On the Complexity of Discrete Feature Selection for Optimal Classification

Jose M. Peña and Roland Nilsson

Abstract—Consider a classification problem involving only discrete features that are represented as random variables with some prescribed discrete sample space. In this paper, we study the complexity of two feature selection problems. The first problem consists in finding a feature subset of a given size k that has minimal Bayes risk. We show that for any increasing ordering of the Bayes risks of the feature subsets (consistent with an obvious monotonicity constraint), there exists a probability distribution that exhibits that ordering. This implies that solving the first problem requires an exhaustive search over the feature subsets of size k. The second problem consists in finding the minimal feature subset that has minimal Bayes risk. In the light of the complexity of the first problem, one may think that solving the second problem requires an exhaustive search over all the feature subsets. We show that, under mild assumptions, this is not true. We also study the practical implications of our solutions to the second problem.

Index Terms—Feature evaluation and selection, Classifier design and evaluation, Machine learning.

1 Introduction

Consider a classification problem involving only discrete features that are represented as random variables with some prescribed discrete sample space. The Bayes classifier over a feature subset is the classifier that outputs the most likely class conditioned on the state of the feature subset. Let the Bayes risk of a feature subset be the error probability of the Bayes classifier over that feature subset. Obviously, every ordering of the Bayes risks of the feature subsets that is possible (i.e. there exists a probability distribution that exhibits that ordering) must comply with the following monotonicity constraint: The supersets of a feature subset cannot have larger Bayes risks than the subset.

In this paper, we study the complexity of two feature selection problems. The first problem consists in finding a feature subset of a given size k that has minimal Bayes risk. We call this problem the k-optimal problem. In Section 3, we prove that any increasing ordering of the Bayes risks of the feature subsets that is consistent with the monotonicity constraint is possible. This implies that solving the k-optimal problem requires an exhaustive search over the feature subsets of size k. As we discuss later, our result strengthens the results in [1, Theorem 1], [11, page 108] and [2, Theorem 32.1].

The second problem that we study in this paper consists in finding the minimal feature subset that has minimal Bayes risk. We call this problem the minimal-optimal problem. One may think that if solving the k-optimal problem requires an exhaustive search over the feature subsets of size k, then solving the minimal-optimal problem requires an exhaustive search over all the feature subsets. We show in Section 4 that, under mild assumptions, this is not true: The minimal-optimal problem can be solved by a backward search, or even without any search by applying a characterization of the solution that we derive. As we discuss later, our result strengthens the result in [5, page 593].

The two methods that we propose to solve the minimal-optimal problem build upon the assumption that the probability distribution over the features and the class is known. In practice, however, this probability distribution is typically unknown and only a finite sample from it is available. We show in Section 5 that our methods can be adapted to finite samples so that they solve the minimal-optimal problem in the large sample limit.

Although the k-optimal problem has received some attention in the literature, the minimal-optimal problem has undoubtedly received much more attention. Therefore, we believe that researchers and practitioners will find more relevant our complexity analysis of the minimal-optimal problem than that of the k-optimal problem. All in all, we believe that both analyses contribute to advance the understanding of feature selection.

2 Preliminaries

Let the set of discrete random variables $X = (X_1, \ldots, X_n)$ represent the features and the discrete random variable Y the class. Assume that every random variable in (X, Y) has a finite sample space of cardinality greater than one. For simplicity, assume that the sample space of every random variable in (X, Y) are the integer numbers $0, 1, \ldots$. For simplicity, we use the juxtaposition of sets to represent their union. For instance, given $S, T \subseteq X$, ST means $S \cup T$. We use $\neg S$ to denote $X \setminus S$.

- *Jose M. Peña is with the Department of Computer and Information Science, Linköping University, Sweden*
  E-mail: jospe@ida.liu.se
- *Roland Nilsson is with the Department of Systems Biology, Harvard Medical School, USA*
  E-mail: nilsson@chgr.mgh.harvard.edu

We use upper-case letters to denote random variables and the same letters in lower-case to denote their states. For instance, s denotes a state of S, st a state of ST, and $\neg s$ a state of $\neg S$. In the expressions $S = 0$ and $s = 0$, 0 represents a vector of zeroes. The expression $s \geq 1$ means that every component of s is greater or equal to 1. We use low-case p to denote a probability distribution, and upper-case P to denote the probability of an event.

The following definitions are taken from [2]. A classifier over X is a function $g : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the sample spaces of X and Y, respectively. The risk of a classifier $g(X)$, denoted as $R(g(X))$, is the error probability of $g(X)$, i.e. $R(g(X)) = P(g(X) \neq Y)$. $R(g(X))$ can also be written as

$$R(g(X)) = \sum_{x,y} p(x,y) 1_{\{g(x) \neq y\}} = 1 - \sum_{x,y} p(x,y) 1_{\{g(x) = y\}}.$$

