LEARNING

Dr. Matthieu R Bloch

Monday, December 6, 2021

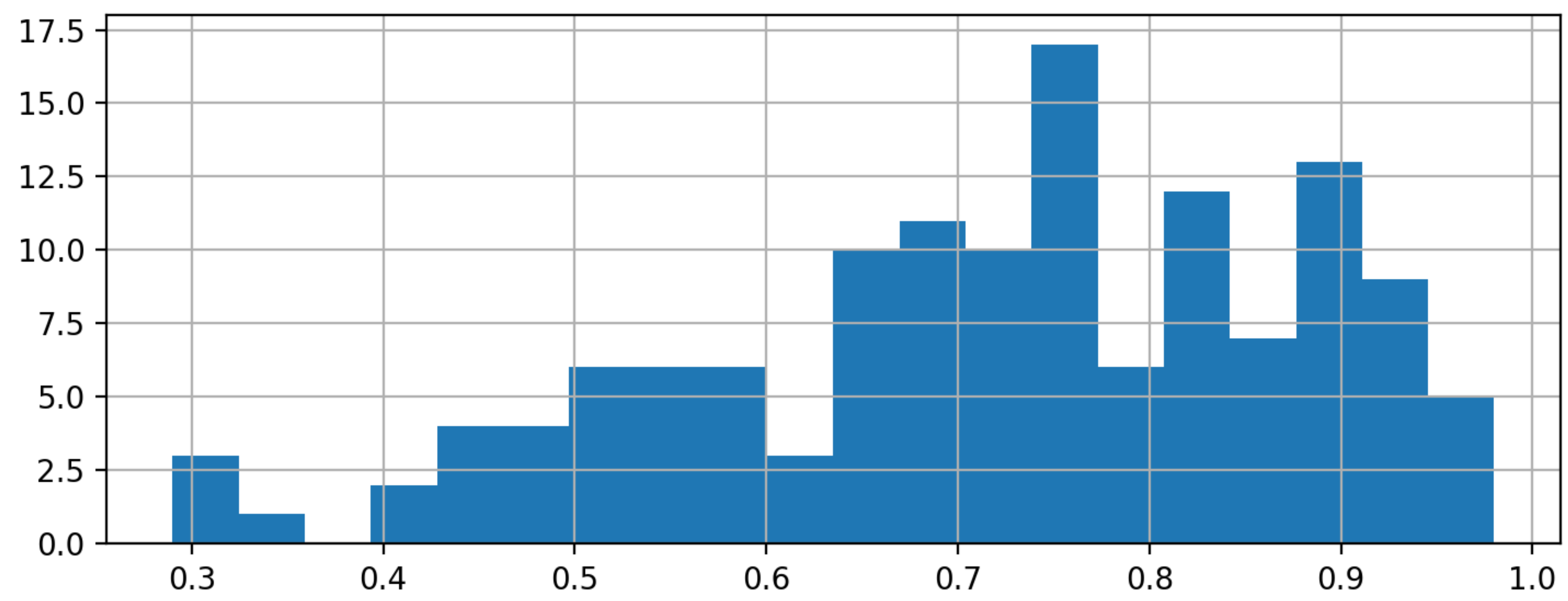
LOGISTICS

General announcements

- Assignment 6 due December 7, 2021 for bonus, deadline December 10, 2021
- Last lecture!
- Let me know what's missing
- Expect an email from me tonight

Midterm 2 statistics

- *Overall:* AVG: 72% - MIN: 29% - MAX: 98%



WHAT WE HAVE LEARNED THIS FALL

Hilbert spaces

- Spaces of functions can be manipulated almost just as easily
- Finite dimensional is fairly natural
- Infinite dimensional can be manipulated just as well using *orthobases*
- With orthobases, vectors in infinite dimensional separates Hilbert spaces are like *square summable sequences*

Regression

- Who knew solving $\mathbf{y} = \mathbf{A}\mathbf{x}$ could be so useful?
- SVD provides lots of insights

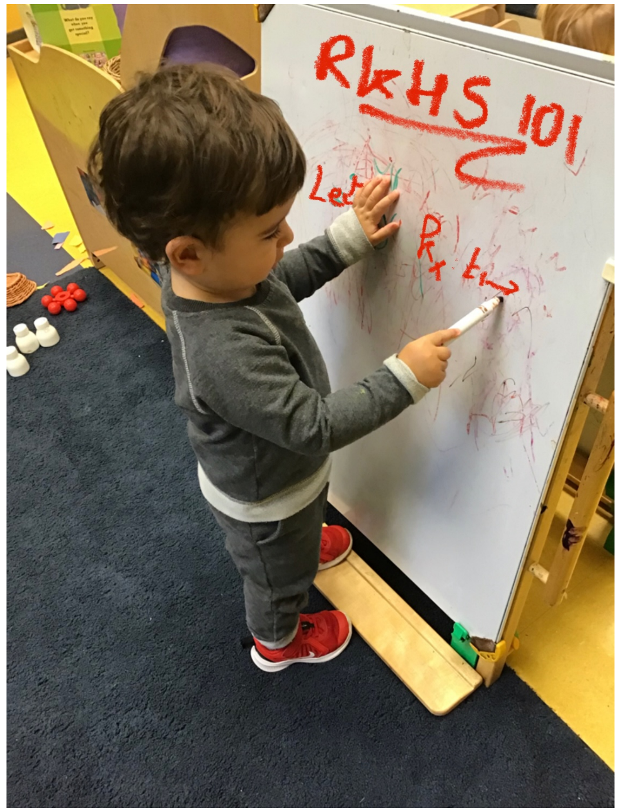
Regression in Hilbert spaces

- Perhaps biggest lesson of the course
- Representer theorem allows us to do regression in infinite dimensional Hilbert spaces
- RKHS provide the kind of Hilbert spaces that naturally embed our data

WHAT'S ON THE AGENDA FOR TODAY?

