Seminar Note 328

Setting: Assume that we are now in a small island and, we want to find some papaya to eat. What we have now is some features of papaya, i.e. color, softness, and we want to predict whether a papaya is tasty or not.

Input: domain set X, here papayas.
Meanwhile, we can assign some features, i.e. color / softness.

Output: What we want to know. Here we call it label set, denoted by Y. For papayas, $Y = \{0, 1\}$.

Training set: finite elements in $X \times Y$, denoted by S.
What the learner has access to.

Our learner will output some prediction rules: $h: X \to Y$
This function h / learner $\triangleq$ predictor, classifier, hypothesis
($A(S)$: hyphothesis that a learner produces after receiving S).

Now, we want to make some assumptions.

① All instances follow some probability distribution D.
($X \times Y \sim D$ or $X \sim D$). Here, $X \sim D$.
② Correct labeling: $f: X \to Y$.
(Sometimes obviously not, i.e. same book, different price).
(Then $P(y|x)$ ...)

Up to now, assume that we have a training set $S = \{(x_i, y_i)\}_{i=1}^{m}$
And a hyphothesis h. How to measure its error?

Generalization error : $L_{D,f}(h) = \underset{x\sim D}{E}(L(h;x,y))$

$= \underset{x\sim D}{E}(L(h;x,f(x)))$,

where L denotes the loss functions.

(Here, since we have : $f: X \to Y$. $x \sim D$, you can see that: Y's probability distribution is already decided by X's distribution D.)

More generally, $L_{D,f}(h) = \underset{x\times y\sim D}{E}(L(h;x,y))$

For this chapter, we use $L(h;x,y) = 1_{h(x)\neq y}$

$\Rightarrow L_{D,f}(h) = \underset{x\sim D}{P}(\{h(x) \neq f(x)\})$.

Remark : Here, assume $h(x), f(x)$ is measurable w.r.t the corresponding probability measure.

However, in real world, we can't know what the generalization error exactly be. So a realistic error must be designed for practical use.

Training error w.r.t the training set S: $L_S(h) \triangleq \frac{|\{x\in S | h(x)\neq f(x)\}|}{|S|}$

(More general: $\frac{\sum_{x\in S} L(h;x,y)}{|S|}$).

(Empirical error / risk)

(Recall : Strong law of large number, $\lim_{|S|\to+\infty} L_S(h) = L_{D,f}(h)$ a.s. if there exist fin'te mean and variance)

Now, we can begin our journey.

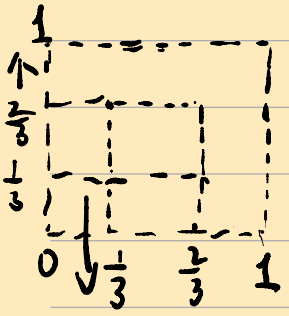
First problem: how to choose the most suitable hypothesis?

Naturally, choose h s.t $L_S(h)$ is minimized.

We call it "Empirical Risk Minimization" paradigm. → Is it enough?

An interesting instance for overfitting

For papayas. assume that we already have make our color and softness numerical variables which take values in $[0,1]$.



Assumption: ① (Color, Softness) ~ Uniform ($[0,1]^2$)

② $\in [\frac{1}{3}, \frac{2}{3}]^2$, tasty, i.e. $y=1$; otherwise not tasty.

③ Our predictor is:

$$h_S(x) = \begin{cases} y_i, & x_i \in S \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, we have: $L_S(h)=0$.

What about $L_{D,f}(h)$? $L_{D,f}(h) = P_{x\sim D}(h(x) \neq f(x))$

$$= \frac{1}{9}.$$

(a.e.'s change doesn't change the probability, $h=0$ a.e.)

Performance on { training data → Good. / True world → May poor } ⇒ Overfitting.

Of course, you may argue that:

In real world, our model / algorithm will not learn such a strange $h_S(x)$.

