Generalization Bounds for Averaged Classifiers

Yoav Freund, Yishay Mansour, Robert E. Shapire

March 16, 2006

presented by Brian McFee for CSE291
“There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns— the ones we don’t know we don’t know.”

— Donald Rumsfeld

- Overfitting occurs when the hypothesis class is too complex relative to the training data.
- We can avoid overfitting by identifying points that are too close to call reliably (known unknowns) and abstaining from prediction.
Averaged Classifiers

- We have a classification problem of mapping I.I.D. points $X$ from some unknown distribution $D$ to a set of labels $Y$.
- Let $\mathcal{H}$ be a class of hypotheses: $h \in \mathcal{H}$ maps a point $x$ to a label $y \in \{-1, +1\}$.
- We want to classify $X$ based on a weighted average of all hypotheses in $\mathcal{H}$.
Generalization Example

- Suppose that the best hypothesis in our class, $h^*$, has 1% error over the unknown distribution $D$, and our training set is such that the generalization error will be 5%.

- If we can choose to abstain from prediction, we can hope that we only predict incorrectly 1% of the time, and the remaining 4% will have no prediction.

Key idea: learn to distinguish the points that we won’t be sure about (4%) from the points that even $h^*$ will get wrong (1%).
Toy Example

Rectangles represent hypotheses and “+” and “-” are the labels.
Weights and Prediction

- Weight for a hypothesis is determined by the exponential weights formula:
  \[ w(h) = e^{-\eta \hat{\epsilon}(h)}, \]
  for some learning rate \( \eta \) and training error \( \hat{\epsilon}(h) \).

- For a data point \( x \), predict \( y \) using the log ratio of weights assigned to hypotheses:
  \[
  \hat{\rho}(x) = \frac{1}{\eta} \ln \left( \frac{\sum_{h:h(x)=+1} w(h)}{\sum_{h:h(x)=-1} w(h)} \right)
  \]
After training on \( m \) samples \((x_i, y_i)\), define the training error of a hypothesis \( h \) as the fraction of mistakes:

\[
\hat{\epsilon}(h) = \frac{1}{m} \sum_{i=1}^{m} 1(h(x_i) \neq y_i).
\]
Note that $|\hat{\rho}(x)|$ indicates the amount of agreement among hypotheses on $x$.

If $|\hat{\rho}(x)|$ is near 0, the prediction is weak.

We can exploit this measure of confidence in the prediction by adding a threshold parameter:

$$\hat{P}(x) = \begin{cases} 
\text{sign}(\hat{\rho}(x)) & \text{if } |\hat{\rho}(x)| > \Delta \\
0 & -\Delta \leq \hat{\rho}(x) \leq \Delta 
\end{cases}$$
Averaged Classifiers
Generalization Bounds

Introduction
Some Definitions...

Log Ratio Prediction

\[ D = 0.50 \]

\[ \frac{w(+1)}{w(-1)} \]

\[ \Delta = 0.50 \]

\[ p(x) \]

\[ w(+1)/w(-1) \]

\[ P(x) = -1 \quad P(x) = 0 \quad P(x) = 1 \]

\[ -5 \quad -4 \quad -3 \quad -2 \quad -1 \quad 0 \quad 1 \quad 2 \quad 3 \quad 3.5 \quad 4 \]

\[ -5 \quad -4 \quad -3 \quad -2 \quad -1 \quad 0 \quad 1 \quad 2 \quad 3 \quad 3.5 \quad 4 \]
How well does \( \hat{P}(x) \) generalize?

Note that \( \hat{\rho}(x) \) is just an approximation of the unknown \( \rho(x) \):

\[
\rho(x) = \frac{1}{\eta} \ln \left( \frac{\sum_{h: h(x) = 1} e^{-\eta \hat{\varepsilon}(h)}}{\sum_{h: h(x) = -1} e^{-\eta \hat{\varepsilon}(h)}} \right)
\]

where \( \hat{\varepsilon}(h) \approx \varepsilon(h) = \Pr_{(x, y) \sim D}[h(x) \neq y] \).