More on learning and Bayes classifiers

Lecture notes 17 and 23



Toddlers can do it!

A SIMPLER SUPERVISED 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

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

The empirical risk becomes

we can compute

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

CAN WE LEARN?

Our objective is to find a hypothesis $h^* = \operatorname{argmin}_{h \in \mathcal{H}} \widehat{R}_N(h)$ that ensures a small risk

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

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

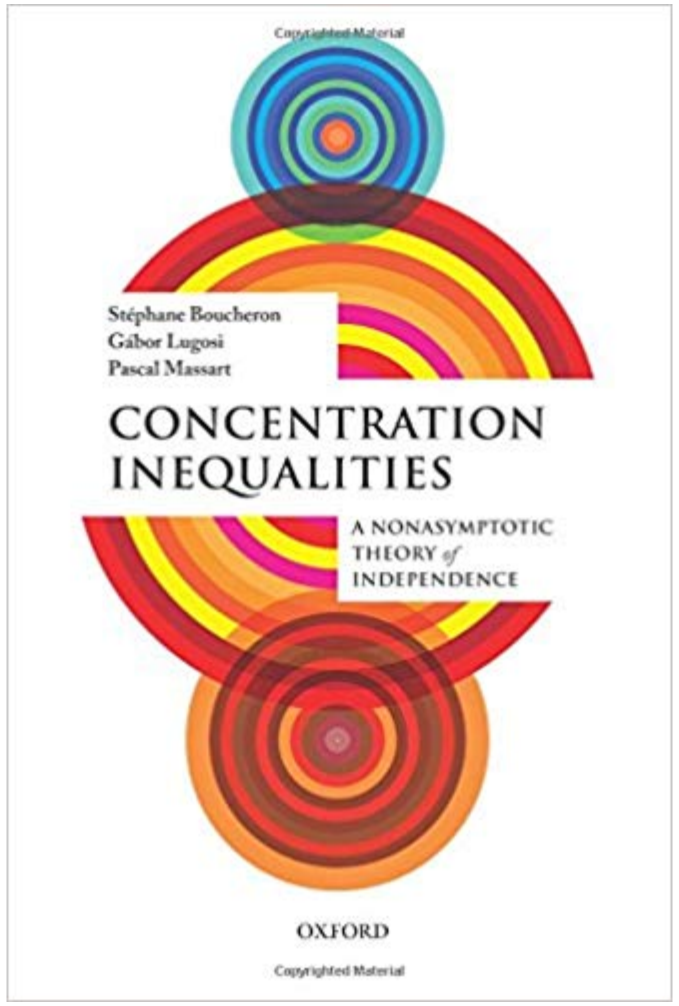
- The empirical risk is a sum of iid random variables

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

iid random variables

- $\mathbb{E}\left[\widehat{R}_N(h_j)\right] = R(h_j)$

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



CAN WE LEARN?

Our objective is to find a hypothesis $h^* = \operatorname{argmin}_{h \in \mathcal{H}} \widehat{R}_N(h)$ that ensures a small risk

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

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

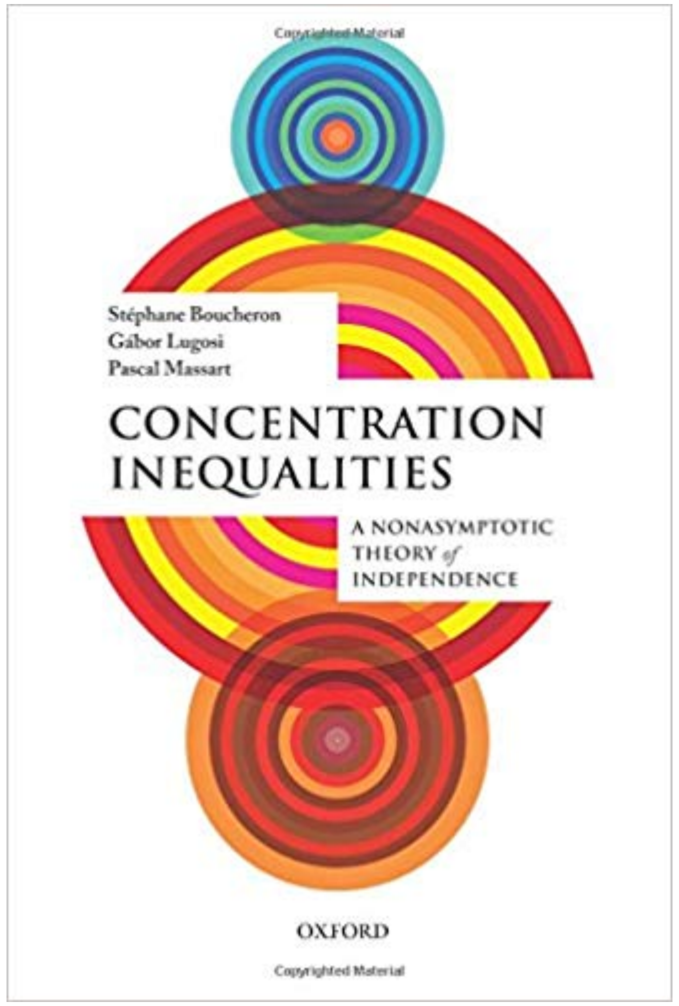
- The empirical risk is a sum of iid random variables

