Analyzing Feature Relevance for Linear Reject Option SVM using Relevance Intervals

Christina Göpfert  
Bielefeld University  
Inspiration 1, 33619 Bielefeld  
cgoepfert@techfak.uni-bielefeld.de

Jan Philip Göpfert  
Bielefeld University  
Inspiration 1, 33619 Bielefeld  
jgoepfert@techfak.uni-bielefeld.de

Barbara Hammer  
Bielefeld University  
Inspiration 1, 33619 Bielefeld  
bhammer@techfak.uni-bielefeld.de

Abstract

When machine learning is applied in safety-critical or otherwise sensitive areas, the analysis of feature relevance can be an important tool to keep the size of models small, and thus easier to understand, and to analyze how different features impact the behavior of the model. In the presence of correlated features, feature relevances and the solution to the minimal-optimal feature selection problem are not unique. One approach to solving this problem is identifying feature relevance intervals that symbolize the range of relevance given to each feature by a set of equivalent models. In this contribution, we address the issue of calculating relevance intervals – a unique representation of relevance – for reject option support vector machines with a linear kernel, which have the option of rejecting a data point if they are unsure about its label.

1 Introduction

Feature selection is a tool commonly used to increase the performance of a machine learning algorithm, to reduce computational and sample complexity, and to obtain smaller models that are easier for humans to understand. Early definitions of feature relevance were given by [7], and feature selection has been an active research area for decades [4], where popular methods include $l_1$-regularization and filter methods based on mutual information [10][11][2]. In the presence of highly correlated variables, minimal-optimal sets may not be unique and relevance cannot be represented in by a binary relevant – irrelevant distinction. Thus, popular feature selection methods do not allow for full insight in this case. This is particularly true if we aim to take into consideration different costs or practicalities of collecting each feature set, or simply need to check whether the decision behavior of highly performing models corresponds to our real world intuition. Under these goals, the problem of interest is the all-relevant problem. So far, it has received far less attention than the minimal-optimal problem, which aims for a single feature subset that allows good classification accuracy. Recent approaches for its solution include importance determination using random forests [9], combined $l_1$- and $l_2$ regularization and various forward-backward selection schemes based on different importance measures [12][3] and, most recently, a mathematical formulation for determining all possible feature relevance combinations for a linear classifier [5]. In this contribution, we extend this formulation to reject option linear classifiers.

Reject option classifiers are a deviation from typical binary classification schemes that are used when the cost of misclassification is high. If the classifier is unsure of the correct label, it declines to
make a low-certainty prediction and instead returns a “reject”, in order to allow further diagnostics or reevaluation, instead of making an uncertain and potentially damaging decision. Nearest neighbor classification with a reject option goes back to [6]. More recently, reject options have been extended to SVMs for the global two-class and local multi-class cases [1, 8].

To allow introspection into these classifiers, it is desirable to analyze which features are especially important to its decisions in general as well as on a particular data point or class. To the best of our knowledge, there is no method to unambiguously analyze feature relevance for reject option classifiers. We propose a method to calculate feature relevance intervals for the special case of linear reject option support vector machines. These relevance intervals provide information about the all-relevant problem as well as allowing the distinction between strongly and weakly relevant features. In the following sections, we first shortly review the main points in feature relevance and feature selection and reject option classifiers, exemplified by reject option SVM, and then present our method, as well as illustrate its results on toy data sets.

2 Feature Relevance

Our setting is a classification problem with features $X_1, \ldots, X_d$ and binary labels $Y$. Regard the feature $X_i$ and let $S_i$ be the set of all features except $X_i$. Then, by the taxonomy introduced in [7], the feature $X_i$ is strongly relevant if it contains information about $Y$ not contained in $S_i$. It is weakly relevant if it contains information about $Y$ not contained in some subset $S \subseteq S_i$ and it is irrelevant if it is neither strongly nor weakly relevant.

The distinction between strong and weak relevance is necessary because it is possible for a feature to be informative but redundant, in the sense that it does contain information on the labeling $Y$, but does not add to the information already contained in $S_i$. In this case, a binary sense of relevant and irrelevant would fail to accurately represent the informativeness of $X_i$.

In applied machine learning, we are often not directly concerned with which features contain information on our target variable, but rather with which features help us solve the classification problem well, i.e. which features contain information on $Y$ that can be leveraged by our choice of model. In this setting, the distinction between strong and weak relevance is equally sensible, because a feature that contains information leverageable by our chosen model may still be made redundant by other features with the same qualities. This is one reason why a unique solution of the minimal-optimal feature selection problem may not exist: exchanging one of the redundant features for another may not affect the quality of the feature subset for qualification. These cases motivate us to consider the all-relevant feature selection problem, which aims to obtain information about all equivalently good subsets. As an example of where this may be important, consider a doctor who, besides the predictive performance offered by a feature set, also has to consider how expensive and invasive the measurement of the features is, or an engineer who has to consider that the accuracy of measuring features deteriorates to different degrees under conditions such as location, lighting and temperature.

In Section 3, we present a method that calculates relevance intervals for the hypothesis class of linear reject option classifiers, thus giving information about which features are necessary to achieve a certain level of classification performance, and which features may be substituted for others.