Notice that: $h_S(x)$ can be interpreted by following:

Consider $P_S(x) = -\prod_{\substack{x_s \in S \\ f(x_s)=1}} (x - x_s)^2$ ($x_s, x \in \mathbb{R}^d$, $d=2$).

Then $P_S(x) \begin{cases} < 0, & \text{if } x \notin S \text{ or } f(x) \neq 1 \\ = 0, & \text{if } x \in S \text{ with } f(x)=1. \end{cases}$

Consider Logistic: $\frac{1}{1+e^{-P_S(x)}} = h_S(x) = \begin{cases} \geq \frac{1}{2}, & \text{if } x \in S \text{ with } f(x)=1 \\ < \frac{1}{2}, & \text{if } x \notin S \text{ or } f(x) \neq 1 \end{cases}$

Aha, seems our training process is going to fail. Is there any methods to rectify the ERM paradigm?

A natural way is to exclude some "bad" hypothesis, or use ERM in the scope of good hyphothesis.

More precisely, we restrict the search space for our hyphothesis, denoted by $\mathcal{H}$. And we use ERM in $\mathcal{H}$, i.e.

$$ERM_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h).$$

So we bias our choice toward a particular set of hyphothesises. (Inductive bias). And $\mathcal{H}$ is usually decided by our previous knowledge about the related problem.

(In interpolation, choose degree $\leq m$...
In this problem, choose $\mathcal{H}$: set of hyphothesises determined by axis aligned rectangles)

And such $\mathcal{H}$ is important obviously, consider what will happen if $\mathcal{H} = \{h: X \to Y \mid |\{h(x)=1\}| < \infty\}$

(You believe only finite papayas are tasty!)

Now, we propose an important conclusion.

Theorem 1: If $|\mathcal{H}| < \infty$, then $ERM_{\mathcal{H}}(S)$ will not overfit.

(Recall: $(-1)^{s} (1+f) 2^{e}$ — 1 digit; 11 digit, $2^{10} \sim 2^{0}$; 52 digit, $\frac{1}{2} \sim (\frac{1}{2})^{52}$; 64 digit.)

In programming, $\mathcal{H}$ = {all hyphothesis determined by axis aligned rectangles} is finite).

Proof: Assumption 1: $\exists h^* \in \mathcal{H}$ s.t $L_{(D,f)}(h^*) = 0$.
Assumption 2: All elements of S are i.i.d. according to the distribution D. So $S \sim D^m$.

Of cource, due to Assumption 1, $P_{x \sim D}(h^*(x) = f(x)) = 1$. So for any set S satisfying assumption 2, we have:

$$P(L_S(h^*) = 0) = P(h^*(x) = f(x), \forall x \in S)$$
$$= \prod_{x \in S} P(h^*(x) = f(x)) = 1.$$

So $L_S(h^*)=0$ a.s. And we omit a.s. in our following discussion.

Let $h_S \in \underset{\mathcal{H}}{\operatorname{argmin}} L_S(h)$, we have: $P(L_S(h_S)=0)=1$.

So with probability 1 over random samples S, $L_S(h_S)=0$.

Notice that $L_{(D,f)}(h_S)$ is a r.v. due to the randomness of S. And not all the possible S will lead to a good hyphothesis by ERM in $\mathcal{H}$.

And we denote the probability of getting such non-representative sample by δ with $1-\delta$ as the confidence parameter.

(Consider in overfitting example. you choose S in $[0,1]^2$ / $[\frac{1}{3},\frac{2}{3}]^2$ / just $[\frac{1}{3},\frac{2}{3}]^2$).

