Examining Loss Functions for Cost-Sensitive Learning

Jacek P. Dmochowski[1], Paul Sajda[2], and Lucas C. Parra[1]

[1]Dept. of Biomedical Engineering, City College of New York, New York, NY, 10031
[2]Dept. of Biomedical Engineering, Columbia University, New York, NY, 10027

Pattern classifiers are essentially decision makers whose ultimate goal is to minimize the expectation of *loss*, or risk:

$$R(\boldsymbol{\theta}) = E\left\{c(\mathbf{x},y)u\left[-yf(\mathbf{x};\boldsymbol{\theta})\right]\right\} = \int\int p(\mathbf{x},y)c(\mathbf{x},y)u\left[-yf(\mathbf{x};\boldsymbol{\theta})\right]d\mathbf{x}dy. \tag{1}$$

where $\boldsymbol{\theta}$ is a vector parameterizing the classifier f, $p(\mathbf{x},y)$ is the joint probability measure on feature $\mathbf{x}$ and class $\mathbf{y}$, $c(\mathbf{x},y)$ denotes the cost of misclassifying $\mathbf{x}$ from class y, and $u(a) = 1,\ a > 0$ and 0 otherwise. In many cases, the cost is feature independent and we have $c(\mathbf{x},y) = c(y)$. The classical zero-one loss is found by setting $c(\mathbf{x},y) = 1,\ \forall \mathbf{x},y$. Given that a host of learning applications involve a (typically rare) class with very high associated misclassification cost (e.g., disease diagnosis, security and defense applications, etc.), it is surprising that the cost-sensitive learning (CSL) problem [1] comprises just a small niche in the learning community, and that the predominant focus is on the zero-one loss case.

Our work [2] is concerned with methods of estimating $\boldsymbol{\theta}$ to minimize risk in cost-sensitive applications. We focus on the empirical risk minimization (ERM) framework, where the empirical estimate of (1) is minimized via a differentiable surrogate loss function L :

$$R_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} c(\mathbf{x}_n,y_n)u\left[-y_n f(\mathbf{x}_n;\boldsymbol{\theta})\right] \approx \frac{1}{N}\sum_{n=1}^{N} L\left(y_n,\mathbf{x}_n,\boldsymbol{\theta}\right), \tag{2}$$

where $\{\mathbf{x}_n,y_n\}$ is a set of training examples. We numerically examine the behavior of the minimizer of (2) in terms of the underlying loss function, the costs $c(\mathbf{x},y)$, the specificity of the model set, and sample size N.

The cost-sensitive learning problem differs non-trivially from the zero-one case. For a given classifier f, the expectation (1) is often dominated by the infrequent but costly class. Making correct decisions on these samples is critical to minimizing loss. One approach to CSL is thus to ensure correct classification on the training samples of the costly class, typically via a weighting by cost; the empirical risk (2) is then dominated by these costly examples. If these examples are quite noisy, the learning algorithm is prone to a large generalization error, and the *effective* sample size is more indicative of the number of costly examples than overall N. Thus, we probe various loss functions on their ability to yield good generalization ability in such cost asymmetric environments.

Assume a binary classification and that our classifier models the log odds of the true posterior probabilities: $f(\mathbf{x}) = \ln\frac{p(+1|\mathbf{x})}{p(-1|\mathbf{x})}$. The optimal (risk minimizing) Bayesian decision rule follows as $\hat{y} = \text{sign}\left[f(\mathbf{x}) - \ln\frac{c(\mathbf{x},-1)}{c(\mathbf{x},+1)}\right]$. A common practice in CSL is thus to estimate the log odds of the class posteriors and then apply the threshold shift heuristically: $\hat{y} = \text{sign}\left[f\left(\mathbf{x};\boldsymbol{\theta}\right) - \ln\frac{c(\mathbf{x},-1)}{c(\mathbf{x},+1)}\right]$. If the estimate $\boldsymbol{\theta}$ is consistent, i.e., $\lim_{N\to\infty} f\left(\mathbf{x};\boldsymbol{\theta}\right) = f(\mathbf{x})$, this procedure is guaranteed to minimize risk asymptotically. It is here that the specification of the model set plays a key role in CSL: in the case of a well-specified model, a threshold shift of the maximum likelihood solution guarantees asymptotic optimality. On the other hand, in the case of misspecification, the parameter vector which minimizes risk varies with the ratio of misclassification costs. This is illustrated for the simplest case in Figure 1, where we show the minimum risk solutions in the case of (unequal covariance) Gaussian classes with a quadratic and linear model set. Optimal decision surfaces are shown for various values of the "probability cost function" [3]: $\text{pcf} = \frac{p(+)c(+)}{p(+)c(+)+p(-)c(-)}$, where $p(y)$ indicates prior class probabilities.

Notice that asymmetric costs $c(\mathbf{x},y)$ effectively scale the probabilities $p(\mathbf{x},y)$ in the expression for the risk (1). One strategy is thus to weight the log-likelihood according to costs:

$$L_{\text{wml}}(\mathbf{x},y,\boldsymbol{\theta}) = L_{\text{ml}}(\mathbf{x},y,\boldsymbol{\theta})^{c(\mathbf{x},y)} = -\log p\left(y|\mathbf{x};\boldsymbol{\theta}\right)^{c(\mathbf{x},y)} = -c(\mathbf{x},y)\log\left[\text{sig}\left(yf(\mathbf{x};\boldsymbol{\theta})\right)\right], \tag{3}$$

