Margins and Feature Analysis

Javed A. Aslam
College of Computer and Information Science
Northeastern University

February 12, 2010

We consider boosting decision stumps and the effect that individual features have on the margin distribution associated with the weighted linear combination that boosting produces.

Suppose that boosting proceeds for T rounds, and in each round t a decision stump h_t is selected and assigned confidence α_t. The weighted linear combination produced by boosting is

$$H(x) = \sum_{t=1}^{T} \alpha_t \, h_t(x)$$

and the margin associated with any instance x is

$$\mathit{margin}(x) = \frac{\ell(x) \cdot H(x)}{\sum_{t=1}^{T} |\alpha_t|} = \frac{\ell(x) \cdot \sum_{t=1}^{T} \alpha_t \, h_t(x)}{\sum_{t=1}^{T} |\alpha_t|}$$

where $\ell(x)$ is the $\{-1, +1\}$ label associated with the instance x.

Now assume that each decision stump is simply a feature-threshold pair. Let F be the total number of unique features used across all T decision stumps, and for any chosen feature f, let N_f be the total number of times that feature f is used. We then have

$$\sum_{f=1}^{F} N_f = T.$$

Finally, let $h_{f,j}$ be the decision stump that corresponds to the j-th use of feature f, and let $\alpha_{f,j}$ be the associated confidence.

We can now redefine $H(x)$ and $\mathit{margin}(x)$ as follows.

$$\begin{aligned} H(x) &= \sum_{f=1}^{F} \sum_{j=1}^{N_f} \alpha_{f,j} \, h_{f,j}(x) \\ \mathit{margin}(x) &= \frac{\ell(x) \cdot \sum_{f=1}^{F} \sum_{j=1}^{N_f} \alpha_{f,j} \, h_{f,j}(x)}{\sum_{t=1}^{T} |\alpha_t|} \end{aligned}$$

Now for any individual feature f, one can consider the weighted linear combination associated with that feature and the "conditional" margin associated with just that weighted linear combination.

$$\begin{aligned} H_f(x) &= \sum_{j=1}^{N_f} \alpha_{f,j} \, h_{f,j}(x) \\ \mathit{margin}_f(x) &= \frac{\ell(x) \cdot \sum_{j=1}^{N_f} \alpha_{f,j} \, h_{f,j}(x)}{\sum_{j=1}^{N_f} |\alpha_{f,j}|} \end{aligned}$$

Now consider the fraction of absolute "confidence" weight associated with any feature f, defined as follows.

$$\gamma_f = \frac{\sum_{j=1}^{N_f} |\alpha_{f,j}|}{\sum_{t=1}^{T} |\alpha_t|} = \frac{\sum_{j=1}^{N_f} |\alpha_{f,j}|}{\sum_{f=1}^{F} \sum_{j=1}^{N_f} |\alpha_{f,j}|}$$

We then have the following theorem.

Theorem 1 *$margin(x) = \sum_{f=1}^{F} \gamma_f \cdot margin_f(x)$.*

Proof:

$$\begin{aligned}
\sum_{f=1}^{F} \gamma_f \cdot margin_f(x) &= \sum_{f=1}^{F} \left(\frac{\sum_{j=1}^{N_f} |\alpha_{f,j}|}{\sum_{t=1}^{T} |\alpha_t|} \right) \cdot margin_f(x) \\
&= \frac{\sum_{f=1}^{F} \left(\sum_{j=1}^{N_f} |\alpha_{f,j}| \right) \cdot margin_f(x)}{\sum_{t=1}^{T} |\alpha_t|} \\
&= \frac{\sum_{f=1}^{F} \left(\sum_{j=1}^{N_f} |\alpha_{f,j}| \right) \cdot \left(\frac{\ell(x) \cdot \sum_{j=1}^{N_f} \alpha_{f,j} \, h_{f,j}(x)}{\sum_{j=1}^{N_f} |\alpha_{f,j}|} \right)}{\sum_{t=1}^{T} |\alpha_t|} \\
&= \frac{\ell(x) \cdot \sum_{f=1}^{F} \sum_{j=1}^{N_f} \alpha_{f,j} \, h_{f,j}(x)}{\sum_{t=1}^{T} |\alpha_t|} \\
&= margin(x)
\end{aligned}$$

□

Thus, we have that

> *the overall margin associated with any instance x is the weighted linear combination of conditional margins, where γ_f are the weights.*

This gives some justification for the use of γ_f as an indicator of the utility of a given feature f. Note, however, that while a feature f may have a large γ_f, it will not contribute to a good overall margin unless $margin_f(x)$ is also large. A better indicator, perhaps, is the fraction of the overall margin that is due to f:

$$\frac{\gamma_f \cdot margin_f(x)}{margin(x)}.$$

Note, however, that this only deals with a single instance x. To combine across all instances, one might be tempted to sum (or average) the above over all x. However, we care more about the entire *margin distribution* and the effect of a feature on this distribution.

Consider the mean of the margin distribution, i.e., the average margin. While the mean does not entirely characterize the margin distribution, it is a decent single-point measure of how "good" the margin distribution is. The average margin is

$$\begin{aligned}
\frac{1}{M} \sum_{i=1}^{M} margin(x_i) &= \frac{1}{M} \sum_{i=1}^{M} \sum_{f=1}^{F} \gamma_f \cdot margin_f(x_i) \\
&= \sum_{f=1}^{F} \left(\gamma_f \frac{1}{M} \sum_{i=1}^{M} margin_f(x_i) \right)
\end{aligned}$$

Thus, the fraction of the average margin due to feature f is

$$\frac{\gamma_f \frac{1}{M} \sum_{i=1}^{M} margin_f(x_i)}{\frac{1}{M} \sum_{i=1}^{M} margin(x_i)} = \gamma_f \cdot \frac{\sum_{i=1}^{M} margin_f(x_i)}{\sum_{i=1}^{M} margin(x_i)}.$$

While the above formula has a convenient interpretation in terms of conditional margins and fractional confidence weights, it can be simplified as follows.

$$
\begin{aligned}
\gamma_f \cdot \frac{\sum_{i=1}^{M} margin_f(x_i)}{\sum_{i=1}^{M} margin(x_i)} &= \frac{\sum_{j=1}^{N_f} |\alpha_{f,j}|}{\sum_{t=1}^{T} |\alpha_t|} \cdot \frac{\sum_{i=1}^{M} \left(\frac{\ell(x_i) \cdot \sum_{j=1}^{N_f} \alpha_{f,j}\, h_{f,j}(x_i)}{\sum_{j=1}^{N_f} |\alpha_{f,j}|} \right)}{\sum_{i=1}^{M} \left(\frac{\ell(x_i) \cdot \sum_{t=1}^{T} \alpha_t\, h_t(x_i)}{\sum_{t=1}^{T} |\alpha_t|} \right)} \\
&= \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_f} \ell(x_i)\, \alpha_{f,j}\, h_{f,j}(x_i)}{\sum_{i=1}^{M} \sum_{t=1}^{T} \ell(x_i)\, \alpha_t\, h_t(x_i)}.
\end{aligned}
$$