Learning with an Unreliable Teacher

Elizabeth M. Sweeney

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

December 11, 2014

Learning with an Unreliable Teacher

Lugosi, Gabor. "Learning with an unreliable teacher." Pattern Recognition 25.1 (1992): 79-87.

Devroye, Luc, Laszlo Gyorfi, and Gabor Lugosi. A probabilistic theory of pattern recognition. Springer Verlag, 1996.

The Classification Task

1. We have random variables (X, Y) such that $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$.
2. Let $g : \mathbb{R}^d \to \{0, 1\}$
3. The error probability is $P_g(X, Y) = \mathbb{P}\{g(X) \neq Y\}$
4. The classification task is to find g that minimizes $P_g(X, Y)$

The Classification Task

1. In practice we have a training sequence $\xi_n = ((X_1, Y_1), (X_2, Y_2), ...(X_n, Y_n))$ where (X_k, Y_k) are iid from (X, Y)
2. We then estimate Y from $g_n(X, \xi_n)$, minimizing $\mathbb{P}\{g_n(X, \xi_n) \neq Y\}$

Learning with an Unreliable Teacher

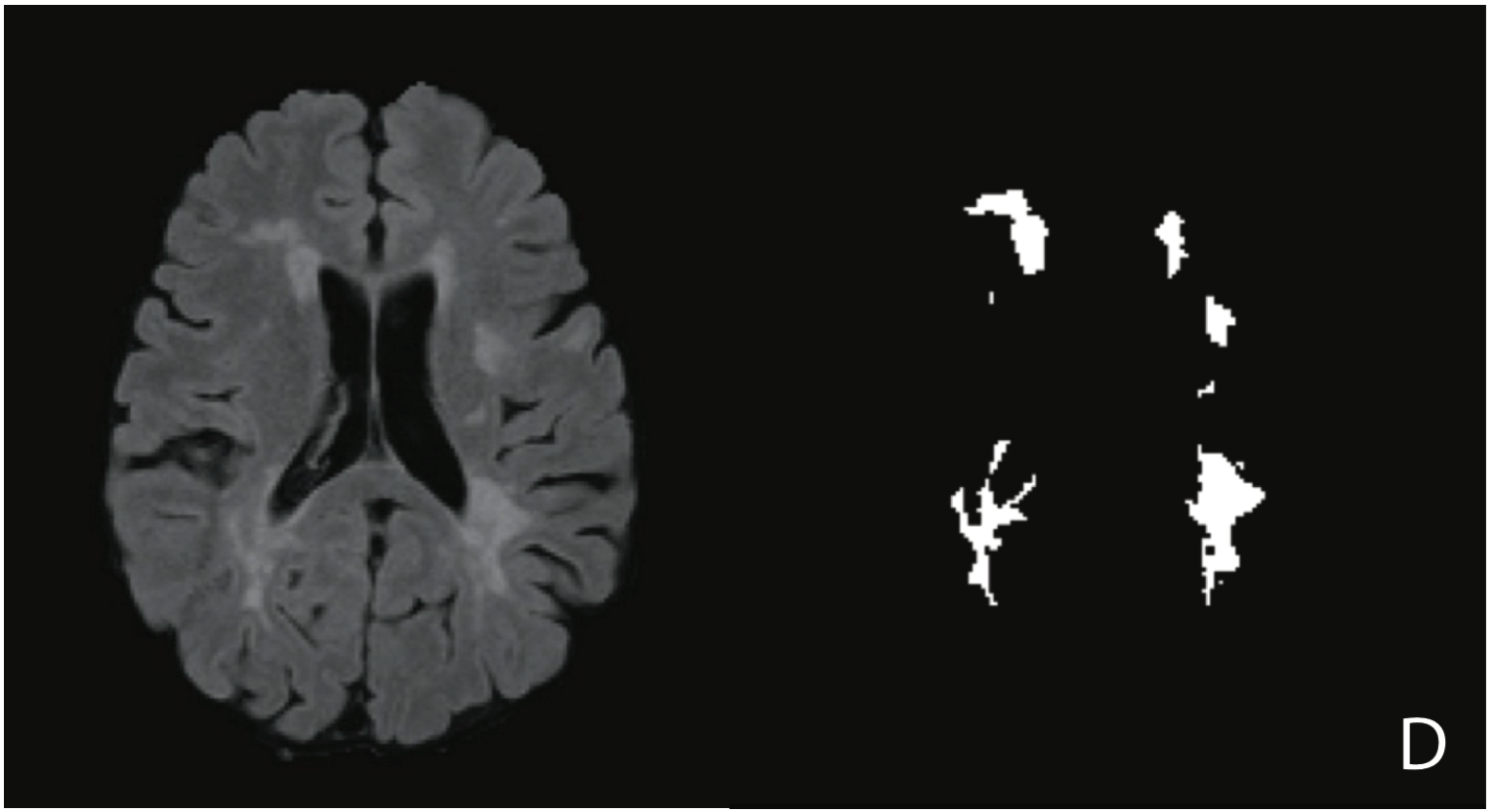
Classification Methods

1. Bayes decision rule
2. One nearest neighbor

Mislabeling Mechanism

1. Discrete Memoryless Channel
2. Misprints in the Training Sequence
3. Consequently Lying Teacher

Some Context



The Bayes Decision Rule

We wish to find that g such that the error probability $P_g(X, Y) = \mathbb{P}\{g(X) \neq Y\}$ is minimized. Let $p_i(x) = \mathbb{P}\{Y = i | X = x\}$, $i \in \{0, 1\}$. The optimal solution is the Bayes decision rule:

$$g^*(x) = \text{argmax}_{0 \leq i \leq 1} p_i(x)$$

Where optimality is defined as $\mathbb{P}\{g^*(X) \neq Y\} \leq \mathbb{P}\{g(X) \neq Y\}$ for all g. We will denote $\mathbb{P}\{g^*(X) \neq Y\}$ as $P^B(X, Y)$, which is called the Bayes risk.

The Bayes Decision Rule