Meanwhile, giving a good example, how to evaluate the quality of h_S ? Choose a parameter ε- accuracy. We say h_S is { success, if $L_{D,f}(h_S) \le \varepsilon$
{ failure, if $L_{D,f}(h_S) > \varepsilon$

When $|S|=m$, we are interested in:

$$\underset{S\sim D^m}{P}(S \mid L_{D,f}(h_S) > \varepsilon) \triangleq \mathbb{P}$$

Let $\mathcal{H}_B = \{h \in \mathcal{H} \mid L_{D,f}(h) > \varepsilon\}$, $\mathbb{P} = \underset{S\sim D^m}{P}(S \mid h_S \in \mathcal{H}_B)$

$$\le \underset{S\sim D^m}{P}\left(\bigcup_{h\in\mathcal{H}_B}(S \mid L_S(h)=0)\right)$$

Recall $L_{D,f}(h) \triangleq P(h(x) \neq f(x))$

(Why $\leq$, notice that: $\operatorname{argmin}_{h} L_S(h)$ may have more than 1 choice).

Hence, $\mathbb{P} \leq \sum_{h \in \mathcal{H}_B} P(S \mid L_S(h)=0)$

$$\leq \sum_{h \in \mathcal{H}_B} (1-\varepsilon)^m \leq |\mathcal{H}_B| e^{-m\varepsilon} \leq |\mathcal{H}| e^{-m\varepsilon}$$

So when $|\mathcal{H}| e^{-\varepsilon m} \leq \delta$, i.e. $m \geq \frac{\log(\frac{|\mathcal{H}|}{\delta})}{\varepsilon}$, we can have:

for <u>any labeling function and distribution D satisfying</u>

assumption 1 and 2. With probability of at least $1-\delta$ over the choice of an i.i.d sample S of size m, we have:

$$L_{D,f}(ERM_{\mathcal{H}}(S)) \leq \varepsilon.$$

(That's probably ($1-\delta$) approximately ($\leq \varepsilon$) correct !
(Probably Approximately Correct learning)!

Why we care about the generalized error - a thing that can't be computed practically?

Though we have only finite instance, what we want to know is infinite. That's - we care about: What's our algorithm's performance encountering unseen data $\sim D$. And the general data has already included the loss from unseen data, in fact, all possible data.

And in reality, even if such correct f exists, (f, D) is unknown for us. PAC is a theoretical result regardless of f, D.

|| What's the ... for $P \sim$

Ch 3.

Just like we discussed before, we define:

Def for PAC learnablity:

For a hyphothesis class $\mathcal{H}$, we call it's PAC learnable if there exist $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ with a learning algorithm, s.t: For $\forall$ distribution over X, and every labeling function $f: X \to \{0,1\}$ with Assumption 1 holds w.r.t $\mathcal{H}, D, f$, the algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d samples w.r.t (D, f) returns a hyphothesis h s.t with probability of at least $1-\delta$ (over the choices of samples), $L_{D,f}(h) \leq \varepsilon$.

Remark: We define $m_{\mathcal{H}}(\varepsilon, \delta)$ to be the minimal integer s.t PAC learnability holds among all possible choices.

Remark: From the i.i.d assumption, you can see that:

① There are always some bad samples which leads to a bad hyphothesis, no matter which learning algorithm you choose (For example, S has only one element), and no matter the size of samples you choose.
(That's why we define δ, though we can't prove such bad samples happening with nonzero probability).

② S is always finite, so we can't expect the full understanding of the real distribution.
(That's why we define ε).

By remark, the function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ determines the sample complexity of learning $\mathcal{H}$.

Restate of Ch2's main theorem: Every finite hypothesis class is PAC learnable with sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(\frac{|\mathcal{H}|}{\delta})}{\varepsilon} \right\rceil$

So far, we have propose too much assumptions, which is too strict in real world. And now we want to remove or relax them:

Giving features x, it's possible that: there are different "labels" y. (e.g. give your all possible information, we can't predict whether you will do some specific thing exactly! Otherwise, you can go to tell anyone's future ...).

However, it's possible to give a conditional probability for "label" y giving information x, i.e. $P(y|x)$.