The Bayes classifier over X, denoted as $g^*(X)$, is the classifier that outputs the most likely class a posteriori, i.e. $g^*(x) = arg\,max_y\, p(y|x)$. Interestingly, the Bayes classifier is optimal. That is, the risk of the Bayes classifier, known as the Bayes risk of X, is minimal [2, Theorem 2.1]. If Y is binary, the Bayes risk of X can be written as

$$R(g^*(X)) = \sum_{x} \min_{y}\ p(x,y).$$

An inducer is a function $I : (\mathcal{X} \times \mathcal{Y})^l \rightarrow \mathcal{G}$, where $\mathcal{G}$ is a set of classifiers over X. An inducer is universally consistent if for every $\varepsilon > 0$, $p(|R(I(D^l)) - R(g^*(X))| > \varepsilon) \rightarrow 0$ as $l \rightarrow \infty$. An estimator of $R(I(D^l))$, denoted as $\widehat{R}(I(D^l))$, is consistent if for every $\varepsilon > 0$, $p(|\widehat{R}(I(D^l)) - R(I(D^l))| > \varepsilon) \rightarrow 0$ as $l \rightarrow \infty$.

3 On the k-Optimal Problem

In this section, we show that solving the k-optimal problem requires an exhaustive search over the subsets of X of size k. First, we prove in the theorem below that any increasing ordering of the Bayes risks of the subsets of X that is consistent with the monotonicity constraint is possible, no matter the cardinality of the sample space of each random variable in (X, Y).

Theorem 1: Let $S_1, \ldots, S_{2^n}$ be an ordering of the subsets of X such that $i < j$ for all $S_j \subset S_i$. Then, there exists a probability distribution p over (X, Y) such that $R(g^*(S_1)) < \ldots < R(g^*(S_{2^n}))$.

Proof: We construct p as follows. We set $p(x, y) = 0$ for all $y \notin \{0, 1\}$ and x. This allows us to treat Y hereinafter as if it were binary. We do so. We set

$$p(X = 0, Y = 0) = \alpha \in (0.5, 1), \quad (1)$$
$$p(X = 0, Y = 1) = 0, \quad (2)$$
$$p(x, Y = 0) = 0 \text{ for all } x \neq 0. \quad (3)$$

Note that $S_1 = X$ and, thus, that $R(g^*(S_1)) = 0$ by Equations 2 and 3. Now, consider the subsets $S_2, \ldots, S_{2^n}$ in order, i.e. consider a subset only after having considered its predecessors in the ordering. Let S_i denote the next subset to consider. Then, $\min_y p(S_i = 0, y) = p(S_i = 0, Y = 1)$ by Equation 1. Furthermore, $p(s_i, Y = 0) = 0$ for all $s_i \neq 0$ by Equation 3 and, thus, $\min_y p(s_i, y) = 0$ for all $s_i \neq 0$. Consequently,

$$R(g^*(S_i)) = \sum_{s_i} \min_{y}\ p(s_i, y) = p(S_i = 0, Y = 1).$$

Furthermore,

$$\begin{aligned} p(S_i = 0, Y = 1) &= \sum_{\neg s_i} p(S_i = 0, \neg s_i, Y = 1) \\ &= \sum_{\{S_j\,:\,S_j \supseteq S_i\}} \sum_{\{\neg s_j\,:\,\neg s_j \geq 1\}} p(S_j = 0, \neg s_j, Y = 1) \\ &= \sum_{\{\neg s_i\,:\,\neg s_i \geq 1\}} p(S_i = 0, \neg s_i, Y = 1) \\ &+ \sum_{\{S_j\,:\,S_j \supset S_i\}} \sum_{\{\neg s_j\,:\,\neg s_j \geq 1\}} p(S_j = 0, \neg s_j, Y = 1). \quad (4) \end{aligned}$$

Since we have already considered $S_1, \ldots, S_{i-1}$, we have already set the probabilities in the second summand in the last equation above as well as those in $R(g^*(S_{i-1}))$. Then, we can now set the probabilities in the first summand in the last equation above to some positive value so that $R(g^*(S_{i-1})) < R(g^*(S_i))$.

Since setting $p(S_i = 0, \neg s_i, Y = 1)$ for all i and $\neg s_i \geq 1$ so that $\sum_i \sum_{\neg s_i \geq 1} p(S_i = 0, \neg s_i, Y = 1) = 1 - \alpha$ is not straightforward, one can initially assign them positive values satisfying the constraints above and, then, normalize them by dividing them by $\frac{\sum_i \sum_{\neg s_i \geq 1} p(S_i = 0, \neg s_i, Y = 1)}{1 - \alpha}$. □

The theorem above implies that no non-exhaustive search method over the subsets of X of size k can always solve the k-optimal problem: For any subset S of X of size k that is not considered by a non-exhaustive search method, there exists a probability distribution such that the Bayes risk of S is smaller than the Bayes risks of the rest of subsets of X of size k. Furthermore, it follows from the proof above that the Bayes risk of S can be made arbitrarily smaller than the Bayes risks of the rest of subsets of X of size k. Therefore, a non-exhaustive search method can perform arbitrarily bad.

