Preprint – Accepted at NIPS Workshop on Transparent and Interpretable Machine Learning in Safety Critical Environments.

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

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

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 $X_1, \dots, X_d$ 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 $S \subsetneq S_i$ 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 4, 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 $(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$, for a linear kernel, the reject option SVM is uniquely determined as the solution of the following convex optimization problem:

$$\begin{aligned} \underset{\vec{w},b,\vec{\xi},\vec{\gamma}}{\arg\min} &\|\vec{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i + \frac{1-2r}{r}\gamma_i\right) \\ s.t.\quad & \xi_i \geq 0, \gamma_i \geq 0, \xi_i \geq 1 - y_i\left(\langle \vec{w}, \vec{x}_i\rangle + b\right), \\ & \gamma_i \geq -y_i\left(\langle \vec{w}, \vec{x}_i\rangle + b\right), \qquad i = 1, \ldots, n. \end{aligned}$$

where the resulting classification rule is given by $f(\vec{x}) = \text{sgn}(\langle \vec{w}, \vec{x}\rangle + b)$ if $|\langle \vec{w}, \vec{x}\rangle + b| > r$ and "reject" otherwise.

The magnitude of the $|w_i|$ determines how much a change in the variable X_i influences the certainty of the classifier: changing the value of the i-th feature by $\frac{2r}{|w_i|}$ or greater where $w_i \neq 0$ can make the difference between a confident prediction of one class or the other, whereas a change by less than $\frac{2r}{|w_i|}$ cannot flip the class prediction, only change a confident prediction to a reject or vice versa. $w_i = 0$ means that the predictions of the classifier are completely independent of the value given to the i-th feature. These observations motivate the use of the weights $|w_i|$ as proxies for the relevance of feature X_i. However, if there is a set of correlated features in the data set, the minimization of the l_2 norm encourages the classifier to distribute weight over this group, making each seem irreplaceable on the one hand, and obscuring the magnitude of their relevance on the other hand, even though "shifting" the weight to a single one out of the group of features would lead to identical classification behavior. We show how to address this issue in the following section.

4 Feature Relevance Intervals for Reject SVM

We have seen that choosing a single, well-performing model and analyzing which features are relevant to this model provides an incomplete picture at best. Instead, we want to consider *all* models with 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. [1]
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 $j = 1, \ldots, d$:

$$\begin{aligned} \min_{\vec{w}',b',\vec{\xi}',\vec{\gamma}'} \Big/ \max_{\vec{w}',b',\vec{\xi}',\vec{\gamma}'} & |w'_j| \\ s.t.\quad & \|\vec{w}'\|_1 \leq \|\vec{w}\|_1 \\ & \sum_{i=1}^{n}\left(\xi'_i + \frac{1-2r}{r}\gamma'_i\right) \leq \sum_{i=1}^{n}\left(\xi_i + \frac{1-2r}{r}\gamma_i\right) \\ & \xi'_i \geq 0, \gamma'_i \geq 0, \xi'_i \geq 1 - y_i\left(\langle \vec{w}', \vec{x}_i\rangle + b'\right), \\ & \gamma'_i \geq -y_i\left(\langle \vec{w}', \vec{x}_i\rangle + b'\right), \qquad i = 1, \ldots, n. \end{aligned}$$

using the original reject SVM solution $(\vec{w}, b, \vec{\xi}, \vec{\gamma})$ 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 [1] 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.