And now, let $X \times Y \sim D$, we define:

$$\begin{cases} L_D(h) = \underset{(x,y)\sim D}{P}(h(x) \neq y) & \text{as generalization error} \\ L_S(h) = \frac{|\{x \in S \mid h(x) \neq y\}|}{|S|} & \text{as empirical error / risk.} \end{cases}$$

To relax / remove assumption 1: we first want to see: What's the possible optimal performance for a hypothesis?

And to better understand the performance, it's necessary to introduce the generalized loss functions when we predict.

Consider: given a hypothesis class $\mathcal{H}$ and some domain $Z = X \times Y$, we say $l : \mathcal{H} \times Z \to \mathbb{R}$ is a loss function. iff

(A) $l \geq 0$ and $l = 0 \Leftrightarrow h(x) = y$ for $h \in \mathcal{H}$, $(x, y) \in X \times Y$

(Different from the book).

So, we have for prediction problems (some semi-supervised / supervised / graph ...)

The general error is always defined by $\underset{X \times Y \sim D}{E}(l(h, (x, y)))$

The empirical error is always defined by $\frac{\sum_{(x,y) \in S} l(h, (x, y))}{|S|}$

(The original def is just $l(h, (x, y)) = \mathbb{1}_{h(x) \neq y}$, i.e. 0-1 loss.)

(For more forms of loss, refer to the book...).

(Here H' is all hypothesises).

It's natural to think: $f_D(x) = \underset{h(x), h \in H'}{\operatorname{argmin}} \int_Y l(h, (x, y)) f_{y|x}(y|x)\, dy$ is optimal. ✓

And for any hypothesis: $h : X \to Y$, we have:

$$L_D(h) = \underset{X \times Y \sim D}{E}(l(h, (x, y)) = \iint_{X \times Y \sim D} l(h, (x, y)) f(x, y)\, dx\, dy \quad \text{(or } dA\text{)}$$

$$= \int_X \int_Y l(h, (x, y)) f(x, y)\, dy\, dx \quad \text{(Fubini)} \quad (f \text{ is the joint p.d.f.})$$

$$= \int_X f_X(x) \int_Y l(h,(x,y)) \frac{f(x,y)}{f_X(x)} dy\, dx \quad (f_X(x),\ \text{p.d.f of } X)$$

$$= \int_X f_X(x) \int_Y l(h,(x,y))\, f_{y|x}(y|x)\, dy\, dx \geq L_D(f_D) \text{ by definition}$$

(Assume: Support $(f(x,y)) =$ Support $(f_{x|x})$ × Support $(f_{Y|y})$, i.e. Support $(f(x,y))$ is a product space).

If $l(h,(x,y)) = 1_{h(x) \neq y}$, then $f_D(x) = \underset{h(x)}{\operatorname{argmin}} \int_Y l(h,|x,y))\, f_{y|x}(y|x)\, dy$

$$= \underset{h(x)}{\operatorname{argmin}} \int_Y 1_{h(x)\neq y}\, f_{y|x}(y|x)\, dy$$

$$= \underset{h(x)}{\operatorname{argmin}}\ P(h|x) \neq y \mid x)$$

$$= \underset{h(x)}{\operatorname{argmin}}\ 1 - P(h(x) = y \mid x)$$

$$= \underset{h(x)}{\operatorname{argmax}}\ P(h(x) = y \mid x)$$

$$= \underset{y}{\operatorname{argmax}}\ P[y|x).$$

Since we don't know about the real distribution, it's hard to define the "global" optimal, i.e. how far is our hyphothesis away from $f_D(x)$. However, it's still possible to define the "local" optimal, i.e. how far is our hyphothesis away from some possible hyphothesis what we choose, i.e. h.

So it's time to define:

(Agnostic PAC learnability for general loss functions).

DEFINITION 3.4 (Agnostic PAC Learnability for General Loss Functions) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set Z and a loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$, if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over Z, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns $h \in \mathcal{H}$ such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon,$$

where $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$.

Ch4

Now, recall what we have:

$L_S(h)$ is the training error w.r.t a training set S

$L_D(h)$ is the generalization error

Since we don't know the distribution D, we can only use $L_S(h)$ based method to improve our algorithm's overall performance directly.