The theorem above strengthens the results in [1, Theorem 1], [11, page 108] and [2, Theorem 32.1]. In particular, [1, Theorem 1] and [11, Theorem 1] prove the same result as the theorem above by constructing a continuous probability distribution that exhibits the desired behavior. Therefore, in these works the features are assumed to be continuous. It is mentioned in [11, page 108] that the result also holds for discrete features: It suffices to find a sufficiently fine discretization of the continuous probability distribution constructed. An alternative proof of the result is provided in [2, Theorem 32.1], where the authors directly construct a discrete probability distribution that exhibits the desired behavior. As a matter of fact, the authors do not only construct the discrete probability distribution but also the sample space of the features. Consequently, the three

papers cited prove the same result as the theorem above for some discrete sample space of the features. However, this sample space may not coincide with the prescribed one. In other words, the three papers cited prove the result for some discrete sample space of the features whereas the theorem above proves it for any discrete sample space of the features, because the sample space of each random variable in (X, Y) can have any cardinality, as long as this is finite and greater than one.

We prove below another interesting result: Some not strictly increasing orderings of the Bayes risks of the subsets of X are impossible though they comply with the monotonicity constraint. This result will be of much help in the next section. We prove first an auxiliary theorem.

Theorem 2: Let p be a probability distribution over (X, Y). Let S and T denote two disjoint subsets of X. If $p(st) > 0$ and $p(Y|st)$ has a single maximum for all st, then $R(g^*(ST)) = R(g(S))$ iff $g^*(ST) = g(S)$.

Proof:

$$
\begin{aligned}
&R(g(S)) - R(g^*(ST)) \\
&= 1 - \sum_{s,y} p(s,y) 1_{\{g(s)=y\}} - 1 + \sum_{st,y} p(st,y) 1_{\{g^*(st)=y\}} \\
&= \sum_{st,y} p(st,y) 1_{\{g^*(st)=y\}} - \sum_{s,y} (\sum_{t} p(st,y)) 1_{\{g(s)=y\}} \\
&= \sum_{st,y} p(st,y) (1_{\{g^*(st)=y\}} - 1_{\{g(s)=y\}}) \\
&= \sum_{\{st\,:\,g^*(st) \neq g(s)\}} p(st, g^*(st)) - p(st, g(s)) \\
&= \sum_{\{st\,:\,g^*(st) \neq g(s)\}} p(st)(p(g^*(st)|st) - p(g(s)|st)).
\end{aligned}
$$

Thus, if $R(g(S)) - R(g^*(ST)) = 0$ then $g^*(st) = g(s)$ for all st, because $p(st) > 0$ and $p(g^*(st)|st) > p(g(s)|st)$ by assumption. On the other hand, if $R(g(S)) - R(g^*(ST)) \neq 0$ then $g^*(st) \neq g(s)$ for some st. □

The assumption that $p(Y|st)$ has a single maximum for all st in the theorem above is meant to guarantee that no tie occurs in $g^*(ST)$.

Theorem 3: Let p be a probability distribution over (X, Y). Let S, T and U denote three mutually disjoint subsets of X. If $p(stu) > 0$ and $p(Y|stu)$ has a single maximum for all stu, then $R(g^*(STU)) = R(g^*(ST)) = R(g^*(SU))$ iff $R(g^*(STU)) = R(g^*(S))$.

Proof: The if part is immediate due to the monotonicity constraint. To prove the only if part, assume to the contrary that $R(g^*(STU)) < R(g^*(S))$. Then, $g^*(s't'u') \neq g^*(s't''u'')$ for some $s't'u'$ and $s't''u''$ such that $t' \neq t''$ or $u' \neq u''$ because, otherwise, for any s, $g^*(stu)$ is constant for all tu and, thus, $g^*(STU)$ reduces to a classifier $g(S)$ such that $R(g(S)) = R(g^*(STU))$ by Theorem 2. This is a contradiction because $R(g(S)) = R(g^*(STU)) < R(g^*(S))$. Then, $g^*(s't'u') \neq g^*(s't''u'')$.

Since $R(g^*(STU)) = R(g^*(ST))$, then $g^*(s't'u') = g^*(s't'u'') = g^*(s't')$ due to Theorem 2. Likewise, since $R(g^*(STU)) = R(g^*(SU))$, then $g^*(s't'u'') = g^*(s't''u'') = g^*(s'u'')$ due to Theorem 2. However, these equalities imply that $g^*(s't'u') = g^*(s't''u'')$ which is a contradiction. □

Consequently, under the assumptions in the theorem above, some not strictly increasing orderings of the Bayes risks of the subsets of X are impossible though they comply with the monotonicity constraint, e.g. $R(g^*(STU)) = R(g^*(ST)) = R(g^*(SU)) < R(g^*(S))$.

4 On the Minimal-Optimal Problem

In this section, we prove that, under mild assumptions on the probability distribution $p(X, Y)$, solving the minimal-optimal problem does not require an exhaustive search over the subsets of X. Specifically, the assumptions are that $p(x) > 0$ and $p(Y|x)$ has a single maximum for all x. The former assumption implies that there are not bijective transformations between feature subsets. To see it, it suffices to note that if there were a bijective transformation between two feature subsets, then the probability that one of the feature subsets is in a state different from the one dictated by the bijective transformation would be zero, which contradicts the assumption of strict positivity. The latter assumption implies that no tie occurs in $g^*(X)$.