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

- $\mathbb{E}\left[\widehat{R}_N(h_j)\right] = R(h_j)$

$\mathbb{P}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| > \epsilon\right)$ 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: BASICS

Lemma (Markov's inequality)

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

$$\mathbb{P}\left(X \geq t\right) \leq \frac{\mathbb{E}\left[X\right]}{t}.$$

Lemma (Chebyshev's inequality)

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

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

Proposition (Weak law of large numbers)

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

$$\forall \epsilon > 0 \quad \mathbb{P}_{\{(\mathbf{x}_i, y_i)\}}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| \geq \epsilon\right) \leq \frac{\operatorname{Var}\left(\mathbf{1}\{h_j(\mathbf{x}_1) \neq y_1\}\right)}{N\epsilon^2} \leq \frac{1}{N\epsilon^2}$$

$$\hookrightarrow \frac{1}{N}\sum_{i=1}^{N} \underbrace{\mathbb{1}\{h_j(x_i) \neq y_i\}}_{\text{iid RV}}$$

BACK TO LEARNING

By the law of large number, we know that

$$\forall\epsilon > 0 \quad \mathbb{P}_{\{(\mathbf{x}_i,y_i)\}}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| \geq \epsilon\right) \leq \frac{\text{Var}\left(\mathbf{1}\{h_j(\mathbf{x}_1) \neq y_1\}\right)}{N\epsilon^2} \leq \frac{1}{N\epsilon^2}$$

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

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

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

Proof: Assume that $\forall h_j \in \mathcal{H} = \{h_i\}_{i=1}^{M}$ $\quad |\hat{R}_N(h_j) - R(h_j)| \leq \epsilon$ for some $\epsilon > 0$ (A)

Then $|\hat{R}_N(h^*) - R(h^*)| \leq \epsilon$

Hence $\mathbb{P}(\forall j \in [\![1;M]\!] \ |\hat{R}_N(h_j) - R(h_j)| \leq \epsilon) \leq \mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| \leq \epsilon)$ (B)

$$1 - \mathbb{P}(\exists j \in [\![1;M]\!] \text{ st } |\hat{R}_N(h_j) - R(h_j)| > \epsilon) \leq 1 - \mathbb{P}(|\hat{R}_N(h^*) - R(h^*)| > \epsilon)$$

A B

hence the result. (*)

BACK TO LEARNING

By the law of large number, we know that

$$\forall \epsilon > 0 \quad \mathbb{P}_{\{(\mathbf{x}_i, y_i)\}}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| \geq \epsilon\right) \leq \frac{\mathrm{Var}\left(\mathbf{1}\{h_j(\mathbf{x}_1) \neq y_1\}\right)}{N\epsilon^2} \leq \frac{1}{N\epsilon^2}$$

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

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

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

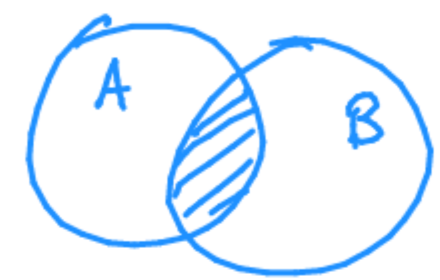
$$\mathbb{P}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geq \epsilon\right) \leq \frac{M}{N\epsilon^2}$$

$$\mathbb{P}\left(\exists j \text{ st } |\hat{R}_N(h_j) - R(h_j)| > \epsilon\right) \leq \sum_{j=1}^{M} \mathbb{P}\left(|\hat{R}_N(h_j) - R(h_j)| > \epsilon\right) \leq \frac{M}{N\epsilon^2}$$

$$\leq \frac{1}{N\epsilon^2}$$

$\mathbb{P}\big(|\hat{R}_N(h_1) - R(h_1)| > \epsilon$ OR
$|\hat{R}_N(h_2) - R(h_2)| > \epsilon$ OR $\dots\big)$

Union bound $\left(\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)\right)$



$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \underbrace{\mathbb{P}(A \cap B)}_{\geq 0}$$

BACK TO LEARNING

By the law of large number, we know that

$$\forall \epsilon > 0 \quad \mathbb{P}_{\{(\mathbf{x}_i, y_i)\}}\left(\left|\widehat{R}_N(h_j) - R(h_j)\right| \geq \epsilon\right) \leq \frac{\mathrm{Var}\left(\mathbf{1}\{h_j(\mathbf{x}_1) \neq y_1\}\right)}{N\epsilon^2} \leq \frac{1}{N\epsilon^2}$$

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

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

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

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

If we choose $N \geq \lceil \frac{M}{\delta\epsilon^2} \rceil$ we can ensure $\mathbb{P}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geq \epsilon\right) \leq \delta$.

- That's a lot of samples!

CONCENTRATION INEQUALITIES: NOT SO BASIC

We can obtain *much* better bounds than with Chebyshev

Lemma (Hoeffding's inequality)

e.g. $[-1/2 ; 1/2]$ then $b_i - a_i = 1$

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

$$\mathbb{P}\left(\left|\underbrace{\frac{1}{N}\sum_{i=1}^N X_i}_{\text{converges in prob. to } 0}\right| \geq \epsilon\right) \leq \underbrace{2\exp\left(-\frac{2N^2\epsilon^2}{\sum_{i=1}^N (b_i - a_i)^2}\right)}.$$

$$= 2\exp\left(-\frac{2N^2\epsilon^2}{N}\right)$$

compare to $\frac{1}{N\epsilon^2}$

CONCENTRATION INEQUALITIES: NOT SO BASIC

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 $X_i \in [a_i; b_i]$ with $a_i < b_i$. 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