We will often not know $p_i(x)$, and so we will estimate the with $q_i(x, \xi)$ with training data ξ and will use the corresponding decision rule

$$g(x, \xi) = \text{argmax}_i q_i(x, \xi)$$

Lemma 1.1

The error of probability of the decision g is close to the Bayes risk if the q_i are good L_1 approximations of the true posterior probabilities

$$\mathbb{P}\{g(X,\xi) \neq Y\} - P^B(X,Y) \leq E\left(\sum_{i=1}^{1} |p_i(X) - q_i(X,\xi)|\right)$$

Lemma 1.3

Let $q_0(x)$ and $q_1(x)$ be real valued measurable functions defined on $\mathbb{R}^d$. Let the decision function g be the following:

$$g(x) = \text{argmax}_i q_i(x)$$

If this maximum is unique almost everywhere then for any sequence of measurable functions $\bar{q}_i^{(n)}(x, s)$ $(i = 0, 1; n = 1, 2, ...)$, for which

$$\lim_{n \to \infty} E\left(\sum_{i=0}^{1} |q_i(X) - \bar{q}_i^{(n)}(X, \xi)| \right) = 0$$

Lemma 1.3

then

$$\lim \left| P_g\left(X, Y\right) - P_{\bar{g}^{(n)}}\left(X, Y\right) \right| = 0$$

where

$$\bar{g}^{(n)}\left(x, s\right) = \text{argmax}_i \bar{q}_i^{(n)}\left(x, s\right)$$

"Every decision based on maximization of measurable functions can be arbitrarily approximated by approximating the function in the L_1 sense "

Proof of Lemma 1.3

Proof: We first note that

$$\begin{aligned} P_g(X,Y) &= \mathbb{P}(g(X) \neq Y) \\ &= 1 - \mathbb{P}(g(X) = Y) \\ &= 1 - E\left[I_{(g(X)=Y)}\right] \end{aligned}$$

The difference between the error of probabilities is then

$$\begin{aligned} |P_g(X,Y) - P_{\bar{g}^{(n)}}(X,Y)| &= \left| E\left[I_{\left(\bar{g}^{(n)}(X,\xi)=Y\right)}\right] - E\left[I_{(g(X)=Y)}\right] \right| \\ &= \left| E\left[I_{\left(\bar{g}^{(n)}(X,\xi)=Y\right)} - I_{(g(X)=Y)}\right] \right| \end{aligned}$$