For our prediction algorithm \( \hat{P}(x) \) to be useful, we need to satisfy two conditions:

1. \( \hat{\rho}(x) \) should be close to the true \( \rho(x) \) with high probability, ie. the effect of approximating \( \varepsilon(h) \) should be bounded.
2. If \( \hat{P}(x) \neq 0 \), the probability of making a mistake should be small.
Theorem 1: \( \hat{\rho}(x) \approx \rho(x) \)

- For any distribution \( D \), any instance \( x \), any \( s \in \{-1, +1\} \), and any \( \lambda, \eta > 0 \),

\[
\Pr_{S \sim D^m} \left[ s(\hat{\rho}(x) - \rho(x)) \geq 2\lambda + \frac{\eta}{8m} \right] \leq 2e^{-2\lambda^2 m}
\]

- Note that this is independent of the hypothesis class \( \mathcal{H} \).

- To show that \( \hat{\rho}(x) \) is close to \( \rho(x) \), we take two steps:
  1. \( \hat{\rho}(x) \) is concentrated around \( E[\hat{\rho}(x)] \),
  2. \( E[\hat{\rho}(x)] \) is close to \( \rho(x) \).
Some Notation...

- Recall

\[ \rho(x) = \frac{1}{\eta} \ln \left( \frac{\sum_{h: h(x) = +1} w(h)}{\sum_{h: h(x) = -1} w(h)} \right) \]

\[ = \frac{1}{\eta} \ln \left( \sum_{h: h(x) = +1} w(h) \right) - \frac{1}{\eta} \ln \left( \sum_{h: h(x) = -1} w(h) \right) \]

- For any \( K \subseteq H \), define

\[ R(K) = \frac{1}{\eta} \ln \left( \sum_{h \in K} w(h) \right) \]
McDiarmid’s Theorem

Let $X_1, \ldots, X_m \in V^m$ be independent random variables. For $f : V^m \to \mathbb{R}$, if

$$|f(x_1, \ldots, x_m) - f(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_m)| \leq c_i,$$

Then for $\lambda > 0$, $s \in \{-1, +1\}$,

$$\Pr [s(f(X_1, \ldots, X_m) - E[f(X_1, \ldots, X_m)]) \geq \lambda] \leq \exp \left( \frac{-2\lambda^2}{\sum_{i=1}^{m} c_i^2} \right)$$
Applying McDiarmid’s Theorem

- Assume some empirical $\hat{R}(K)$ learned from $m$ samples.
- If we replace $(x_i, y_i)$ with $(x'_i, y'_i)$ in the sample data, we get $\hat{R}'(K)$.
- Then the difference is bounded by the penalty for making a single mistake:

$$\hat{R}'(K) - \hat{R}(K) = \frac{1}{\eta} \ln \left( \frac{\sum_{h \in K} e^{-\eta \hat{\epsilon}'(h)}}{\sum_{h \in K} e^{-\eta \hat{\epsilon}(h)}} \right)$$

$$\leq \frac{1}{\eta} \ln \left( \max_{h \in K} e^{-\eta (\hat{\epsilon}'(h) - \hat{\epsilon}(h))} \right)$$

$$= \max_{h \in K} (\hat{\epsilon}(h) - \hat{\epsilon}'(h)) \leq \frac{1}{m}$$
McDiarmid’s Theorem (continued)

Setting \( c_i = \frac{1}{m} \) for all \( i \), we get:

\[
\Pr \left[ s \left( \hat{R}(K) - E \left[ \hat{R}(K) \right] \right) \geq \lambda \right] \leq \exp \left( \frac{-2\lambda^2}{\sum_{i=1}^{m} \frac{1}{m^2}} \right) = e^{-2\lambda^2 m}
\]

So \( \hat{R}(K) \) is concentrated around \( E \left[ \hat{R}(K) \right] \).
Averaged Classifiers

Generalization Bounds

\[ \hat{\rho}(x) \approx \rho(x) \]
\[ P(x) \approx h^*(x) \]

\[ E[\hat{R}(K)] \approx R(K) \text{ (Lower Bound)} \]

- By convexity of \( \ln \left( \sum e^{x_i} \right) \), we can apply Jensen’s inequality:

\[
\eta E[\hat{R}(K)] = E \left[ \ln \left( \sum_{h \in K} e^{-\eta \hat{\varepsilon}(h)} \right) \right] \geq \ln \left( \sum_{h \in K} e^{-\eta E[\hat{\varepsilon}(h)]} \right) \\
= \ln \left( \sum_{h \in K} e^{-\eta \varepsilon(h)} \right) = \eta R(K)
\]

- In the other direction, we get

\[
E[\hat{R}(K)] \leq R(K) + \frac{\eta}{8m}
\]

- So

\[
R(K) \leq E[\hat{R}(K)] \leq R(K) + \frac{\eta}{8m}
\]
Averaged Classifiers
Generalization Bounds

\[ \hat{\rho}(x) \approx \rho(x) \]
\[ P(x) \approx h^*(x) \]

Back to Theorem 1...

