In the following, we respond to all the reviewers' questions that will be addressed in the paper's final version together with all their suggestions.

Reviewer #1

1. The optimization of 0-1 loss instead of a surrogate loss brings the learning process one step closer to the original goal of minimizing expected 0-1 loss. The results in the paper show that optimizing 0-1 loss can lead to enhanced performance guarantees (much tighter bounds) since the learning process directly provides bounds on the expected 0-1 loss (probability of error). We would also like to point out that the bounds' tightness is shown not only in Fig.1 but also in "LB" and "UB" columns of Table 1.

2. Learning techniques in Theorem 1 and performance guarantees in Theorems 2 and 3 are obtained addressing the minimax in line 105 using Lagrange duality. For instance, parameters $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ correspond to the Lagrange multipliers of constraints given by $\mathbf{a}$ and $\mathbf{b}$ in (1), respectively. (See also Answer 4 to Reviewer #4)

3. cvxpy can be installed following the instructions in https://www.cvxpy.org/install/. We will also include a readme file with detailed installation steps in the implementation files.

Reviewer #2

1. As the Reviewer mentions, optimization based on stochastic gradient descent (SGD) approaches can increase efficiency especially for large-scale training. The learning techniques proposed in the paper can be addressed using variants of SGD methods. In particular, primal-dual subgradient descent methods can enable efficient iterative optimization using subgradients of objective and constraints functions. In addition, the expression for $\mathbf{a}$ and $\mathbf{b}$ in (2) given by sample averages leads to an objective function in (3) that is amenable for stochastic subgradient descent methods.

2. All methods (proposed MRC and competing techniques) were implemented using their default parameters and settings in all datasets for a fair and transparent experimental comparison.

3. The role of the feature map in the presented methods is similar to that in conventional linear classifiers such as SVMs and logistic regression. However, as the Reviewer mentions, the paper offers new insights for feature mappings' design including their role in determining the probability distributions considered (uncertainty set) in (1) together with the trade-offs for dimensionality, variability, and training size in the generalization bounds of Theorem 3.

4. The main criteria for choosing the UCI datasets was to select frequently used datasets for binary and multi-class problems. Those with large number of samples were used for comparison with performance bounds in Fig.1 over one instantiation in terms of training size up to 10,000 samples, while the others (with less than 1000 samples) were used for comparison with both state-of-the-art techniques and performance bounds in Table 1 using 10-fold cross-validation.

Reviewer #4

1. We would like to clarify that the finite cardinality of the instance space does not lead to any practical limitation or computational burden. In the paper, instance spaces are taken to be finite only for technical convenience in the proof of Theorem 1. Infinite instance spaces would require to use heavier tools from variational analysis in such proof, but the corresponding MRC methods would not change. The methods presented do not create or compute matrices describing probability distributions and classification rules. Such methods obtain expectation estimates $(\mathbf{a}, \mathbf{b})$ using (2) and MRCs parameters ($\boldsymbol{\mu}_a$, $\boldsymbol{\mu}_b$, ν) through optimization problem (3) that has dimensionality and number of constraints given by the feature mapping used, independently of the size of the instance space.

2. Uncertainty sets are chosen to be determined by linear constraints in (1) so that Lagrange duality enables to obtain efficient learning techniques through Theorem 1. In addition, the uncertainty sets proposed can be guaranteed (with prob. $> 1 - \delta$) to include the true data-generating distribution using $(\mathbf{a}, \mathbf{b})$ given by expectations' confidence intervals at level $1 - \delta$. Such condition can be achieved by using parameter λ as given in Theorem 3 (line 183) or using other statistical methods that obtain expectations' confidence intervals from i.i.d. samples.

3. MRCs' generalization depends on the uncertainty set that is determined by the feature map in (1). Theorem 3 shows MRCs' generalization in terms of feature map characteristics such as dimensionality and variability. We will describe how to interpret such results in terms of uncertainty sets, e.g., increased dimensionality reduces uncertainty set size.

4. We will include a short description of Theorem 1 proof to show the intuition behind such result. In particular, the minimax problem addressed by MRCs is equivalent to optimization problem (3) by using Lagrange duality and maximin equivalence. Parameters $\boldsymbol{\mu}_a, \boldsymbol{\mu}_b$, and ν are the Lagrange multipliers corresponding to the linear constraints defining the uncertainty set, and constraints in (3) come from the conjugate of the objective in the maximin problem.

5. We will describe that the role of the feature map in the presented methods is similar to that in conventional linear classifiers such as SVMs and logistic regression. In particular, MRCs are determined by a linear-affine combination of the feature map as shown in (4). Threshold-based features are used in experimentation section to plainly show the potential of the new approach. More sophisticated feature maps such as those given by kernel-based embeddings in conventional techniques can be analogously used (see short discussion in lines 236-239 and footnote 1).

6. For fair experimental comparison, we implemented all the methods (including the proposed MRCs) using their default settings and parameters in all the datasets. For the experimentation carried-out in the paper, we consider that it is more transparent not to use cross-validation or tuning in any method.