Proof of Lemma 1.3

Then we have that

$$\begin{aligned}\left|E\left[I_{\left(\bar{g}^{(n)}(X,\xi)=Y\right)}-I_{(g(X)=Y)}\right]\right| &\leq E\left[\left|I_{\left(\bar{g}^{(n)}(X,\xi)=Y\right)}-I_{(g(X)=Y)}\right|\right] \\ &= E\left[I_{\left(\bar{g}^{(n)}(X,\xi)\neq g(X)\right)}\right] \\ &= \mathbb{P}\{g(X)\neq \bar{g}^{(n)}(X,\xi)\}\end{aligned}$$

Proof of Lemma 1.3

Therefore

$$\begin{aligned} |P_g(X,Y) - P_{\bar{g}^{(n)}}(X,Y)| &\leq \mathbb{P}\{g(X) \neq \bar{g}^{(n)}(X,\xi)\} \\ &= \mathbb{P}\{g(X) = 1, \bar{g}^{(n)}(X,\xi) = 0\} \\ &+ \mathbb{P}\{g(X) = 0, \bar{g}^{(n)}(X,\xi) = 1\} \end{aligned}$$

Proof of Lemma 1.3

By symmetry, we need only show that as $n \to \infty$

$$\mathbb{P}\{g(X) = 1, \bar{g}^{(n)}(X, \xi) = 0\} \to 0$$

For all $\delta > 0$ we have that

$$\begin{aligned}
\mathbb{P}\{g(X) = 1, \bar{g}^{(n)}(X, \xi) = 0\} &= \mathbb{P}\{q_1(X) \geq q_0(X), \bar{q}_1^n(X, \xi) \leq \bar{q}_0^n(X, \xi)\} \\
&\leq \mathbb{P}\{|q_1(X) - q_0(X)| < \delta\} \\
&\quad + \mathbb{P}\left\{\sum_{i=0}^{1} |q_i(X) - \bar{q}_i^{(n)}(X, \xi)| \geq \delta\right\}
\end{aligned}$$

Proof of Lemma 1.3

By symmetry, we need only show that as $n \to \infty$

$$\mathbb{P}\{g(X) = 1, \bar{g}^{(n)}(X, \xi) = 0\} \to 0$$

For all $\delta > 0$ we have that

$$\begin{aligned}\mathbb{P}\{g(X) = 1, \bar{g}^{(n)}(X, \xi) = 0\} &= \mathbb{P}\{q_1(X) \geq q_0(X), \bar{q}_1^n(X, \xi) \leq \bar{q}_0^n(X, \xi)\} \\ &\leq \mathbb{P}\{|q_1(X) - q_0(X)| < \delta\} \\ &+ \mathbb{P}\left\{\sum_{i=0}^{1} |q_i(X) - \bar{q}_i^{(n)}(X, \xi)| \geq \delta\right\}\end{aligned}$$

Learning with an Unreliable Teacher

Gabor Lugosi 11:39 AM (1 hour ago)

to me

Dear Elizabeth,

This was so long ago that I totally forgot the details. I checked the step of the proof you ask about and you are right, it doesn't seem to make any sense. Luckily, I think the lemma is still true, it shouldn't be difficult to fix it.

Thanks for your interest! Best regards,

Gabor

Proof of Lemma 1.3

We now show that for any $\epsilon > 0$, we can chose δ and N such that if $n > N$ we have

$$\mathbb{P}\{|q_1(X) - q_0(X)| < \delta\} + \mathbb{P}\left\{\sum_{i=0}^{1} |q_i(X) - \bar{q}_i^{(n)}(X, \xi)| \geq \delta\right\} < \epsilon$$