Before proving the main result of this section, we prove that the solution to the minimal-optimal problem is unique.

Theorem 4: Let p be a probability distribution over (X, Y). If $p(x) > 0$ and $p(Y|x)$ has a single maximum for all x, then the solution to the minimal-optimal problem is unique.

Proof: A solution to the minimal-optimal problem is any minimal feature subset that has minimal Bayes risk. It is obvious that one such subset always exists. Assume to the contrary that there exist two such subsets, say S^* and S_*. Then, $R(g^*(X)) = R(g^*(S^*)) = R(g^*(S_*))$. Since $S^* = (S^* \cap S_*)(S^* \setminus S_*)$ and $S_* \subseteq (S^* \cap S_*)(X \setminus S^*)$, the monotonicity constraint implies that $R(g^*(X)) = R(g^*((S^* \cap S_*)(S^* \setminus S_*))) = R(g^*((S^* \cap S_*)(X \setminus S^*)))$. Since $X = (S^* \cap S_*)(S^* \setminus S_*)(X \setminus S^*)$, $R(g^*(X)) = R(g^*(S^* \cap S_*))$ by Theorem 3. However, this contradicts that S^* and S_* are minimal with respect to having minimal Bayes risk. □

Hereinafter, S^* denotes the unique solution to the minimal-optimal problem. We prove below that the backward search (BS) method in Table 1 solves the minimal-optimal problem. Let S denote the estimate of S^*. BS first initializes S to X. Then, it chooses any $X_i \in S$ such that $R(g^*(S \setminus X_i)) = R(g^*(S))$ and removes it from S. The method keeps removing features from S while possible.

Theorem 5: Let p be a probability distribution over (X, Y). If $p(x) > 0$ and $p(Y|x)$ has a single maximum for all x, then the backward search (BS) method in Table 1 solves the minimal-optimal problem.

Proof: Assume that no feature can be removed from S and, thus, that BS halts. At that point, S has minimal Bayes risk, i.e. $R(g^*(S)) = R(g^*(X))$, by how BS works.

TABLE 1
Backward search (BS) method

```
1   S = X
2   i = 1
3   while i ≤ n do
4     if X_i ∈ S then
5       if R(g*(S \ X_i)) = R(g*(S)) then
6         S = S \ X_i
7         i = 1
8       else
9         i = i + 1
10    else
11      i = i + 1
12  return S
```

TABLE 2
One-shot (OS) method

```
1   S = X
2   i = 1
3   while i ≤ n do
4     if R(g*(¬X_i)) = R(g*(X)) then
5       S = S \ X_i
6     i = i + 1
7   return S
```

Moreover, S is minimal with respect to having minimal Bayes risk. To see it, assume to the contrary that there exists some $T \subset S$ such that $R(g^*(T)) = R(g^*(S)) = R(g^*(X))$. Then, $R(g^*(S \setminus X_i)) = R(g^*(S))$ with $X_i \in S \setminus T$, due to the monotonicity constraint because $T \subset S \setminus X_i \subset S$. However, this contradicts that no more features can be removed from S.

Finally, note that if S is minimal with respect to having minimal Bayes risk, then $S = S^*$ by Theorem 4. □

If X contains more than two features, then the theorem above implies that solving the minimal-optimal problem does not require an exhaustive search over the subsets of X. Recall from the previous section that solving the k-optimal problem requires an exhaustive search over the subsets of X of size k. One may think that such an exhaustive search would not be necessary if one makes the same assumptions as in the theorem above. Unfortunately, this is not true: The probability distribution constructed in the proof of Theorem 1 satisfies those assumptions, because of Equations 1 and 3 and because the probabilities in the first summand of Equation 4 are set to positive values.

BS removes features from S in certain order: It always removes the feature with the smallest index that satisfies the conditions in lines 4 and 5. However, removing any other feature that satisfies these conditions works equally well, because the proof of the theorem above does not depend on this question. However, the study of this question led us to an interesting finding: The features that satisfy the conditions in lines 4 and 5 in the first iteration of BS, i.e. when $S = X$, are exactly the features that will be removed from S in all the iterations. The theorem below proves this fact.

Theorem 6: Let p be a probability distribution over (X, Y). If $p(x) > 0$ and $p(Y|x)$ has a single maximum for all x, then $X_i \in S^*$ iff $R(g^*(\neg X_i)) \neq R(g^*(X))$ or, alternatively, $X_i \in S^*$ iff $g^*(\neg X_i) \neq g^*(X)$.

Proof: It suffices to prove the first equivalence in the theorem, because the second follows from the first by Theorem 2.

Consider any $X_i \notin S^*$. By Theorem 5, BS removes X_i from S at some point. At that point, $R(g^*(S \setminus X_i)) = R(g^*(S)) = R(g^*(X))$ by how BS works. Moreover, $R(g^*(S \setminus X_i)) = R(g^*(X))$ implies $R(g^*(\neg X_i)) = R(g^*(X))$ by the monotonicity constraint since $S \setminus X_i \subset \neg X_i \subset X$.

