# LEARNING MAY WORK…

Matthieu R Bloch

January 10, 2019



## LOGISTICS

Registration update

Lecture videos on Canvas
- Media gallery
- Please keep coming to class!

Self-assessment online here
- Due Friday January 18, 2019 (11:59PM EST) (Friday January 25, 2019 for DL)

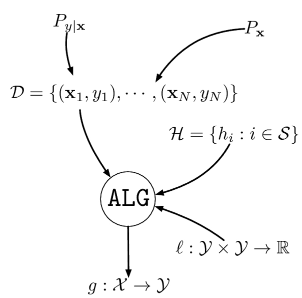
Lecture slides and notes
- I will make every effort to post ahead of time


http://www.phdcomics.com



## RECAP: COMPONENTS OF SUPERVISED MACHINE LEARNING

1. A *dataset* $\mathcal{D} \triangleq \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$
   - $\{\mathbf{x}_i\}_{i=1}^N$ *drawn i.i.d. from an unknown probability distribution* $P_{\mathbf{x}}$ on $\mathcal{X}$
   - $\{y_i\}_{i=1}^N$ are the corresponding targets $y_i \in \mathcal{Y} \triangleq \mathbb{R}$
2. An *unknown conditional distribution* $P_{y|\mathbf{x}}$
   - $P_{y|\mathbf{x}}$ models $f : \mathcal{X} \to \mathcal{Y}$ *with noise*
3. A *set of hypotheses* $\mathcal{H}$ as to what the function could be
4. A *loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ capturing the "cost" of prediction
5. An *algorithm* ALG to find the best $h \in \mathcal{H}$ that explains $f$




Learning model



## RECAP: THE SUPERVISED LEARNING PROBLEM

Learning is not *memorizing*
- Our goal is *not* to find $h \in \mathcal{H}$ that accurately assigns values to elements of $\mathcal{D}$
- Our goal is to find the *best* $h \in \mathcal{H}$ that accurately *predicts* values of *unseen* samples

Consider hypothesis $h \in \mathcal{H}$. We can easily compute the *empirical risk* (a.k.a. *in-sample* error)

$$\widehat{R}_N(h) \triangleq \frac{1}{N}\sum_{i=1}^N \ell(y_i, h(\mathbf{x}_i))$$

What we really care about is the *true risk* (a.k.a. *out-sample* error)

$$R(h) \triangleq \mathbb{E}_{\mathbf{x}y}(\ell(y, h(\mathbf{x})))$$

*Question #1:* Can *generalize*?
- For a given $h$, is $\widehat{R}_N(h)$ close to $R(h)$?

*Question #2:* Can we learn *well*?
- Given $\mathcal{H}$, the *best* hypothesis is $h^\sharp \triangleq \operatorname{argmin}_{h\in\mathcal{H}} R(h)$
- Our algorithm can only find $h^* \triangleq \operatorname{argmin}_{h\in\mathcal{H}} \widehat{R}_N(h)$
- Is $\widehat{R}_N(h^*)$ close to $R(h^\sharp)$?
- Is $R(h^\sharp) \approx 0$?

## WHY THE QUESTIONS MATTERS

*Quick demo:* nearest neighbor classification



## A SIMPLER LEARNING PROBLEM

Consider a special case of the general supervised learning problem

1. Dataset $\mathcal{D} \triangleq \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$
   - $\{\mathbf{x}_i\}_{i=1}^N$ *drawn i.i.d. from unknown* $P_{\mathbf{x}}$ on $\mathcal{X}$
   - $\{y_i\}_{i=1}^N$ labels with $\mathcal{Y} = \{0, 1\}$ (binary classification)
2. Unknown $f : \mathcal{X} \to \mathcal{Y}$, no noise.
3. Finite set of hypotheses $\mathcal{H}$, $|\mathcal{H}| = M < \infty$
   - $\mathcal{H} \triangleq \{h_i\}_{i=1}^M$
4. Binary loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+ : (y_1, y_2) \mapsto \mathbf{1}\{y_1 \neq y_2\}$

In this very specific case, the true risk simplifies

$$R(h) \triangleq \mathbb{E}_{\mathbf{x}y}(\mathbf{1}\{h(\mathbf{x}) \neq y\}) = \mathbb{P}_{\mathbf{x}y}(h(\mathbf{x}) \neq y)$$

The empirical risk becomes

$$\widehat{R}_N(h) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{h(\mathbf{x}_i) \neq y\}$$



## CAN WE LEARN?

Our objective is to find a hypothesis $h^*$ that ensures a small risk

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{R}_N(h)$$

For a *fixed* $h_j \in \mathcal{H}$, how does $\widehat{R}_N(h_j)$ compares to $R(h_j)$?