Proof of Lemma 1.3

We have that the maximum of $q_i(x)$ is unique almost everywhere, therefore

$$|q_1(X) - q_0(X)| > 0$$

with probability 1 and so a δ can be selected such that

$$\mathbb{P}\{|q_1(X) - q_0(X)| < \delta\} < \frac{\epsilon}{2}$$

Proof of Lemma 1.3

By the L_1 convergence of the function $\bar{q}_i^{(n)}(x, s)$

$$\lim_{n\to\infty} \mathbb{P}\left\{\sum_{i=0}^{1} |q_i(X) - \bar{q}_i^{(n)}(X, \xi)| \geq \delta\right\} = 0$$

therefore, there exists an N such that for every $n > N$ we have

$$\mathbb{P}\left\{\sum_{i=0}^{1} |q_i(X) - \bar{q}_i^{(n)}(X, \xi)| \geq \delta\right\} < \frac{\epsilon}{2}$$

□

Labeling Errors in the Training Set

1. We deal with the case where instead of knowing Y_k we have the labels $Z_k \in \{0, 1\}$ which are erroneous labels
2. Let (X_k, Y_k, Z_k) be iid from (X, Y, Z)
3. We now wish to estimate Y given the data X from the training set $\eta_n = ((X_1, Z_1), (X_2, Z_2), ...(X_n, Z_n))$

Bayes Decision Rule with Mislabeled Training Data

Let $q_i(x) = \mathbb{P}\{Z = i|X = x\}$. We assume that we have an L_1 consistent estimator of $q_i(x)$ and we will denote this estimator by $q_{in}(x) = q_{in}(x, \eta_n)$. By L_1 consistency we will mean

$$\lim_{n\to\infty} E\left(\sum_{i=0}^{1} |q_i(X) - q_{in}(X)|\right) = 0$$

and the corresponding decision rule will be

$$g_n(x) = \text{argmax}_i q_{in}(x)$$

which is an approximation of $g(x) = \text{argmax}_i q_i(x)$

Discrete Memoryless Channel

Here we assume that the true labels Y_k are transmitted over a binary memoryless channel to get the training labels Z_k.

$$\begin{aligned} a_{ji} &= \mathbb{P}\{Z = i | Y = j, X = x\} \\ &= \mathbb{P}\{Z = i | Y = j\} \end{aligned}$$

Discrete Memoryless Channel

Let

$$A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

then $\mathbf{q}(x) = \begin{pmatrix} q_0(x) & q_1(x) \end{pmatrix}$ and $\mathbf{p}(x) = \begin{pmatrix} p_0(x) & p_1(x) \end{pmatrix}$. Then we have that

$$\mathbf{q}(x) = \mathbf{p}(x) A$$

Known Discrete Memoryless Channel

We focus on the case where the discrete memoryless channel is known. Let $\mathbf{q}_n(x) = \begin{pmatrix} q_{0n}(x) & q_{1n}(x) \end{pmatrix}$. Let $\mathbf{p}_n(x)$ be the solution to the equation $\mathbf{q}_n(x) = \mathbf{p}_n(x) A$. Then define

$$f_n(x) = \text{argmax}_i p_{in}(x)$$

Theorem 2.2

As long as $p + q \neq 1$ the decision f_n is asymptotically optimal:

$$P_{f_n}(X, Y) - P^B(X, Y) \leq \left| \frac{1 + |p - q|}{1 - p - q} \right| E\left(\sum_{i=0}^{1} |q_i(X) - q_{in}(X)| \right)$$

Proof Theorem 2.2

We begin by introducing the following notation

$$\delta_{in}(x) = q_i(x) - q_{in}(x)$$
$$\gamma_{in}(x) = p_i(x) - p_{in}(x)$$

And let $\Delta_n = \begin{pmatrix} \delta_{0n}(x) & \delta_{1n}(x) \end{pmatrix}$ and $\Gamma_n = \begin{pmatrix} \gamma_{0n}(x) & \gamma_{1n}(x) \end{pmatrix}$. Then, as $\mathbf{q}(x) = \mathbf{p}(x) A$ and $\mathbf{q}_n(x) = \mathbf{p}_n(x) A$, it follows that

$$\Delta_n(x) = \Gamma_n(x) A$$

Proof Theorem 2.2

Assuming A is invertible, we have

$$\Delta_n(x) A^{-1} = \Gamma_n(x)$$

We then take the L_1 norm of both sides of the equation

$$\sum_{i=0}^{1} |p_i(x) - p_{in}(x)| = \sum_{i=0}^{1} |q_i(x) - q_{in}(x)| \|A^{-1}\|_1$$

where

$$\|A^{-1}\|_1 = \left| \frac{1 + |p - q|}{1 - p - q} \right|$$

Proof Theorem 2.2

Taking the expectation of both sides gives us

$$E\sum_{i=0}^{1}\left|p_i\left(x\right)-p_{in}\left(x\right)\right|=\left|\frac{1+\left|p-q\right|}{1-p-q}\right|E\sum_{i=0}^{1}\left|q_i\left(x\right)-q_{in}\left(x\right)\right|$$