Now, consider any $X_i \in S^*$. By Theorem 5, $X_i \in S$ when BS halts. At that point, $R(g^*(S \setminus X_i)) \neq R(g^*(S)) = R(g^*(X))$ by how BS works. Moreover, $R(g^*(X)) \neq R(g^*(S \setminus X_i))$ implies $R(g^*(X)) = R(g^*(S)) \neq R(g^*(\neg X_i))$ by Theorem 3 and the fact that $R(g^*(X)) = R(g^*(S))$. □

The theorem above implies that the minimal-optimal problem can be solved without performing a search over the subsets of X: It suffices to apply the characterization of S^* in the theorem. We call this method the one-shot (OS) method. Tables 2 shows its pseudocode.

It is worth mentioning that an unproven characterization of S^* is proposed in [5, page 593]. Although in [5] the features are assumed to be continuous, the authors claim that their characterization also applies to discrete features. Specifically, the authors state without proof that $X_i \in S^*$ iff $P(g^*(\neg X_i, X_i) \neq g^*(\neg X_i, X_i')) > 0$ where X_i and X_i' are two representations of the i-th feature that are independent and identically distributed conditioned on $\neg X_i$. Intuitively, one can think of X_i and X_i' as two identical sensors measuring the state of the i-th feature. Note that the only independence assumed is that between X_i and X_i' conditioned on $\neg X_i$ and, thus, no independence is assumed between the random variables in $(\neg X_i, X_i, Y)$ or between the random variables in $(\neg X_i, X_i', Y)$. The theorem below proves the correctness of this alternative characterization of S^*.

Theorem 7: Let p be a probability distribution over (X, Y). If $p(x) > 0$ and $p(Y|x)$ has a single maximum for all x, then $X_i \in S^*$ iff $P(g^*(\neg X_i, X_i) \neq g^*(\neg X_i, X_i')) > 0$ where X_i and X_i' are independent and identically distributed conditioned on $\neg X_i$.

Proof: Let p' denote the probability distribution over $(\neg X_i, X_i', Y)$. That X_i and X_i' are independent and identically distributed conditioned on $\neg X_i$ implies that, for any $\neg x_i$, $p'(\neg x_i, X_i' = z) = p(\neg x_i, X_i = z)$ for all z state of X_i and X_i'. We represent this coincidence by the expression $p' = p$.

By Theorem 6, it suffices to prove that $P(g^*(\neg X_i, X_i) \neq g^*(\neg X_i, X_i')) = 0$ iff $g^*(\neg X_i) = g^*(X)$. We first prove that $P(g^*(\neg X_i, X_i) \neq g^*(\neg X_i, X_i')) = 0$ iff, for any $\neg x_i$, $g^*(\neg x_i, x_i)$ is constant for all x_i. The if part is immediate because $p' = p$ implies that, for any $\neg x_i$, $g^*(\neg x_i, x_i')$ is also constant for all x_i'. To prove the only if part, assume to the contrary that, for some $\neg x_i$, $g^*(\neg x_i, x_i)$ is not constant for all x_i.

Then, $p' = p$ implies that $g^*(\neg x_i, x_i) \neq g^*(\neg x_i, x_i')$ for some state $\neg x_i x_i x_i'$. Note that this state has probability $p(\neg x_i)p(x_i|\neg x_i)p'(x_i'|\neg x_i)$, which is greater than zero because $p(x) > 0$ for all x and $p' = p$. Then, $P(g^*(\neg X_i, X_i) \neq g^*(\neg X_i, X_i')) > 0$ which is a contradiction. Consequently, $P(g^*(\neg X_i, X_i) \neq g^*(\neg X_i, X_i')) = 0$ iff, for any $\neg x_i$, $g^*(\neg x_i, x_i)$ is constant for all x_i.

Moreover, for any $\neg x_i$, $g^*(\neg x_i, x_i)$ is constant for all x_i iff $g^*(X)$ coincides with some classifier $g(\neg X_i)$. We now prove that the latter statement is true iff $g^*(X) = g^*(\neg X_i)$. The if part is trivial. To prove the only if part, assume to the contrary that $g^*(X)$ coincides with some classifier $g(\neg X_i)$ such that $g(\neg X_i) \neq g^*(\neg X_i)$. Then, $R(g^*(\neg X_i)) < R(g(\neg X_i)) = R(g^*(X))$ by Theorem 2. However, this contradicts the monotonicity constraint. □

Finally, note that our characterization of S^* in Theorem 6 as $X_i \in S^*$ iff $g^*(\neg X_i) \neq g^*(X)$ resembles the definition of strongly relevant features introduced in [4, Definition 5]: X_i is strongly relevant iff $p(Y|\neg X_i) \neq p(Y|X)$. Note, however, that our characterization of S^* involves the Bayes classifier whereas the definition of strongly relevant involves the posterior distribution of Y. This is why S^* does not coincide with the set of strongly relevant features in general, as the following example illustrates.

Example 1: Let X and Y be two binary random variables. Let $p(x) > 0$ and $p(Y = 0|x) = x/3$ for all x. Then, X is strongly relevant though $X \notin S^*$, because it affects the posterior distribution of Y but not enough so as to affect the Bayes classifier, which is $g^*(x) = 1$ for all x.