Observe that for $h_j \in \mathcal{H}$

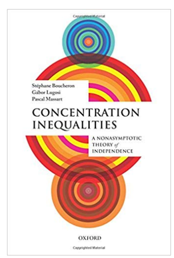
- The empirical risk is a sum of iid random variables

$$\widehat{R}_N(h_j) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{h_j(\mathbf{x}_i) \neq y\}$$

- $\mathbb{E}(\widehat{R}_N(h_j)) = R(h_j)$

$\mathbb{P}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| > \epsilon\right)$ is a statement about the deviation of a normalized sum of iid random variables from its mean

We're in luck! Such bounds, a.k.a, known as *concentration inequalities*, are a well studied subject






## CONCENTRATION INEQUALITIES 101

**Lemma (Markov's inequality)**

Let $X$ be a *non-negative* real-valued random variable. Then for all $t > 0$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

**Lemma (Chebyshev's inequality)**

Let $X$ be a real-valued random variable. Then for all $t > 0$

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

**Proposition (Weak law of large numbers)**

Let $\{X_i\}_{i=1}^N$ be i.i.d. real-valued random variables with finite mean $\mu$ and finite variance $\sigma^2$. Then

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{N\epsilon^2} \qquad \lim_{N\to\infty} \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \geq \epsilon\right) = 0.$$

# BACK TO LEARNING

By the law of large number, we know that

$$\forall \epsilon > 0 \quad \mathbb{P}_{\{(\mathbf{x}_i, y_i)\}} \left( \left| \widehat{R}_N(h_j) - R(h_j) \right| \geq \epsilon \right) \leq \frac{\mathrm{Var}(\mathbf{1}\{h_j(\mathbf{x}_1) \neq y\})}{N\epsilon^2} \leq \frac{1}{N\epsilon^2}$$

Given enough data, we can *generalize*

How much data? $N = \frac{1}{\delta\epsilon^2}$ to ensure $\mathbb{P}\left( \left| \widehat{R}_N(h_j) - R(h_j) \right| \geq \epsilon \right) \leq \delta$.

That's not quite enough! We care about $\widehat{R}_N(h^*)$ where $h^* = \operatorname{argmin}_{h \in \mathcal{H}} \widehat{R}_N(h)$

- If $M = |\mathcal{H}|$ is large we should expect the existence of $h_k \in \mathcal{H}$ such that $\widehat{R}_N(h_k) \ll R(h_k)$

$$\mathbb{P}\left( \left| \widehat{R}_N(h^*) - R(h^*) \right| \geq \epsilon \right) \leq ?$$

$$\mathbb{P}\left( \left| \widehat{R}_N(h^*) - R(h^*) \right| \geq \epsilon \right) \leq \mathbb{P}\left( \exists j : \left| \widehat{R}_N(h^*) - R(h^*) \right| \geq \epsilon \right)$$

$$\mathbb{P}\left( \left| \widehat{R}_N(h^*) - R(h^*) \right| \geq \epsilon \right) \leq \frac{M}{N\epsilon^2}$$

We need $N \geq \frac{M}{\delta\epsilon^2}$ to ensure $\mathbb{P}\left( \left| \widehat{R}_N(h^*) - R(h^*) \right| \geq \epsilon \right) \leq \delta$.



# CONCENTRATION INEQUALITIES 102

We can obtain *much* better bounds than with Chebyshev

**Lemma (Hoeffding's inequality)**

Let $\{X_i\}_{i=1}^N$ be i.i.d. real-valued zero-mean random variables such that $X_i \in [a_i; b_i]$. Then for all $\epsilon > 0$

$$\mathbb{P}\left( \left| \frac{1}{N} \sum_{i=1}^{N} X_i \right| \geq \epsilon \right) \leq 2 \exp\left( -\frac{2N^2\epsilon^2}{\sum_{i=1}^{N}(b_i - a_i)^2} \right).$$

In our learning problem

$$\forall \epsilon > 0 \quad \mathbb{P}\left( \left| \widehat{R}_N(h_j) - R(h_j) \right| \geq \epsilon \right) \leq 2\exp(-2N\epsilon^2)$$

$$\forall \epsilon > 0 \quad \mathbb{P}\left( \left| \widehat{R}_N(h^*) - R(h^*) \right| \geq \epsilon \right) \leq 2M\exp(-2N\epsilon^2)$$

We need $N \geq \frac{1}{2\epsilon^2}\left(\log M + \log \frac{2}{\delta}\right)$

$M$ can be quite large (almost exponential in $N$) and, with enough data, we can generalize $h^*$.

How about learning $h^\sharp \triangleq \operatorname{argmin}_{h \in \mathcal{H}} R(h)$?



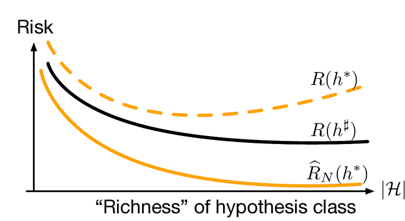
# LEARNING CAN WORK!

**Lemma.**

If $\forall j \in \mathcal{H} \left| \widehat{R}_N(h_j) - R(h_j) \right| \leq \epsilon$ then $\left| R(h^*) - R(h^\sharp) \right| \leq 2\epsilon$.

How do we make $R(h^\sharp)$ small?

- Need bigger hypothesis class $\mathcal{H}$! (could we take $M \to \infty$?)
- Fundamental trade-off of learning



# WHAT IS A GOOD HYPOTHESIS?

Ideally we want $|\mathcal{H}|$ small so that $R(h^*) \approx R(h^\sharp)$ and get lucky s that $R(h^*) \approx 0$

In general this is *not* possible

- Remember, we usually have to learn $P_{y|\mathbf{x}}$, not a function $f$

Next time

- What is the optimal binary classification hypothesis class?
- How small can $R(h^*)$ be?