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

$$\forall \epsilon > 0 \quad \underbrace{\mathbb{P}\left(\left|\widehat{R}_N(h^*) - R(h^*)\right| \geq \epsilon\right)}_{\leq \delta} \leq 2M\exp(-2N\epsilon^2) \qquad \left(\text{compare to } \frac{M}{N\epsilon^2}\right)$$

CONCENTRATION INEQUALITIES: NOT SO BASIC

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 $X_i \in [a_i; b_i]$ with $a_i < b_i$. Then for all $\epsilon > 0$
>
> $$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N X_i\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2N^2\epsilon^2}{\sum_{i=1}^N (b_i - a_i)^2}\right).$$

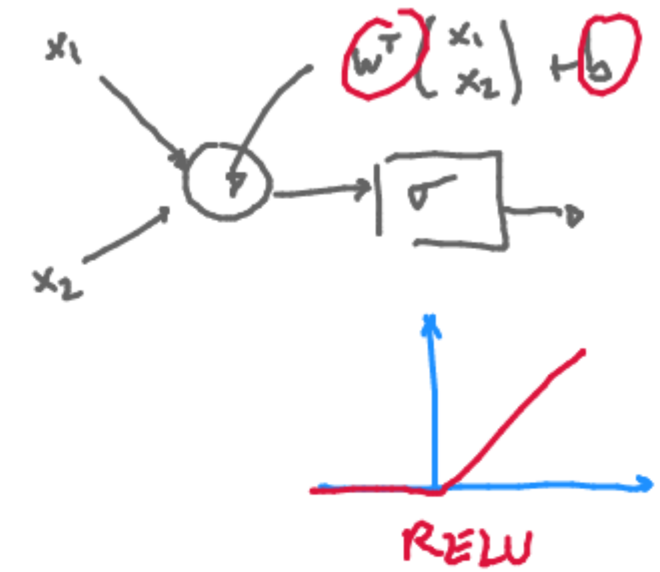
In our learning problem

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

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

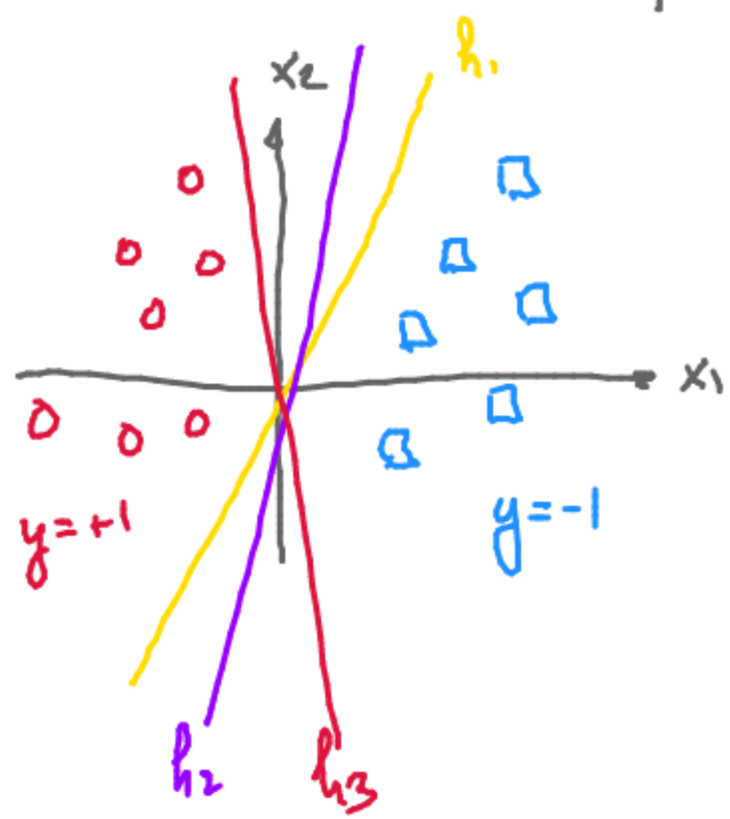
We can now choose $N \geq \lceil \frac{1}{2\epsilon^2}\left(\ln \frac{2M}{\delta}\right)\rceil$

Note: what about infinite classes of models (e.g. neural network)



$w, b \in \mathbb{R}$

It is possible to extend the results to infinite classes



h_1 has empirical risk $\hat{R}_N(h_1) = 0$

h_2 ———— $\hat{R}_N(h_2) = 0$

h_3 ———— $\hat{R}_N(h_3) = 0$

CONCENTRATION INEQUALITIES: NOT SO BASIC

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 $X_i \in [a_i; b_i]$ with $a_i < b_i$. 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 can now choose $N \geq \lceil \frac{1}{2\epsilon^2}\left(\ln \frac{2M}{\delta}\right)\rceil$

