Margin bounds for arbitrary classifiers

Steve Kroon

Computer Science Division
Stellenbosch University

4 September 2009
1. Margin bounds
2. Generalized margin bounds
3. Examples
4. Conclusion
1. Margin bounds

2. Generalized margin bounds

3. Examples

4. Conclusion
Aim: To construct confidence intervals on average loss of a classifier without using a separate, independent test sample.

Benefit: More data can be used for training the classifier.

Obtained by bounding deviation of mean loss from some statistic with high probability.
• Margin bound: applies to binary classification based on thresholding real-valued outputs.

• Bounds deviation of true risk from sample margin risk.

• Classical result (Bartlett, 1998): Suppose an independent $m$-sample is drawn from a distribution $D$, and $\mathcal{H}$ is a class of real-valued functions. Then, with probability at least $1 - \delta$, every $h \in \mathcal{H}$ has

$$r_D(h, L_0) < r_S(h, L_\gamma) + \sqrt{\frac{2}{m} \ln \frac{2N_\infty(\frac{\gamma}{2}, \mathcal{H}, 2m)}{\delta}}.$$ 

• Other improvements possible, but outside scope (uniform over $\gamma$, other ghost-sample sizes, realizable case).
\[ r_D(h, L_0) < r_S(h, L_{\gamma}) + \sqrt{\frac{2}{m}} \ln \frac{2N_{\infty}(\frac{\gamma}{2}, \mathcal{H}, 2m)}{\delta}. \]

- \( r_P(h, L_{\gamma}) \), called the \( \gamma \)-margin risk, is probability an input-output pair \((x, y)\) sampled from \( P \) satisfies \( yh(x) < \gamma \).
- Think of successful classification with this loss function as achieving a margin of \( \gamma \), where margin is \( yh(x) \).
- On the real line, margin can be thought of as signed distance of \( h(x) \) on the correct side of the decision boundary at zero.
- \( N_{\infty}(\frac{\gamma}{2}, \mathcal{H}, 2m) \) - number of \( \frac{\gamma}{2} \)-distinct functions in \( \mathcal{H} \) when restricted to some set of \( 2m \) input points.
Example: 2-D linear discriminant analysis
1. Margin bounds

2. Generalized margin bounds

3. Examples

4. Conclusion
No need to restrict ourselves to the real line. Just need a decision boundary in some space $\mathcal{E}$.
Bartlett’s result (and others) apply to any class of real-valued functions.
We can induce a real-valued function class by composing each function $h$ into $\mathcal{E}$ with a real-valued function $d : \mathcal{E} \rightarrow \mathbb{R}$.
To be useful: $d$ should measure some notion of signed distance from the decision boundary (zero on boundary, opposite signs on opposite sides of boundary).
Then apply margin bound to $d \circ \mathcal{H}$.
Example: squashed function classes can be seen as $d : \mathbb{R} \rightarrow \mathbb{R}$.
Problem: complicating the function class, then trying to estimate covering numbers in $\mathbb{R}$.

When $|d(\cdot)|$ is set-distance based on a pseudometric, the same argument applied on the real line can be applied in $\mathcal{E}$ directly.

Still use covering numbers of $\mathcal{H}$ — dependency on $d$ now via the underlying pseudometric.

Leads to a more general definition of margin.
Let $d'$ be a pseudometric for $\mathcal{E}$.
Let the decision boundary be $E \subseteq \mathcal{E}$.
Let $d(e) = d'(e, E), \ e \in \mathcal{E}$.
Let $g : \mathcal{E} \rightarrow \{-1, 1\}$ indicate what the prediction is for each point in $\mathcal{E}$.

Generalized margin of $(x, y)$ is $yg(h(x))d(h(x))$.

Existing margin bounds apply almost verbatim, but distances for covering numbers are now based on $d'$, i.e. $\mathcal{N}(\frac{\gamma}{2}, \mathcal{H}, d'_\infty, 2m)$.
1. Margin bounds
2. Generalized margin bounds
3. Examples
4. Conclusion
Consider the spherical decision boundary \( \{ e : e - \eta_0 = r \} \) in Euclidean space \( \mathbb{R}^n \).

\( \eta_0 \) might represent some “ideal prototype” of one class.

Then \( d(e) = \| e - \eta_0 \| - r \).

Could also get normal margin bound by composing with \( d \).

Special case: \( \varepsilon \)-insensitive prediction (still to come).
Margin bounds
Generalized margin bounds
Examples
Conclusion

Steve Kroon
Margin bounds for arbitary classifiers
Margin bounds
Generalized margin bounds
Examples
Conclusion
(j, N)-voting committee with thresholding members.

Predict 1 if at least j members predict 1, otherwise predict 0.

Let $e \in \mathbb{R}^n$ represent the vector of unthresholded predictions.

Use the Manhattan metric (i.e. 1-norm)

Decision boundary: set of orthant boundaries between orthants with $j$ positive coordinates and those with $j - 1$.

Let $n(e)$ be the number of nonnegative coordinates of $e$.

Let $e^\star$ be the vector obtained by re-ordering the components of $e$ in descending order.

New margin is $d(e) = \begin{cases} 
\sum_{k=j}^{n(e)} e^*_k & \text{if } n(e) \geq j \\
- \sum_{k=n(e)+1}^{j} e^*_k & \text{if } n(e) < j 
\end{cases}$.
Example: a (2, 2)-voting machine:
- $\varepsilon$-insensitive prediction.
- No loss when prediction $h(x)$ is within distance $\varepsilon$ of actual $y$.
- Seems decision boundary must depend on $y$.
- Consider a modified, but equivalent, problem: input is $(x, y)$ pair.
- Modified hypothesis $h'$ from $h$: $h'(x, y) = |h(x) - y|$.
- Decision boundary is at $\varepsilon$: wrong side is when $h'(x, y) > \varepsilon$. 
Margin bounds for arbitrary classifiers
1. Margin bounds
2. Generalized margin bounds
3. Examples
4. Conclusion
Margin concept more widely applicable than original results indicate.

Possibility of using alternative metrics for thresholding classifiers which are more suitable to the problem.

Complication: unusual metrics require covering numbers defined in terms of these unusual metrics.