Partition \( \mathcal{H} \) into \( H_+ \) and \( H_- \): the subsets that classify an instance \( x \) as +1 and −1 respectively.

Then

\[
\rho(x) - \hat{\rho}(x) = R(H_+) - R(H_-) - \hat{R}(H_+) + \hat{R}(H_-) \\
= R(H_+) - \hat{R}(H_+) - (R(H_-) - \hat{R}(H_-))
\]

We know that

1. \( \Pr \left[ R(H_+) - \hat{R}(H_+) > \lambda \right] \leq e^{-2\lambda^2 m} \),
2. \( \Pr \left[ \hat{R}(H_-) - R(H_-) > \lambda + \frac{\eta}{8m} \right] \leq e^{-2\lambda^2 m} \),

so the probability of both occurring is bounded:

\[
\Pr \left[ s(\rho(x) - \hat{\rho}(x)) \geq \lambda + \lambda + \frac{\eta}{8m} \right] \leq 2e^{-2\lambda^2 m}.
\]
Theorem 2: \( \hat{P}(x) \approx \text{sign}(\rho(x)) \)

- If \( \hat{P}(x) \neq 0 \) (we make a prediction), then \( |\hat{\rho}(x)| \geq \Delta \).
- By Theorem 1, we can set \( \Delta \) such that the probability of incorrect prediction is bounded:
  \[
  \Pr_{(x,y) \sim D} [ |\hat{\rho}(x)| \geq \Delta \land \text{sign}(\hat{\rho}(x)) \neq \text{sign}(\rho(x))] \leq \delta
  \]
  where \( \delta \) is a function of \( \Delta, \eta, m \).
- This holds with probability \( 1 - \delta \) over the random training set.
Parameters: \( \text{sign}(\rho(x)) \approx h^*(x) \)

- For \( q \in (0, 1/2) \) and \( \delta > 0 \), we can set the parameters as follows:

  \[
  \eta = \ln(8|\mathcal{H}|)m^{1/2-q} \quad \Delta = 2\sqrt{\frac{1}{m} \ln(\sqrt{2}/\delta)} + \frac{\ln(8|\mathcal{H}|)}{8m^{1/2+q}}
  \]

- For \( m \geq 8 \),

  \[
  \Pr_{(x,y) \sim D} [\text{sign}(\rho(x)) \neq y] = 2\varepsilon(h^*) + O \left( \frac{\ln(m)}{m^{1/2-q}} \right)
  \]

  \[
  \Pr_{(x,y) \sim D} [|\hat{\rho}(x)| > \Delta \wedge \text{sign}(\hat{\rho}(x)) \neq y] = 2\varepsilon(h^*) + O \left( \frac{\ln(m)}{m^{1/2-q}} \right) + \delta
  \]
We can also bound the probability of abstaining from prediction.

If

\[ m \geq \left[ 8 \sqrt{\ln \frac{\sqrt{2}}{\delta} \ln(8|\mathcal{H}|)} \right]^{1/q}, \]

then

\[ \Pr_{(x,y) \sim D} [y \rho(x) \leq 2\Delta] = 5\varepsilon(h^*) + O \left( \frac{\sqrt{-\ln \delta + \ln |\mathcal{H}|}}{m^{1/2-q}} \right) \]
Conclusions

- The empirical log ratio $\hat{\rho}(x)$ is a good approximation of the true ratio $\rho(x)$ (from Theorem 1).
- $\rho(x)$ is stable and independent of the size of the hypothesis class (from McDiarmid’s theorem).
- $\rho(x)$ behaves almost as well as the best hypothesis $h^* \in \mathcal{H}$. 
References