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

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

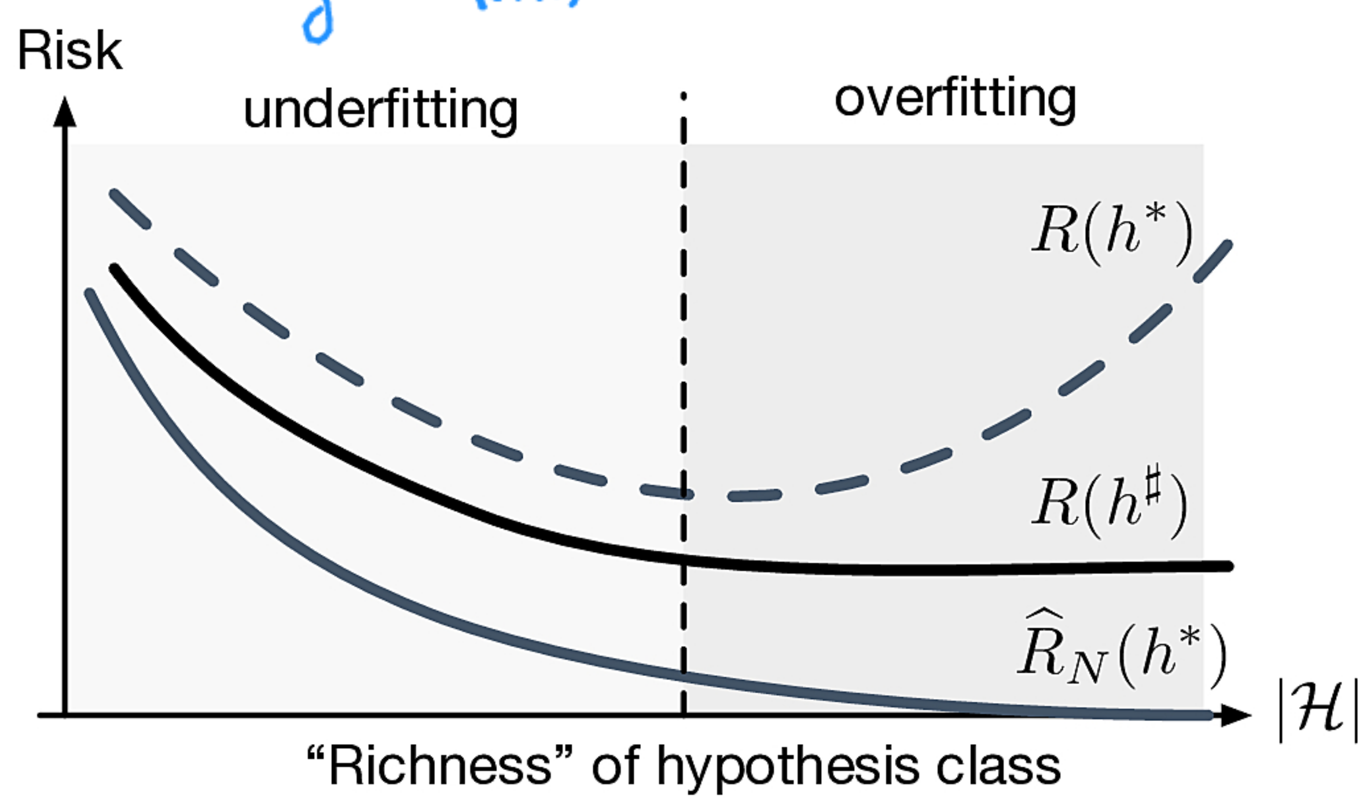
LEARNING CAN WORK!

Lemma.

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

$\operatorname{argmin} \widehat{R}_N(h)$ $\qquad$ $\operatorname{argmin}_h R(h)$

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



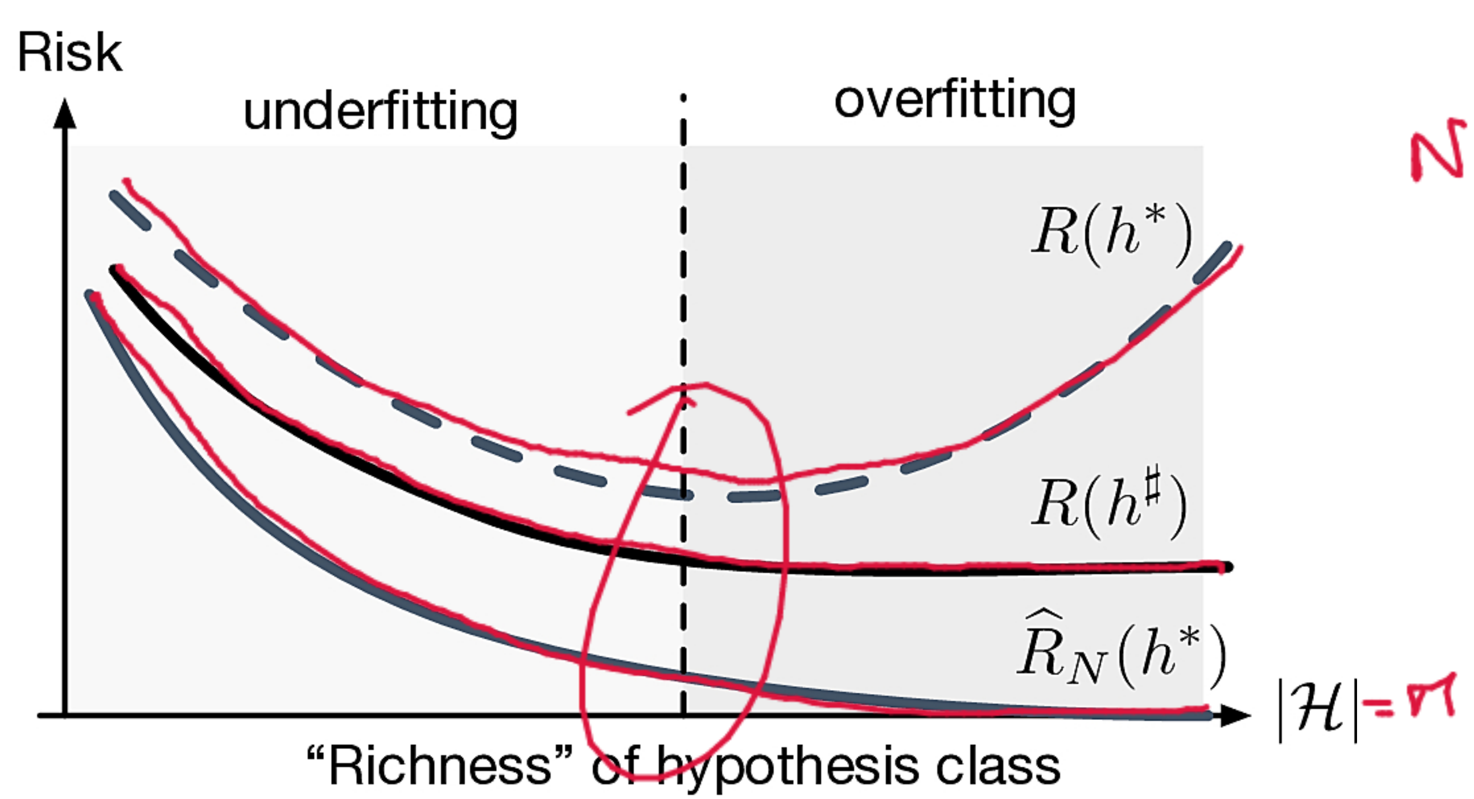
LEARNING CAN WORK!