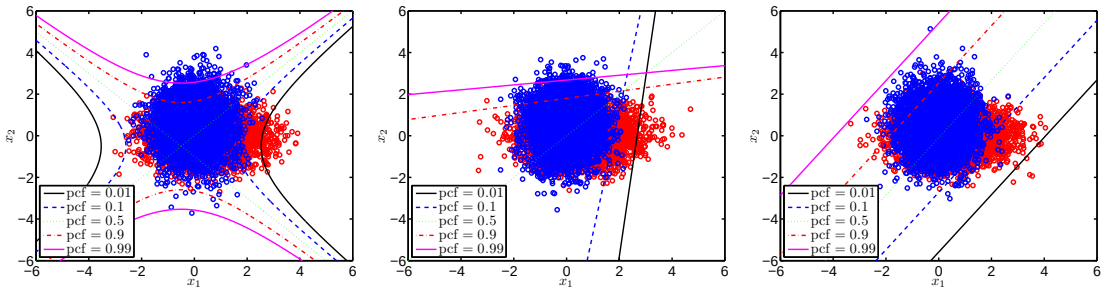
where $\text{sig}(a) = \frac{1}{1+e^{-a}}$ is the logistic sigmoid.

We have proposed a direct approximation of (1) with the following non-convex surrogate:

$$L_{\text{sig}}(\mathbf{x},y,\boldsymbol{\theta}) = c(\mathbf{x},y)\text{sig}\left[-yf(\mathbf{x};\boldsymbol{\theta})\right], \tag{4}$$

which corresponds to a sigmoidal approximation to the step function u. Using the inequality $z \le -\log(1-z),\ z \le 1$, one can show that this loss is upper bounded by the negative weighted log likelihood:

$$L_{\text{sig}}(\mathbf{x},y,\boldsymbol{\theta}) \le L_{\text{wml}}(\mathbf{x},y,\boldsymbol{\theta}). \tag{5}$$

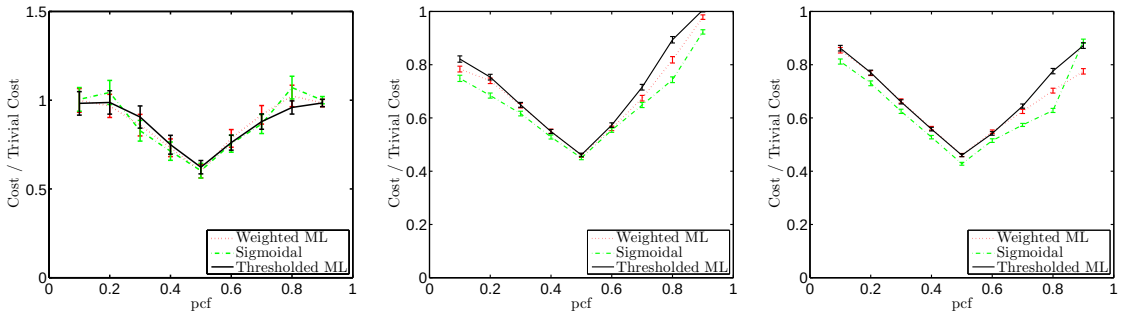


(a) minimum risk quadrics (b) minimum risk planes (c) thresholded maximum likelihood

Figure 1: Minimum risk classification of Gaussian data with unequal covariance matrices: (a) A quadratic model set in which a threshold shift of a cost-insensitive (pcf = 0.5) classifier yields an optimal classification, (b) the optimum linear classifier for various costs, (c) a suboptimal thresholded maximum likelihood solution.

Thus, minimizing the negative weighted log likelihood corresponds to minimizing a convex upper bound on the cost-sensitive sigmoidal loss.

Figure 2 shows the cross-validated risk on three data sets from the UCI database for the loss functions presented above. In the figure, we normalize the risk (1) with respect to the "trivial risk", or the risk yielded by the data-independent classifier: $\hat{y} = \text{sign}\,[p(+1)c(+1) - p(-1)c(-1)]$, which selects the more prevalent class in the case of zero-one loss. All methods employed ℓ_2 regularization with hyperparameters tuned using (nested) cross-validation. In general, we have found that with small sample sizes (relative to the dimensionality of the problem), cost-sensitive



(a) Transfusion, $N = 648$ (b) Magic, $N = 19020$ (c) Adult, $N = 32561$

Figure 2: Cost curves (with standard errors of the means) for various real data sets with synthetic costs. Negative weighted log likelihood and sigmoidal risk perform

loss functions offer only mild improvements over the thresholded ML approach (subfigure a). In these "small N" cases, cost-sensitive approaches fail to yield accurate estimates of the minimum risk solution, and the model-based thresholded ML, while suboptimal, provides comparable risk. On the other hand, given enough training data (subfigures b and c), the weighted ML and sigmoidal estimates provide substantial gains over thresholded ML, with the sigmoidal risk estimator typically outperforming weighted ML. This is somewhat surprising given prior work on zero-one risk with convex loss functions [4]. Note that while the normalized risk is typically lowest in the zero-one case (pcf = 0.5), the greatest "savings" over thresholded ML are often observed at extreme values of the pcf, or in those cases where the costs are very skewed. The determination of loss functions which can significantly outperform thresholded ML in the small N case remains an open and challenging problem.

References

[1] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978.

[2] J. P. Dmochowski, P. Sajda, and L. C. Parra, "Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds," *Journal of Machine Learning Research*, 11:3313–3332, 2010.

[3] C. Drummond and R. C. Holte, "Explicitly representing expected cost: an alternative to ROC representation," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 198–207, ACM Press.

[4] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, Classification, and Risk Bounds," *Journal of the American Statistical Association*, vol. 101, pp. 138-156, 2006.

Topic: learning theory
Preference: poster

Presenting author: either Jacek Dmochowski or Lucas Parra
Corresponding e-mail address: `jdmochowski@ccny.cuny.edu`