set. Implementation details are provided in our supplementary materials.

Table 2: Comparison of unsupervised algorithms.

	S1	S2	S3	USPS	Isolet	HAPT	Sonar	COIL20	Wine
No FS	77.9	75.0	87.0	97.3	92.8	97.1	86.5	98.8	98.3
UFFS k	11	12	11	93	309	88	8	331	7
UFFS	**80.3**	**76.8**	**86.2**	**97.0**	**91.7**	**95.6**	**81.3**	98.8	97.5
LS	55.1	61.2	71.0	95.6	88.6	89.8	77.4	**98.9**	97.5
MCFC	56.6	59.0	65.8	93.9	90.1	94.5	77.4	94.0	**99.2**
UDFS	64.0	60.6	64.3	80.8	90.2	78.0	77.9	98.0	98.3

Table 2 shows the average of the resulted classification accuracies for each algorithm. The second row is the resulted accuracy without any feature selection. The third row is k which is the number of non-redundant features declared by the UFFS. Observe that, in synthetic data sets, k is very close to 10 which is the actual number of non-redundant features. The resulted accuracy by the UFFS is very close or greater than the accuracy without feature selection. This implies that the UFFS detects almost all the redundant features. Further, it significantly outperforms other algorithms in the synthetic and many real data sets. This result shows that the UFFS performs well on data sets with nonlinear redundancies.

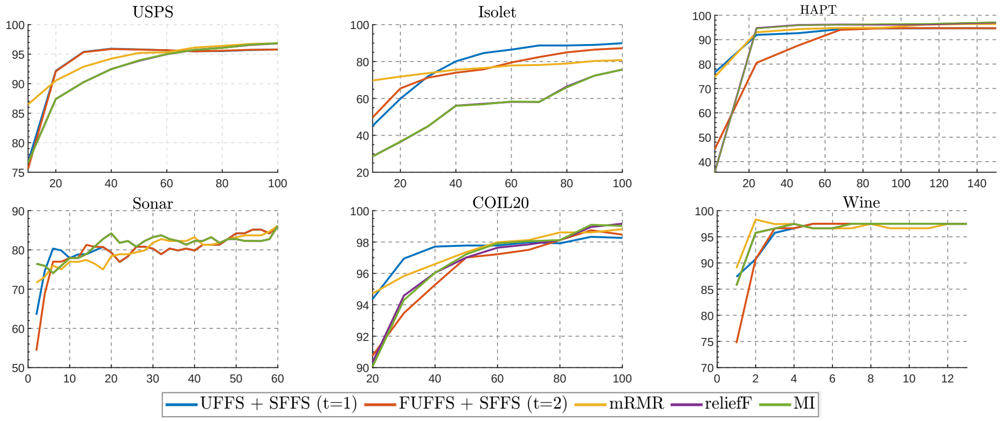

Figure 1: Classification accuracy (y-axis) versus number of selected features (x-axis).

Supervised setting: In this part, we compare SFFS with ReliefF [19], mRMR [27], and MI [22] on the real data sets in Table 1. We first run an UFFS ($t = 3$) to extract the non-redundant features before running the supervised versions SFFS ($t = 1$ and $t = 2$). As a performance measure, we perform a 5-fold cross validation with feature selection and the SVM classifier described above in a pipeline. In our supplementary materials, we explain the implementation details. Figure 1 shows the average classification accuracy for various values of selected features (k). It is observed that our SFFS improves upon other methods on some ranges of k and has comparable performance on the other values of k, but we have reduced computational complexity. For instance, in Isolet data set, we observe a dominant performance by our SFFS for $k > 40$ as compared to other algorithms. In COIL20, we observe a notiable performance improvement for $k \in [25, 50]$. Note that SFFS with $t = 1$ and $t = 2$ are overlapping in these data sets and for many values of k. We note here that the SFFS with $t = 1$ has a running time linear in data size and of order $O(nd)$.

5 Broader Impact

During the past decade, with the proliferation of data-driven technologies, sophisticated learning models incorporating large dimensional data have been implemented. The focus of this work, the problem of selecting a small set of features, is one of profound importance in such a context of high-dimensional data to reduce the burden on the computational resources and to have a better interpretation of the models from the cluttered set of features.