So, we can only understand $L_D(h)$ from the understanding of $L_S(h)$. This motivates us to introduce about some relation of $L_D(h)$ and $L_S(h)$.

Def: A training S is called ε-representative w.r.t domain Z, hypothesis $\mathcal{H}$, loss functions ℓ, distribution D, if

$|L_D(h) - L_S(h)| \leq \varepsilon$ for all $h \in \mathcal{H}$

It's directly that:

If a training set S is $\frac{\varepsilon}{2}$-representative (w.r.t $Z, D, l, \mathcal{H}$), then any output of $ERM_{\mathcal{H}}(S)$ — denoted by h_S, satisfying:

$$L_D(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h) + \frac{\varepsilon}{2} \leq L_D(h) + \varepsilon \text{ for all } h \in \mathcal{H}$$
$$\Rightarrow L_D(h_S) \leq \min_{\mathcal{H}} L_D(h) + \varepsilon$$

So, if with probability of at least $1-\delta$ for all possible choice of training set to be $\frac{\varepsilon}{2}$-representative, we can conclude the ε-δ result. This leads to:

Def: A hyphothesis class $\mathcal{H}$ is called uniform convergence property (w.r.t domain Z and loss function l) if there exists a function $m_{\mathcal{H}}^{UC}: (0,1)^2 \to \mathbb{N}$, s.t for any $\varepsilon, \delta \in (0,1)$ and any distribution D over Z, if $|S| \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ examples drawn i.i.d according D, then $P(S \text{ is } \varepsilon\text{-representative}) \geq 1-\delta$.

(A) It's obviously that: For uniform convergance class $\mathcal{H}$, it's agnostically PAC learnable with sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\varepsilon}{2}, \delta)$

(The corresponding algorithis is ERM).

Now, we propose the main theorem.

(Thm): Any finite classes are agnostic PAC learnable.

Proof: It suffices to prove that any finite class is of uniform convergence.

Lemma: For i.i.d r.v. $X_1 \sim X_m$ with $E(X_i) = \mu$, $P(a \le X_i \le b) = 1$.
Then $\forall \varepsilon > 0$, $\mathbb{P}\left(\left|\frac{\sum_{i=1}^m X_i}{m} - \mu\right| > \varepsilon\right) \le 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$ (Hoeffding ineq).

Lemma's proof: Notice that:

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^m X_i}{m} - \mu\right| \ge \varepsilon\right) = \mathbb{P}\left(\frac{\sum_{i=1}^m X_i}{m} - \mu \ge \varepsilon\right) + \mathbb{P}\left(\frac{\sum_{i=1}^m X_i}{m} - \mu \le -\varepsilon\right)$$

Bound $\mathbb{P}\left(\frac{\sum_{i=1}^m X_i}{m} - \mu \ge \varepsilon\right)$ first.

$$\mathbb{P}\left(\frac{\sum_{i=1}^m X_i}{m} - \mu \ge \varepsilon\right) = \mathbb{P}\left(e^{\lambda \frac{\sum_{i=1}^m (X_i - \mu)}{m}} \ge e^{\lambda\varepsilon}\right)$$

Lemma. lemma: X is a r.v. with $P(a \le X \le b) = 1$, $E(X) = 0$, then
$E(e^{\lambda X}) \le e^{\frac{\lambda^2 (b-a)^2}{8}}$ for $\lambda > 0$.

Pf: $e^{\lambda x}$ is cvx for $\lambda > 0 \Rightarrow e^{\lambda x} \le \frac{b-x}{b-a} e^{\lambda a} + e^{\lambda b} \frac{x-a}{b-a}$.

$$\Rightarrow E(e^{\lambda X}) \le \frac{b}{b-a} e^{\lambda a} - e^{\lambda b} \frac{a}{b-a} = e^{\lambda a}\left(\frac{b}{b-a} + e^{\lambda(b-a)} \frac{-a}{b-a}\right)$$

Since $E(X) = 0$, we have: $a \le 0$, hence, $\frac{-a}{b-a} \le 1$, $\frac{b}{b-a} \le 1$.