Applying Lemma 1.1 will give us that

$$P_{f_n}\left(X,Y\right)-P^B\left(X,Y\right)\leq E\sum_{i=0}^{1}\left|p_i\left(x\right)-p_{in}\left(x\right)\right|$$

and therefore

$$P_{f_n}\left(X,Y\right)-P^B\left(X,Y\right)\leq\left|\frac{1+\left|p-q\right|}{1-p-q}\right|E\left(\sum_{i=0}^{1}\left|q_i\left(X\right)-q_{in}\left(X\right)\right|\right)$$

Consequently Lying Teacher

For the consequently lying teacher, we have that for $h : \mathbb{R}^d \to \{0, 1\}$

$$Z = h(X)$$

And in this case

$$q_i(x) = \mathbb{P}\{Z = i | X = x\} = \mathbb{P}\{h(X) = i | X = x\}$$

and so

$$q_i(x) = \begin{cases} 1 & \text{if h(x) = i} \\ 0 & \text{otherwise} \end{cases}$$

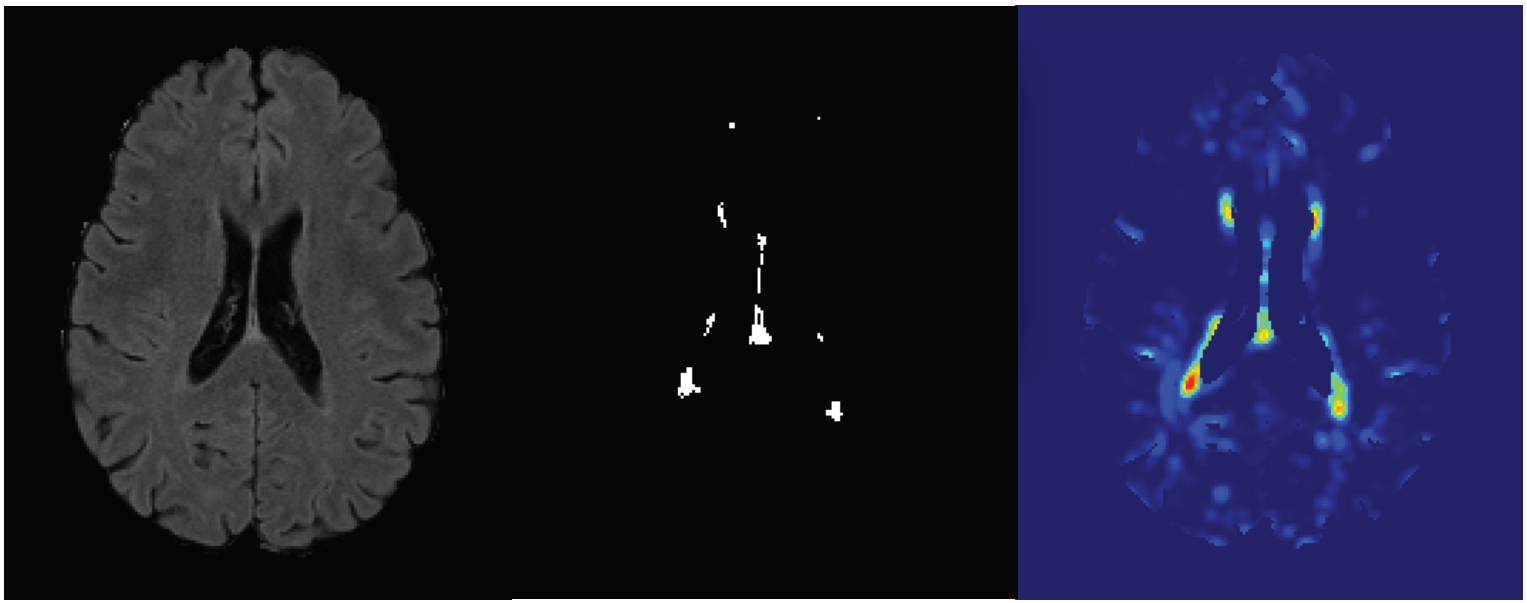
therefore $g = h$ and we have that

$$\mathbb{P}\{g(X) \neq Y\} = \mathbb{P}\{Y \neq Z\}$$

Conclusion

1. With the discrete memoryless channel for mislabeling, we can reach Bayes optimal classification error
2. With the consequently lying teacher, as long as there is mislabeling, we never reach Bayes optimal

Consequently Lying Teacher



Supplemental Materials

1. Optimality of the Bayes Decision Rule
2. Misprints in the Training Sequence

Optimality of the Bayes Decision Rule

Proof: Let $p_i(x) = \mathbb{P}\{Y_i = i | X = x\}$. For any classifier g we have that

$$\begin{aligned}
\mathbb{P}\{g(X) \neq Y | X = x\} &= 1 - \mathbb{P}\{g(X) = Y | X = x\} \\
&= 1 - \mathbb{P}\{g(X) = 0, Y = 0 | X = x\} \\
&\quad - \mathbb{P}\{g(X) = 1, Y = 1 | X = x\} \\
&= 1 - I(g(x) = 0) p_0(x) - I(g(x) = 1) p_1(x) \\
&= 1 - I(g(x) = 0) p_0(x) - I(g(x) = 1)(1 - p_0(x))
\end{aligned}$$

Optimality of the Bayes Decision Rule

Then we have that

$$
\begin{aligned}