As machine learning continues to impact decisions in societal establishments such as healthcare, the justice system, and financial institutions, there raises an urge to ensure that the learning models are interpretable and transparent in the decision process. To this end, our approach for feature selection has the potential to increase the interpretability of the learning models without causing too much reduction in performance. We provide the flexibility to balance the interpretability-completeness tradeoff with our algorithms.

Feature selection may introduce biases in machine learning tasks by restricting attention to a small feature set. The strong guarantees (both theoretical and experimental) of our methods ensure that the features selected by our algorithm represent the entire dataset with good accuracy and minimizes biases in feature selection. We also characterize the limitations of our algorithm with non-asymptotic results to determine the number of samples required to achieve certain levels of accuracy.

In addition to interpretability, our work has the potential to reduce the cost of collecting data and running certain machine learning models. For instance, in the healthcare systems medical diagnoses often involve several examinations some of which are expensive and not covered by typical healthcare plans. Our feature selection algorithms can be used to choose the most characterizing subset of the diagnosing tests without compromising much on the prediction accuracy, and later make clinical decisions by restricting to the selected tests.

In our settings, we did not directly address the fairness in feature selection. Thus, there is a possibility that our algorithms are not sensible enough to certain ethical matters and more work is needed in this direction.

References

[1] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967.

[2] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, July 1994.

[3] E. Blais, R. O'Donnell, and K. Wimmer. Polynomial regression under arbitrary product distributions. *Machine learning*, 80(2-3):273–294, 2010.

[4] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271, 1997.

[5] D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 333–342, 2010.

[6] N. Cesa-Bianchi, Y. Mansour, and O. Shamir. On the complexity of learning with kernels. In *Conference on Learning Theory*, pages 297–325, 2015.

[7] G. Chastaing, F. Gamboa, C. Prieur, et al. Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012.

[8] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan. Kernel feature selection via conditional covariance minimization. In *Advances in Neural Information Processing Systems*, pages 6946–6955, 2017.

[9] T. A. Courtade and G. R. Kumar. Which Boolean functions maximize mutual information on noisy inputs? *IEEE Trans. Inf. Theory*, 60(8):4515–4525, 2014.

[10] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

[11] D. Dua and C. Graff. UCI machine learning repository, 2017.

[12] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Lecture Notes in Computer Science*, pages 63–77. Springer Berlin Heidelberg, 2005.

[13] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[14] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *Advances in neural information processing systems*, pages 507–514, 2006.

[15] M. Heidari, S. S. Pradhan, and R. Venkataramanan. Boolean functions with biased inputs: Approximation and noise sensitivity. In *Proc. IEEE Int. Symp. Information Theory (ISIT)*, pages 1192–1196, July 2019.

[16] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

[17] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[18] G. Kalai. Noise sensitivity and chaos in social choice theory. Technical report, Hebrew University, 2005.

[19] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In W. R. Swartout, editor, *Proceedings of the 10th National Conference on Artificial Intelligence, San Jose, CA, USA, July 12-16, 1992*, pages 129–134. AAAI Press / The MIT Press, 1992.

[20] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2):273–324, 1997.

[21] D. Koller and M. Sahami. Toward optimal feature selection. Technical report, Stanford InfoLab, 1996.

[22] A. Kraskov, H. Stögbauer, and P. Grassberger. Erratum: Estimating mutual information [phys. rev. e 69, 066138 (2004)]. *Physical Review E*, 83(1):019903, 2011.

[23] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

[24] E. Mossel, R. O'Donnell, and R. A. Servedio. Learning functions of k relevant variables. *J. Comput. Syst. Sci*, 69(3):421–434, 2004.

[25] E. Mossel, R. O'Donnell, and R. P. Servedio. Learning juntas. In *Proc. ACM Symp. on Theory of Computing*, pages 206–212, 2003.

[26] R. O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[27] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[28] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical modelling and computational experiments*, 1(4):407–414, 1993.

[29] J. R. Vergara and P. A. Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24(1):175–186, 2014.

[30] R. d. Wolf. *A Brief Introduction to Fourier Analysis on the Boolean Cube*. Number 1 in Graduate Surveys. Theory of Computing Library, 2008.

[31] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou. L2, 1-norm regularized discriminative feature selection for unsupervised. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[32] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, 5(Oct):1205–1224, 2004.