It should be noted, however, that every feature in S^* is strongly relevant. See [5, Theorem 8] for a proof of this statement for continuous features. The proof also applies to discrete features. Yet another feature subset of importance in classification is the so-called Markov boundary introduced in [6, page 97]: The Markov boundary is the minimal feature subset M such that $p(Y|M) = p(Y|X)$. When $p(x) > 0$ for all x, the Markov boundary coincides with the strongly relevant features. See [5, Theorem 10] for a proof of this statement for continuous features. The proof also applies to discrete features. Therefore, S^* does not coincide with the Markov boundary in general either.

When facing a classification problem for the first time, the practitioner should decide whether it will suffice to predict a class label for each new instance or whether it will also be needed to assess the confidence in the class label predicted. Some may say that this is not a decision the practitioner can make but an intrinsic characteristic of the classification problem at hand. In any case, the practitioner should determine the feature subset on which to build the classifier. As we have discussed above, S^* and the Markov boundary M do not coincide in general. Therefore, it is crucial to choose the right feature subset in order to solve the classification problem optimally. If only the label of the class predicted is needed when classifying a new instance, then one should go for S^* because it is the minimal feature subset that allows to build a classifier with minimal risk, i.e. $R(g^*(S^*)) = R(g^*(X))$. If a measure of the confidence in the class label predicted is required, then one should go for M, which as mentioned above coincides with the strongly relevant features when $p(x) > 0$ for all x, because it is the minimal feature subset such that $p(Y|M) = p(Y|X)$.

5 BS AND OS IN PRACTICE

It is shown in [9] that for a feature selection algorithm to solve a feature selection problem, the algorithm should be custom designed for specific classes of classifiers and performance measures. Of course, these classes of classifiers and performance measures must be aligned with the feature selection problem at hand. Clearly, these conditions are satisfied by BS and OS and the feature selection problem that they address, i.e. the minimal-optimal problem: The algorithms and the problem are defined in terms of the same classifier (the Bayes classifier) and performance measure (the risk of a classifier). We have proved in Theorems 5 and 6 that BS and OS solve the minimal-optimal problem. Recall that BS and OS assume that one has access to the probability distribution $p(X, Y)$ so that the Bayes risks of different feature subsets can be computed. Unfortunately, in practice, one does not have access to this probability distribution but to a sample from it of finite size l, here denoted as D^l. Therefore, in order to use BS and OS in practice, we make the following modifications:

- We replace the condition $R(g^*(S \setminus X_i)) = R(g^*(S))$ in Table 1 with the condition $\widehat{R}(I(D^l_{S\setminus X_i})) \leq \widehat{R}(I(D^l_S)) + \tau$, and
- we replace the condition $R(g^*(\neg X_i)) = R(g^*(X))$ in Table 2 with the condition $\widehat{R}(I(D^l_{\neg X_i})) \leq \widehat{R}(I(D^l)) + \tau$,

where I is an inducer, $\widehat{R}$ is a risk estimator, D^l_T is the data in D^l for the features in $T \subseteq X$, and $\tau > 0$ is a parameter that enables to discard X_i if this does not harm performance significantly. This parameter enables to control the trade-off between precision and recall, i.e. the smaller τ the higher recall but the lower precision. We call the methods resulting from the two modifications above, respectively, FBS and FOS, where the F stands for finite sample.

As we have discussed above, if FBS and FOS are to solve the minimal-optimal problem, then I and $\widehat{R}$ must be aligned with the Bayes classifier and the risk of a classifier, respectively. A reasonable interpretation of being aligned may be that the former converge to the latter asymptotically. The theorem below proves that, under this interpretation, FBS and FOS solve the minimal-optimal problem asymptotically, i.e. the probability that they do not return S^* converges to zero as the sample size tends to infinity. We call this property of an algorithm consistency.

Theorem 8: Let p be a probability distribution over (X, Y) such that $p(x) > 0$ and $p(Y|x)$ has a single maximum for all x. If I is an universally consistent

inducer and $\widehat{R}$ is a consistent risk estimator, then there exists some $\eta > 0$ such that FBS and FOS are consistent for all $\tau \in (0, \eta)$.

Proof: The proof is a straightforward adaptation of that of [5, Theorem 11]. We start by proving the theorem for FBS. Let S_j and T_j denote the content of S and $S \setminus X_i$, respectively, when line 5 in Table 1 is executed for the j-th time. Let $\eta = 1$ if $R(g^*(T_j)) - R(g^*(S_j)) = 0$ for all j. Otherwise, let $\eta = \min_j\{R(g^*(T_j)) - R(g^*(S_j)) : R(g^*(T_j)) - R(g^*(S_j)) > 0\}$. Let $\tau \in (0, \eta)$. Since BS returns S^* by Theorem 5, if FBS does not return S^*, then there exists some feature that is in the output of FBS but not in the output of BS, or that is in the output of BS but not in the output of FBS. In other words, if FBS does not return S^*, then there exists some j such that either $R(g^*(T_j)) - R(g^*(S_j)) = 0$ whereas $\widehat{R}(I(D^l_{T_j})) - \widehat{R}(I(D^l_{S_j})) > \tau$, or $R(g^*(T_j)) - R(g^*(S_j)) \geq \eta$ whereas $\widehat{R}(I(D^l_{T_j})) - \widehat{R}(I(D^l_{S_j})) \leq \tau$. Let $\varepsilon \in (0, \min\{\tau, \eta - \tau\})$. Then,