3 Reject Option SVM

Reject option classification is a common approach in applications where the cost of misclassification is high enough that rejecting a data point is preferable to classifying it if the probability of misclassification is large. If we let $r$ be the highest probability of misclassification that we find acceptable, we can define a reject option loss function that assigns a cost of 1 to a misclassification, 0 to a correct classification, and a cost of $r$ to a reject. Bartlett et al. [1] introduced a piecewise linear surrogate of the cost function optimized by the optimal reject Bayes classifier that generalizes the hinge loss commonly used for training support vector machines and showed how to use it in the training of reject option SVMs. Specifically, given data $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$, for a linear kernel, the reject option SVM is uniquely determined as the solution of the following convex optimization problem:
We have seen that choosing a single, well-performing model and analyzing which features are relevant is an important step in understanding the behavior of the classifier. However, often it is necessary to deduce acceptable classification performance and infer feature relevance for each one, in order to deduce which features are substitutable for one another, should this be necessary out of economic or efficiency concerns. This can be achieved by the following procedure, which is an extension of our work in [5]:

1. Pick a baseline reject SVM, e.g. by solving the regularized empirical risk minimization problem proposed by Bartlett et al. [11]

2. Consider all reject option SVMs with similar costs and $l_1$-norm of the weight vector $\vec{w}$ as the baseline SVM. This allows weights to shift between correlated features while maintaining the performance on the observed data. The range of $|w_i|$ that occur in these SVMs are a proxy for the range of relevance that can sensibly be assigned to the $i$-th feature.

Step 2 can be formalized by the following optimization problems (a minimization and a maximization problem, both with the same objective function and constraints), for each feature $j = 1, \ldots, d$:

$$\min_{\vec{w}, \vec{b}, \vec{\xi}, \vec{\gamma}} \max_{\vec{w}'}, \vec{b}', \vec{\xi}', \vec{\gamma}' \left\| w_j' \right\|_1$$

s.t. $\left\| \vec{w}' \right\|_1 \leq \left\| \vec{w} \right\|_1$

$$\sum_{i=1}^{n} \left( \xi_i' + \frac{1 - 2r}{r} \gamma_i' \right) \leq \sum_{i=1}^{n} \left( \xi_i + \frac{1 - 2r}{r} \gamma_i \right)$$

$$\xi_i' \geq 0, \gamma_i' \geq 0, \xi_i' \geq 1 - y_i (\langle \vec{w}', \vec{x}_i \rangle + b')$$

$$\gamma_i' \geq -y_i (\langle \vec{w}', \vec{x}_i \rangle + b'), \quad i = 1, \ldots, n.$$ using the original reject SVM solution ($\vec{w}, \vec{b}, \vec{\xi}, \vec{\gamma}$) as a baseline. Here, we upper bound the margin violation of the hypotheses ($\vec{w}', b'$) by the margin violations of the baseline classifier in order to ensure that our hypotheses perform as well concerning the surrogate loss function of [11] as our base hypothesis. We do not lower bound the margin violations in order to admit improvements on the baseline hypothesis.

These optimization problems can be solved using linear programs, and thus the results are unique and can be computed in polynomial time. If $w_{min}$ and $w_{max}$ are the solutions of the minimization and maximization problems, respectively, then $[w_{min}, w_{max}]$ is exactly the range of magnitudes of a weight vector such that weights are only shifted compared to the baseline, with no adverse effects on the reject option loss.
Figure 1: The left column shows the baseline SVM. The middle column shows a linear classifier that assigns minimal weight to the first feature (plotted horizontally), while maintaining the $l_1$-norm and value of the modified hinge loss of the baseline classifier. The right column shows a linear classifier giving maximal weight to the first feature, again maintaining the $l_1$-norm and value of the modified hinge loss of the baseline classifier. The top, middle and bottom rows depict the first, seconds and third datasets, respectively. Rejected samples are shown in gray; classified samples are shown in black or white. The decision boundary is indicated by a black line.

5 Experiments

We present experimental results of our method on three toy datasets that are two-dimensional (to allow straight-forward visualization) and result from independent measurements of the same feature, i.e. the features are designed contain redundant information, but the duplication may become useful due to independent noise. In the first two datasets, noise gets more pronounced the closer a sample lies to the decision boundary. In the first dataset, the first feature is noisier than the second, while in the second dataset, both are subject to equal amounts of noise. Both datasets simulate cases where measuring a feature is more complicated in critical cases, e.g. in medical settings. The third dataset is composed of samples that are fewer towards the decision boundary, which is akin to populations in which the vast majority of samples has characteristics expressed strongly in one of two ways.

The results of our method, applied to the three datasets, are displayed in Figure 1. Even though measuring twice helps alleviate the decline in classification accuracy and certainty caused by noise, our method detects when a single measurement suffices for classification performance close to optimal – however, it does not simply reduce the feature’s weight to zero or increase it to the maximum in
all cases. For example, for the first dataset, since the first feature is noisier than the second, relying solely on the first is not possible.

Notice how, in the majority of cases, assigned classes are not flipped, but the choice of rejected points differs strongly among these classifiers with the same generalized hinge loss. In the future, we will investigate the behavior of the set of rejected points and of the true reject loss in the extremal cases.

6 Conclusion

We have presented a method of calculating feature relevance intervals for linear reject option SVM. In a first analysis, our method has produced results that align with the underlying ground truth as well as our expectation. The next steps are a comparative analysis with other feature selection methods on toy data with known ground truth, an evaluation on real world data, and a closer investigation of the true loss of the “equivalent” hypotheses we consider.

Acknowledgement

Funding in the frame of the CITEC cluster of excellence is gratefully acknowledged.

References