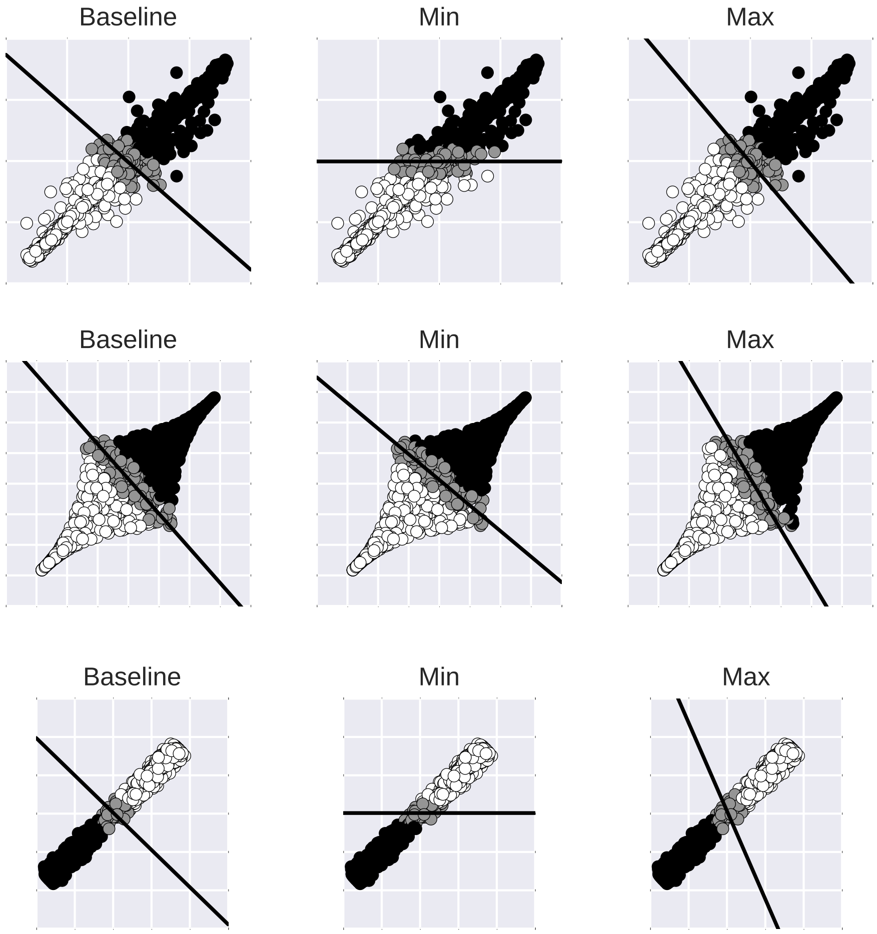

Figure 1: The left column shows the baseline SVM. The middle column shows a linear classifier that assigns minimal weight to the first feature (plotted horizontally), while maintaining the l_1-norm and value of the modified hinge loss of the baseline classifier. The right column shows a linear classifier 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. Code for our method and the experiments can be found at `github.com/janphigoe/rejectferel`.

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

[1] P. L. Bartlett and M. H. Wegkamp. Classification with a Reject Option using a Hinge Loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

[2] B. Frénay, G. Doquire, and M. Verleysen. Theoretical and Empirical Study on the Potential Inadequacy of Mutual Information for Feature Selection in Classification. *Neurocomputing*, 112:64–78, nov 2013.

[3] B. Frénay, D. Hofmann, A. Schulz, M. Biehl, and B. Hammer. Valid interpretation of feature relevance for linear data mappings. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 149–156, dec 2014.

[4] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, nov 2003.

[5] C. Göpfert, L. Pfannschmidt, J. P. Göpfert, and B. Hammer. Interpretation of Linear Classifiers by Means of Feature Relevance Bounds. *Neurocomputing ESANN Special Issue*, 2017.

[6] M. E. Hellman. The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, 6(3):179–185, July 1970.

[7] R. Kohavi and G. H. John. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2):273–324, nov 1997.

[8] J. Kummert, B. Paassen, J. Jensen, C. Göpfert, and B. Hammer. Local reject option for deterministic multi-class SVM. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9887 LNCS, pages 251–258, 2016.

[9] W. R. Rudnicki, M. Wrzesień, and W. Paja. All Relevant Feature Selection Methods and Applications. In U. Stańczyk and L. C. Jain, editors, *Feature Selection for Data and Pattern Recognition*, Studies in Computational Intelligence, pages 11–28. Springer Berlin Heidelberg, jan 2015.

[10] R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[11] H. Zou. An improved 1-norm SVM for simultaneous classification and variable selection. pages 675–681, 2007.

[12] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, nov 2005.