Generalizing the margin concept to arbitrary classifiers

Kroon, Steve  
*Stellenbosch University, Computer Science Division*  
Private Bag X1  
Matieland (7602), South Africa  
E-mail: kroon@sun.ac.za

Steel, Sarel  
*Stellenbosch University, Department of Statistics and Actuarial Science*  
Private Bag X1  
Matieland (7602), South Africa  
E-mail: sjst@sun.ac.za

Many binary classification problems are tackled using *thresholding classifiers*, i.e. an instance is classified by thresholding a real-valued response calculated from the instance. Margin bounds are based on the intuition that the further the real value calculated differs from the threshold, the higher the confidence one can have in the resulting classification.

Margin bounds came to light following the proposal of the support vector machine and the subsequent investigation of their excellent classification performance. The pioneering work in margin bounds was predominantly due to Peter Bartlett, John Shawe-Taylor and their collaborators. The following bound (Bartlett, 1998) is an example of a margin bound for thresholding linear classifiers:

\[
P_{S \sim D} \left\{ \sup_{h \in \mathcal{H}} \frac{e_D(g_h) - e_\gamma^g(h)}{\sqrt{e_D(g_h)}} > \epsilon \right\} < 4 \mathcal{N} \left( \frac{\gamma^2}{2}, \mathcal{H} \right) \exp \left( -\frac{\epsilon^2 m}{4} \right).
\]

The result provides an exponentially decaying, distribution-independent bound on the probability of drawing an \(m\)-sample of input-output pairs \((x_i, y_i)\) such that that the relative deviation between the true classification error of the thresholded \(h\) and the sample \(\gamma\)-error is large for any classification rule \(h \in \mathcal{H}\). For this case, the strategy \(g\) applied to the hypothesis \(h\) is thresholding at zero, so that \(g_h(x) \equiv g(h(x)) = \text{sgn}(h(x))\), and the true classification error of \(g_h\) is \(e_D(g_h) = P_{(x,y) \sim D} \{ g_h(x) \neq y \}\). Meanwhile, the sample \(\gamma\)-error \(e_\gamma^g(h)\) is independent of the strategy, and is the proportion of sample points not achieving a margin of at least \(\gamma\), where the margin of a point \((x, y)\) is defined for the result above by \(\rho(x, y) = yh(x)\).

A similar bound to the above, applicable to \(\epsilon\)-insensitive regression, was formulated in Anthony and Bartlett (1994), but not as a margin bound. This early regression result was later reformulated as a margin bound on a thresholded classifier in Shawe-Taylor et al. (1998). This parallel led us to investigate the extension of the margin concept, resulting in the general margin for binary classifiers presented in this paper. In what follows, we present the generalized margin, noting that the results for traditional margin bounds still apply directly to the more general margin concept, and present two examples of the general margin concept.

A more general margin The margin of the point \((x, y)\) in the example above can be rewritten as \((yg_h(x))|h(x)|\). We can then view the margin’s sign as an indication of whether \(g_h\) classifies the point correctly, while the magnitude of the margin is the distance of \(h(x)\) from the threshold (zero).

The generalized margin concept considers the margin as a distance on the “correct side” of the decision boundary. In the previous example, the hypotheses \(h\) map onto the real line, and the point zero defines a decision boundary between inputs classified as \(-1\) or \(1\) by any hypothesis \(h\). Thus, the distance \(h(x)\) is on the “correct side” of the decision boundary is \(\rho(x, y) = yh(x)\). More generally,
however, we can use other strategies, and hypotheses need not map onto the real line. However, at some stage, we must rely on a strategy $g$ to classify a new $x$ based on the output $h(x)$. Supposing the hypotheses map into a metric space $(\mathcal{E},d)$, then $g$ induces a partition of $\mathcal{E}$ for any $h$. Using this partition, we can formalize our margin concept, defining the margin as follows:

$$\rho(x, y) = (yg_h(x))d(h(x), \{\eta \in \mathcal{E} : g(\eta) \neq g_h(x)\})$$

**First example** For this example, consider a classifier which calculates a point $h(x)$ in an arbitrary space $\mathcal{E}$, and then classifies $x$ based on whether $h(x)$ lies within distance $\varepsilon$ of some target value $\eta_0$. In this case, the strategy is $g(\eta) = I(d(\eta, \eta_0) > \varepsilon)$. With a little attention, our margin definition in (2) above yields the expression $\rho(x, y) = y(d(h(x), \eta_0) - \varepsilon)$.

This example is very similar to the formulation of the $\varepsilon$-insensitive regression bound as a margin bound. Note that by appropriate selection of the metric $d$, we can get a variety of shapes for our decision surface. For example, using the Manhattan metric in $\mathbb{R}^N$ yields a box-shaped region, with bounds determined by the distance that mapped training points lie from the box surface.

**Second example** Consider a voting classifier composed of $N$ thresholded classifiers, each thresholding at zero. The voting classifier predicts 1 if $j$ or more of the $N$ base classifiers predict 1. Thus, we can view any selection of the $N$ base classifiers as a map $h$ into $\mathbb{R}^N$, while the strategy $g$ maps $\mathbb{R}^N$ to $\{-1, 1\}$. Clearly, $g$ partitions $\mathbb{R}^N$ into two sets of orthants, those with $j$ or more non-negative co-ordinates (mapped to 1), and those with less than $j$ non-negative co-ordinates (mapped to $-1$).

Let us employ the Manhattan metric $d_1$ for $\mathbb{R}^N$, and consider a point $\eta \in \mathbb{R}^N$ such that $g(\eta) = 1$. Then $\eta$ has $i \geq j$ non-negative co-ordinates. Any other point classified differently from $\eta$ must have a distance from $\eta$ of at least $\varepsilon$ of the sum of the $(i - j) + 1$ smallest non-negative co-ordinates of $\eta$. Similarly comparing a point $\eta$ such that $g(\eta) = -1$ to any point classified differently shows that the distance between them is at least the negated sum of the $i - j$ smallest negative co-ordinates of $\eta$. Since these distances can be approximated arbitrarily closely by appropriate selection of the other point, we can use these expressions to define the margin.

The result is a simple formula for the margin, which can be employed in conjunction with existing margin bounds, yielding performance bounds in terms of the unthresholded outputs of the individual classifiers in the voting machine.

**Conclusion** In conclusion, note that the example margin bound (1) is only a representative of a large number of margin bounds. Furthermore, to be practically useful issues such as margin unification needs to be addressed. This paper highlights that many margin bounds are more widely applicable than their initial statement in the literature indicates. For a further discussion of the generalized margin concept and references on other issues, the interested reader is referred to Kroon (2008).

**REFERENCES**