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



N fixed

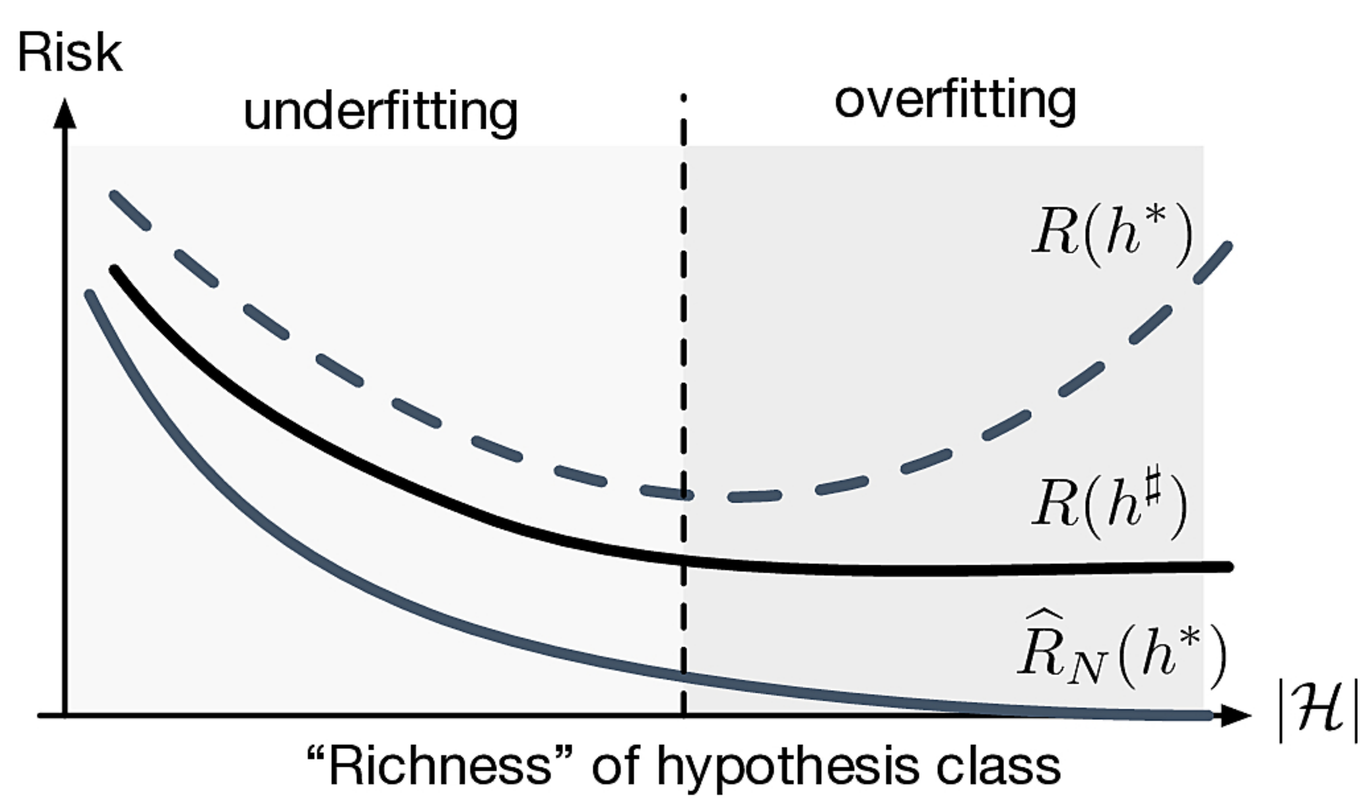
LEARNING CAN WORK!