$$
\begin{aligned}
&P(\text{FBS does not return } S^*)\\
&\leq P(\bigvee_j |R(g^*(T_j)) - R(g^*(S_j)) - (\widehat{R}(I(D^l_{T_j})) - \widehat{R}(I(D^l_{S_j})))| > \tau\\
&\vee |R(g^*(T_j)) - R(g^*(S_j)) - (\widehat{R}(I(D^l_{T_j})) - \widehat{R}(I(D^l_{S_j})))| \geq \eta - \tau)\\
&\leq P(\bigvee_j |R(g^*(T_j)) - R(g^*(S_j)) - (\widehat{R}(I(D^l_{T_j})) - \widehat{R}(I(D^l_{S_j})))| > \varepsilon)\\
&\leq P(\bigvee_j |R(g^*(T_j)) - \widehat{R}(I(D^l_{T_j}))| + |\widehat{R}(I(D^l_{S_j})) - R(g^*(S_j))| > \varepsilon)\\
&\leq P(\bigvee_j |R(g^*(T_j)) - R(I(D^l_{T_j})) + R(I(D^l_{T_j})) - \widehat{R}(I(D^l_{T_j}))|\\
&\quad + |\widehat{R}(I(D^l_{S_j})) - R(I(D^l_{S_j})) + R(I(D^l_{S_j})) - R(g^*(S_j))| > \varepsilon)\\
&\leq P(\bigvee_j |R(g^*(T_j)) - R(I(D^l_{T_j}))| + |R(I(D^l_{T_j})) - \widehat{R}(I(D^l_{T_j}))|\\
&\quad + |\widehat{R}(I(D^l_{S_j})) - R(I(D^l_{S_j}))| + |R(I(D^l_{S_j})) - R(g^*(S_j))| > \varepsilon)\\
&\leq P(\bigvee_j |R(g^*(T_j)) - R(I(D^l_{T_j}))| > \frac{\varepsilon}{4}\\
&\quad \vee |R(I(D^l_{T_j})) - \widehat{R}(I(D^l_{T_j}))| > \frac{\varepsilon}{4}\\
&\quad \vee |\widehat{R}(I(D^l_{S_j})) - R(I(D^l_{S_j}))| > \frac{\varepsilon}{4}\\
&\quad \vee |R(I(D^l_{S_j})) - R(g^*(S_j))| > \frac{\varepsilon}{4})\\
&\leq \sum_j P(|R(g^*(T_j)) - R(I(D^l_{T_j}))| > \frac{\varepsilon}{4})\\
&\quad + P(|R(I(D^l_{T_j})) - \widehat{R}(I(D^l_{T_j}))| > \frac{\varepsilon}{4})\\
&\quad + P(|\widehat{R}(I(D^l_{S_j})) - R(I(D^l_{S_j}))| > \frac{\varepsilon}{4})\\
&\quad + P(|R(I(D^l_{S_j})) - R(g^*(S_j))| > \frac{\varepsilon}{4}).
\end{aligned}
$$

Note that the four probabilities in the last expression above converge to zero for all j as l tends to infinity: The first and forth probabilities because I is universally consistent, and the second and third probabilities because $\widehat{R}$ is consistent. Consequently, $P(\text{FBS does not return } S^*)$ converges to zero as l tends to infinity.

The proof above also applies to FOS if S_j and T_j denote the content of X and $\neg X_i$, respectively, when line 4 in Table 2 is executed for the j-th time. □

Luckily, there are many universally consistent inducers and consistent risk estimators, among them some of the most commonly used inducers and risk estimators. For instance, two examples of universally consistent inducer are support vector machines [8] and the k-nearest neighbor method [2, Theorem 6.4]. Two examples of consistent risk estimator are the counting risk estimator [2, Corollary 8.1] and the cross-validation risk estimator [2, Chapter 24]. Furthermore, note that the number of iterations that FBS and FOS perform is smaller than n^2 in the case of the former and exactly n in the case of the latter, where n is the number of features in X. Therefore, the running time of FBS and FOS is polynomial in n, provided that the inducer and the risk estimator in them are also polynomial in n. For example, let the inducer be the k-nearest neighbor method run on the first half of D^l, and let the risk estimate be the counting risk estimate on the second half of D^l, i.e. the fraction of errors on the second half of D^l. This inducer and this risk estimator are polynomial in n. Consequently, FBS and FOS with this inducer and this risk estimator prove that there exist algorithms for solving the minimal-optimal problem that are both polynomial and consistent.