&\mathbb{P}\{g(X) \neq Y | X = x\} - \mathbb{P}\{g^*(X) \neq Y | X = x\} \\
&= p_0(x)\left(I\left(g^*(x) = 0\right) - I\left(g(x) = 0\right)\right) + \left(1 - p_0(x)\right)\left(I\left(g^*(x) = 1\right)\right. \\
&\left. - I\left(g(x) = 1\right)\right) \\
&= p_0(x)\left(I\left(g^*(x) = 0\right) - I\left(g(x) = 0\right)\right) + \left(1 - p_0(x)\right)\left(\left(1 - I\left(g^*(x) = 0\right)\right)\right. \\
&\left. - \left(1 - I\left(g(x) = 0\right)\right)\right) \\
&= p_0(x)\left(I\left(g^*(x) = 0\right) - I\left(g(x) = 0\right)\right) + \left(p_0(x) - 1\right)\left(I\left(g^*(x) = 0\right)\right. \\
&\left. - I\left(g(x) = 0\right)\right) \\
&= \left(2p_0(x) - 1\right)\left(I\left(g(x)^* = 0\right) - I\left(g(x) = 0\right)\right)
\end{aligned}
$$

Optimality of the Bayes Decision Rule

And it follows that

$$(2p_0(x) - 1)\left(I\left(g(x)^* = 0\right) - I\left(g(x) = 0\right)\right) \geq 0$$

as

$$g^*(x) = \text{argmax}_{0 \leq i \leq 1} \mathbb{P}\{Y = i | X = x\}$$

therefore when $I\left(g^*(x) = 0\right) = 1$, $p_0(x) > \frac{1}{2}$ and when $I\left(g^*(x) = 0\right) = 0$, $p_0(x) < \frac{1}{2}$. The proof is completed by integrating both sides with respect to $\mu(dx)$. $\square$

Misprints in the Training Sequence

The labels Z_i take an arbitrary value with probability p that is independent from X_i and Y_i

$$Z_i = \beta_i Y_i + (1 - \beta_i) W_i$$

with $\beta_i \in (0, 1)$, $W_i \in (0, 1)$ and

$$\mathbb{P}(\beta_i = 0) = p$$

Note that W_i may depend of X_i and Y_i

Theorem 2.4

When we have misprints in the training sequence the error probability is bounded in the following way for the Bayes decision rule

$$P_g(X, Y) \leq P^B(X, Y) \frac{1}{1 - 2p}$$