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



PROBABLY APPROXIMATELY CORRECT LEARNABILITY

Definition. (PAC learnability)
A hypothesis set $\mathcal{H}$ is (agnostic) PAC learnable if there exists a function $N_{\mathcal{H}} :]0;1[^2 \to \mathbb{N}$ and a learning algorithm such that:

- for very $\epsilon, \delta \in]0;1[$,
- for every $P_{\mathbf{x}}, P_{y|\mathbf{x}}$,
- when running the algorithm on at least $N_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples, the algorithm returns a hypothesis $h \in \mathcal{H}$ such that

$$\mathbb{P}_{\mathbf{x}y}\left(\left|R(h) - R(h^{\sharp})\right| \leq \epsilon\right) \geq 1 - \delta$$

The function $N_{\mathcal{H}}(\epsilon, \delta)$ is called *sample complexity*

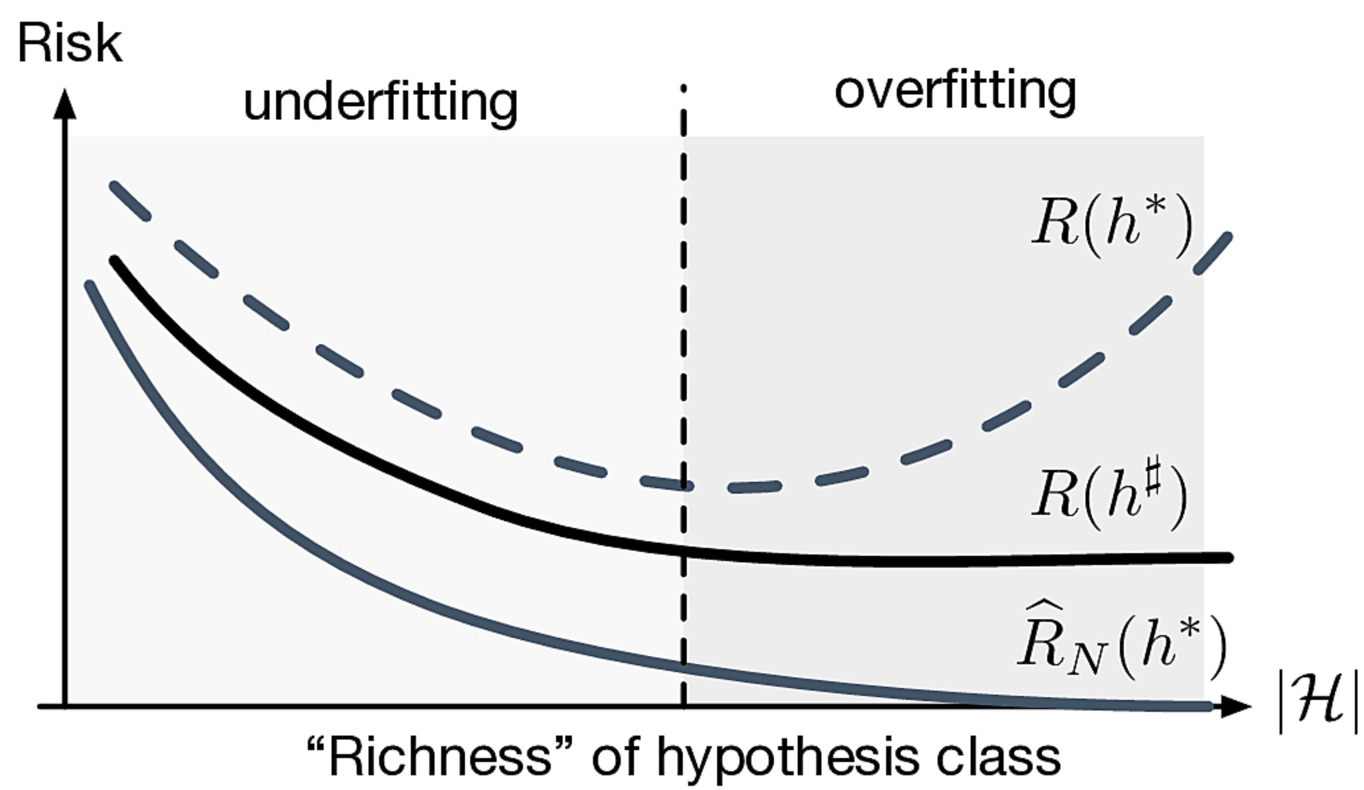
We have effectively already proved the following result

Proposition.

A finite hypothesis set $\mathcal{H}$ is PAC learnable with the Empirical Risk Minimization algorithm and with sample complexity

$$N_{\mathcal{H}}(\epsilon, \delta) = \lceil \frac{2\ln(2|\mathcal{H}|/\delta)}{\epsilon^2} \rceil$$

WHAT IS A GOOD HYPOTHESIS SET?



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

Questions

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

SUPERVISED LEARNING MODEL

We revisit the supervised learning setup (*slight* change in notation)

1. Dataset $\mathcal{D} \triangleq \{(X_1, Y_1), \cdots, (X_N, Y_N)\}$
   - $\{X_i\}_{i=1}^N$ *drawn i.i.d. from unknown* P_X on $\mathcal{X} = \mathbb{R}^d$
   - $\{Y_i\}_{i=1}^N$ labels with $\mathcal{Y} = \{0, 1, \cdots, K-1\}$ (multiclass classification)
2. Unknown $P_{Y|X}$
3. Binary loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+ : (y_1, y_2) \mapsto \mathbf{1}\{y_1 \neq y_2\}$

The risk of a classifier h is

$$R(h) \triangleq \mathbb{E}_{XY}\left[\mathbf{1}\{h(X) \neq Y\}\right] = \mathbb{P}_{XY}\left(h(X) \neq Y\right)$$

We will not directly worry about $\mathcal{H}$, but rather about $R(\hat{h}_N)$ for some $\hat{h}_N$ that we will estimate from the data

BAYES CLASSIFIER

What is the *best* risk (smallest) that we can achieve?