Let $h = \lambda(b-a)$.

Let $p = \frac{-a}{b-a} \in (0,1)$, we have: $E(e^{\lambda X}) \le e^{-hp + \log(1-p+pe^h)}$.

Let $F(h) = -hp + \log(1-p+pe^h)$.

$F(0) = 0$, $F'(0) = -p + \left.\frac{pe^h}{1-p+pe^h}\right|_{h=0} = 0$, $F''(h) = \frac{pe^h(1-p)}{(1-p+pe^h)^2} \le \frac{1}{4}$ (AM-GM)

Hence, $\tilde{F}(h) \le \frac{1}{8}h^2$

Hence, $E(e^{\lambda X}) \le e^{\frac{\lambda^2(b-a)^2}{8}}$ ✓

Hence, $\mathbb{P}\left(e^{\frac{\sum_{i=1}^m (X_i - \mu)}{m}} \ge e^{\varepsilon}\right) \le \frac{E\left(e^{\frac{\sum_{i=1}^m (X_i-\mu)}{m}}\right)}{e^{\varepsilon}} \le \frac{\prod_{i=1}^m E\left[e^{\frac{\lambda(X_i-\mu)}{m}}\right]}{e^{\lambda\varepsilon}}$

$$\le e^{-\lambda\varepsilon + \frac{\lambda^2(b-a)^2}{8m}} \le e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$

$\left(= \text{取} \lambda = \frac{4m\varepsilon}{(b-a)^2}\right)$.

Similarly for the other side: $(X_i' = -X_i)$ ✓

Hence: to prove:

$$\mathbb{P}(\{S \mid \forall h \in \mathcal{H}, |L_S(h) - L_D(h)| \le \varepsilon\}) \ge 1 - \delta$$

$S \sim D^m$

$$\Leftarrow \mathbb{P}(\{S \mid \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \varepsilon\}) \le \delta$$

$$LHS \le \sum_{h \in \mathcal{H}} P(\{S \mid |L_S(h) - L_D(h)| > \varepsilon\})$$

Assume $|\ell(h, (x, y))| \le M$ for all $(x, y) \in Z \sim D$ and $h \in \mathcal{H}$.

Hence, $LHS \le |\mathcal{H}| \cdot 2e^{-\frac{2m\varepsilon^2}{M^2}}$

$$\text{A } 2|\mathcal{H}| e^{-\frac{2m\varepsilon^2}{M^2}} \le \delta \Leftrightarrow m \ge \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right) M^2}{2\varepsilon^2}$$ ✓

This leads to:

COROLLARY 4.6 *Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0,1]$ be a loss function. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity*

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil .$$

Furthermore, the class is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil .$$

(We just need to let $\ell: \mathcal{H} \times Z \to [0, M]$).

Rmk: You may say: In reality, "finite" hyphothesis class is almost impossible.

Consider: You have a hyphothesis class that is parametrized by d parameters. (i.e. NNs, OLS, Regression ....., covering many practical things in ML). When we realize this hyphothesis, we just tune these parameters to find out the near-optimal one.

Due to the floating point's property, $(-1)^{S}(1+f)\,2^{E}$ (S: 1 bit, E: 11 bits, f: 52 bits), we only have 2^{64} different choice for each parameter when using computer. Hence $|\mathcal{H}| \leq 2^{64d}$

Using above thm. the sample complexity is bounded by:

$$\frac{2M^2 \log\left(\frac{2^{64d+1}}{\delta}\right)}{\epsilon^2} \leq \frac{2M^2 \log\left(\frac{e^{64d \cdot 2}}{\delta}\right)}{\epsilon^2} \leq \frac{2M^2\left(64d + \log\frac{2}{\delta}\right)}{\epsilon^2}$$