As discussed above, FBS may be slower than FOS: The number of iterations of the former is quadratic in n whereas the number of iterations of the latter is linear in n. However, FBS may be more reliable than FOS: Since S gets smaller with each feature discarded, the result of the check $\widehat{R}(I(D^l_{S\setminus X_i})) \leq \widehat{R}(I(D^l_S)) + \tau$ in FBS is more reliable than the result of the check $\widehat{R}(I(D^l_{\neg X_i})) \leq \widehat{R}(I(D^l)) + \tau$ in FOS. Therefore, FBS can be said to be slower but more reliable than FOS and viceversa. It is up to the practitioner to decide which of the two algorithms suits the application at hand, depending on the amount of learning data available and the running time requirements.

The theorem above provides evidence of why feature selection algorithms that run backward as FBS does, i.e. they initialize the estimate S to X and then proceed by removing features from S (e.g. [3]), usually work well in practice. All in all, FBS and FOS are not meant to be applied in practice, though they may be. Thus, their empirical evaluation is out of the scope of this paper. The reason is that the estimates $\widehat{R}(I(D^l_{S\setminus X_i}))$, $\widehat{R}(I(D^l_S))$, $\widehat{R}(I(D^l_{\neg X_i}))$ and $\widehat{R}(I(D^l))$ may be unreliable if the number of features is large and the data available scarce. We have developed FBS and FOS as a proof-by-example of the existence of time efficient and asymptotically correct

algorithms for solving the minimal-optimal problem. It is our hope that FBS and FOS provide foundation for designing algorithms that, in addition to being time efficient and asymptotically correct as FBS and FOS, are data efficient as well. We are convinced that such algorithms must run forward, i.e. initializing the estimate S to the empty set and then adding features to it until it coincides with S^*. Forward search methods are common when searching for the Markov boundary, e.g. [7], [10]. This is why we plan to investigate the conditions under which forward search methods aiming at finding S^* are asymptotically correct.

6 Conclusions

In this paper, we have reported some theoretic results on feature selection that have important practical implications. Specifically, we have proved the following theoretic results:

- Any increasing ordering of the Bayes risks of the feature subsets that is consistent with the monotonicity constraint is possible, no matter the cardinality of the sample space of the features and the class. This implies that finding the feature subset of a given size that has minimal Bayes risk requires an exhaustive search over the feature subsets of that size. Up to now, [1], [2], [11] have frequently been miscited as evidence for the intractability of this feature selection problem (recall Section 3).
- Finding the minimal feature subset that has minimal Bayes risk, i.e. S^*, is a tractable feature selection problem since it does not require an exhaustive search over feature subsets. We have proposed two algorithms to solve this problem: BS that runs backward, and OS that takes a one-shot approach based on a characterization of S^* that we have derived.

The results above are theoretic in the sense that they build upon the assumption that the probability distribution of the features and the class, i.e. $p(X, Y)$, is known. Unfortunately, in practice, one does not have access to this probability distribution but to a finite sample from it. We have adapted BS and OS to finite samples resulting in two algorithms, FBS and FOS, that converge to S^* asymptotically and whose running time is polynomial in the number of features. This result provides evidence of why feature selection algorithms that run backward as FBS does, e.g. [3], usually work well in practice. In any case, the aim of this paper was not to develop algorithms that are competitive in practice, but to demonstrate that there are principled ways of developing time efficient and asymptotically correct algorithms. We hope that our results provide foundation for developing feature selection algorithms that are not only time efficient and asymptotically correct but also data efficient and, thus, competitive in practice. We are convinced that such algorithms must run forward. We plan to investigate the assumptions that allow to develop such algorithms. Of course, the assumptions should be as mild as possible. However, it is unlikely that they will be as mild as the assumptions made to develop BS, OS, FBS and FOS, namely that $p(x) > 0$ and $p(Y|x)$ has a single maximum for all x.

Acknowledgements

This work is funded by the Swedish Research Council (ref. VR-621-2005-4202) and CENIIT at Linköping University (ref. 09.01). We thank the Associate Editor and the anonymous Referees for their insightful comments.

References

[1] Cover, T. and Van Campenhout, J. On the Possible Orderings in the Measurement Selection Problem. *IEEE Transactions on Systems, Man, and Cybernetics*, 7:657-661, 1977.

[2] Devroye, L., Györfi, L. and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[3] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, 46:389-422, 2002.

[4] Kohavi, R. and John, G. H. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97:273-324, 1997.

[5] Nilsson, R., Peña, J. M., Björkegren, J. and Tegnér, J. Consistent Feature Selection for Pattern Recognition in Polynomial Time. *Journal of Machine Learning Research*, 8:589-612, 2007.

[6] Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[7] Peña, J. M., Nilsson, R., Björkegren, J. and Tegnér, J. Towards Scalable and Data Efficient Learning of Markov Boundaries. *International Journal of Approximate Reasoning*, 45:211-232, 2007.

[8] Steinwart, I. On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research*, 2:67-93, 2001.

[9] Tsamardinos, I. and Aliferis, C. Towards Principled Feature Selection: Relevancy, Filters and Wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[10] Tsamardinos, I., Aliferis, C. F. and Statnikov, A. Algorithms for Large Scale Markov Blanket Discovery. In *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, pages 376-380, 2003.

[11] Van Campenhout, J. The Arbitrary Relation between Probability of Error and Measurement Subset. *Journal of the American Statistical Association*, 75:104-109, 1980.