- Assume that we actually know P_X and $P_{Y|X}$
- Denote the *a posteriori* class probabilities of $\mathbf{x} \in \mathcal{X}$ by

$$\eta_k(\mathbf{x}) \triangleq \mathbb{P}\left(Y = k | X = \mathbf{x}\right)$$

- Denote the *a priori* class probabilities by

$$\pi_k \triangleq \mathbb{P}\left(Y = k\right)$$

Lemma (Bayes classifier)

The classifier $h^{\mathrm{B}}(\mathbf{x}) \triangleq \mathbf{argmax}_{k\in[0;K-1]}\eta_k(\mathbf{x})$ is optimal, i.e., for *any* classifier h, we have $R(h^{\mathrm{B}}) \leq R(h)$.

$$R(h^{\mathrm{B}}) = \mathbb{E}_X\left[1 - \max_k \eta_k(X)\right]$$

Terminology

- h^B is called the *Bayes classifier*
- $R_B \triangleq R(h^B)$ is called the *Bayes risk*

OTHER FORMS OF THE BAYES CLASSIFIER

$h^{\mathrm{B}}(\mathbf{x}) \triangleq \operatorname{argmax}_{k\in[0;K-1]}\eta_k(\mathbf{x})$

$h^{\mathrm{B}}(\mathbf{x}) \triangleq \operatorname{argmax}_{k\in[0;K-1]}\pi_k p_{X|Y}(\mathbf{x}|k)$

For $K = 2$ (binary classification): log-likelihood ratio test

$$\log \frac{p_{X|Y}(\mathbf{x}|1)}{p_{X|Y}(\mathbf{x}|0)} \gtrless \log \frac{\pi_0}{\pi_1}$$

If all classes are equally likely $\pi_0 = \pi_1 = \cdots = \pi_{K-1}$

$$h^{\mathrm{B}}(\mathbf{x}) \triangleq \operatorname*{argmax}_{k\in[0;K-1]} p_{X|Y}(\mathbf{x}|k)$$

> **Example (Bayes classifier)**
>
> Assume $X|Y = 0 \sim \mathcal{N}(0,1)$ and $X|Y = 1 \sim \mathcal{N}(1,1)$. The Bayes risk for $\pi_0 = \pi_1$ is $R(h^{\mathrm{B}}) = \Phi(-\frac{1}{2})$ with $\Phi \triangleq \text{Normal CDF}$

In practice we do *not* know P_X and $P_{Y|X}$

- *Plugin methods*: use the *data* to learn the distributions and plug result in Bayes classifier

NEAREST NEIGHBOR CLASSIFIER

Back to our training dataset $\mathcal{D} \triangleq \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$

Definition. ((1) nearest neighbor classifier)

The *nearest-neighbor* (NN) classifier is $h^{\mathrm{NN}}(\mathbf{x}) \triangleq y_{\mathrm{NN}(\mathbf{x})}$ where $\mathrm{NN}(\mathbf{x}) \triangleq \mathbf{argmin}_i \|\mathbf{x}_i - \mathbf{x}\|$

Risk of NN classifier conditioned on $\mathbf{x}$ and $\mathbf{x}_{\mathrm{NN}(\mathbf{x})}$

$$R_{\mathrm{NN}}(\mathbf{x}, \mathbf{x}_{\mathrm{NN}(\mathbf{x})}) = \sum_k \eta_k(\mathbf{x}_{\mathrm{NN}(\mathbf{x})})(1 - \eta_k(\mathbf{x})) = \sum_k \eta_k(\mathbf{x})(1 - \eta_k(\mathbf{x}_{\mathrm{NN}(\mathbf{x})})).$$

- How well does the average risk $R_{\mathrm{NN}} = R(h^{\mathrm{NN}})$ compare to the Bayes risk for large N?

Lemma.

Let $\mathbf{x}$, $\{\mathbf{x}_i\}_{i=1}^N$ be i.i.d. $\sim P_{\mathbf{x}}$ in a separable metric space $\mathcal{X}$. Let $\mathbf{x}_{\mathrm{NN}(\mathbf{x})}$ be the nearest neighbor of $\mathbf{x}$. Then $\mathbf{x}_{\mathrm{NN}(\mathbf{x})} \to \mathbf{x}$ with probability one as $N \to \infty$

Theorem (Binary NN classifier)

Let $\mathcal{X}$ be a separable metric space. Let $p(\mathbf{x}|y=0), p(\mathbf{x}|y=1)$ be such that, with probability one, $\mathbf{x}$ is either a continuity point of $p(\mathbf{x}|y=0)$ and $p(\mathbf{x}|y=1)$ or a point of non-zero probability measure. As $N \to \infty$,

$$R(h^{\mathrm{B}}) \leq R(h^{\mathrm{NN}}) \leq 2R(h^{\mathrm{B}})(1 - R(h^{\mathrm{B}}))$$

K NEAREST NEIGHBORS CLASSIFIER

Can drive the risk of the NN classifier to the Bayes risk by *increasing* the size of the neighborhood

- Assign label to $\mathbf{x}$ by taking majority vote among K nearest neighbors $h^{K\text{-NN}}$

$$\lim_{N\to\infty} \mathbb{E}\left[R(h^{K\text{-NN}})\right] \leq \left(1 + \sqrt{\frac{2}{K}}\right) R(h^{\mathrm{B}})$$

Definition.

Let $\hat{h}_N$ be a classifier learned from N data points; $\hat{h}_N$ is *consistent* if $\mathbb{E}\left[R(\hat{h}_N)\right] \to R_B$ as $N \to \infty$.

Theorem (Stone's Theorem)

If $N \to \infty$, $K \to \infty$, $K/N \to 0$, then $h^{K\text{-NN}}$ is consistent

Choosing K is a problem of *model selection*

- Do *not* choose K by minimizing the empirical risk on training:

$$\widehat{R}_N(h^{1\text{-NN}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{h_1(\mathbf{x}_i) = y_i\} = 0$$

- Need to rely on estimates from model selection techniques (more later!)