Lecture 2: PAC Learnability

CSE 427: Machine Learning

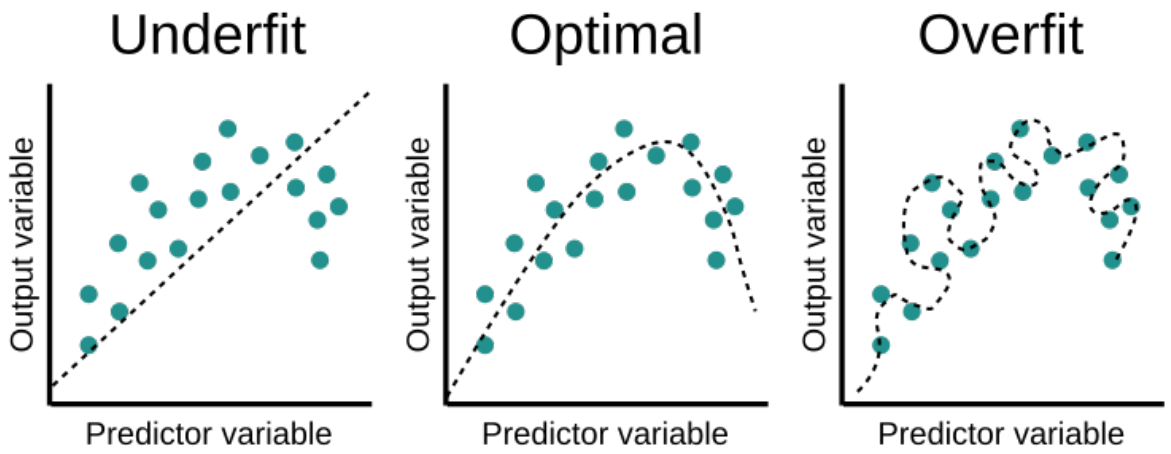
Md Zahidul Hasan

Lecturer, Computer Science
BRAC University

Spring 2023

The Risk of Empirical Risk Minimization

Since the testing data isn't available to the learner during training, the only tool to measure performance is the empirical error. But the training data could give us a very misleading picture of the whole domain. Thus the learner could be misled trying to wholly adapt to the training samples. The learning paradigm that tries to minimize empirical risk is called the ERM paradigm. In the image below, we can see that the rightmost predictor most accurately describes the training data. But the true distribution of the data might not follow this peculiar curve. So, to be safe, the second predictor is our best option.



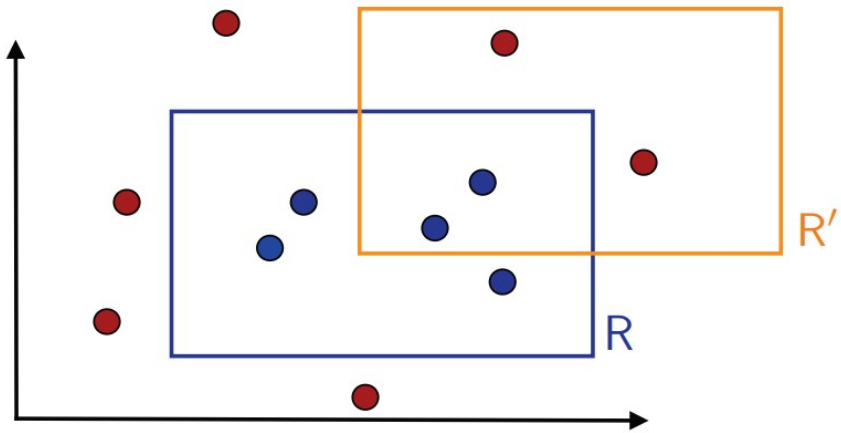
ERM with Inductive Bias

In the previous slide, we saw that the ERM approach could be very misleading. If we put no restriction on the set of possible hypotheses, we could end up with the most bizarre hypothesis that's just perfect for the training data but fails miserably during testing. So, only if we could find a good class of non-bizarre hypotheses called H, we could just choose one that minimizes the empirical error over the training set S. And that hypothesis is called ERM_H.

$$ERM_H(S) \in \arg\min_{h \in H} R_S(h)$$

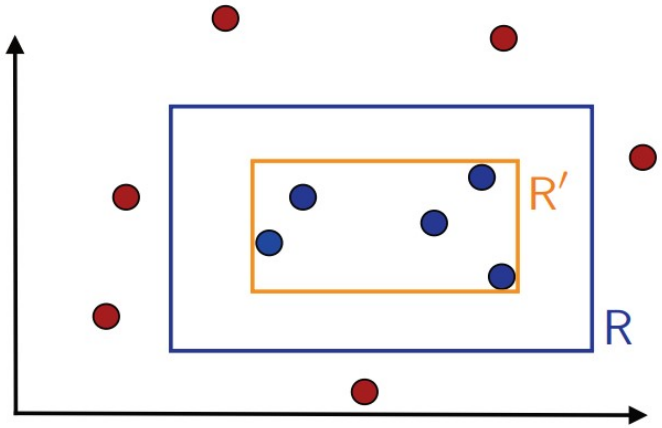
This choice of hypotheses introduces a bias towards a particular set of predictors. This is called inductive bias. ERM with inductive bias is one of the remedies for overfitting. The more restricted a hypothesis class is, the more biased our learner is. The less restricted a hypothesis class is, the more likely it is that we will end up with an overfitting hypothesis. In this course, we will learn how to find a good hypothesis set.

ERM Application: Axix-aligned Rectangles



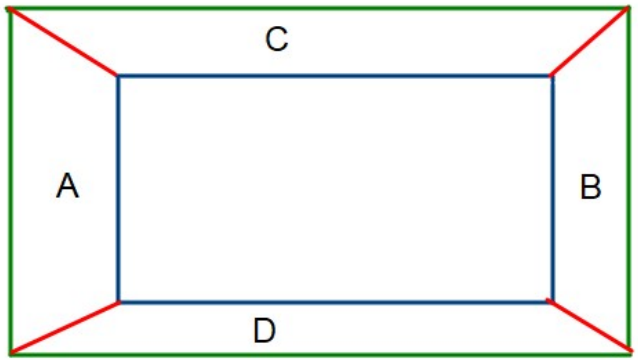
Suppose, a military school is looking for applicants that have their weight and height within a specific range. Let's say an accepted candidate's height is x and their weight is y. So, for some x_1, x_2, y_1, y_2, $x_1 \leq x \leq x_2$ and $y_1 \leq y \leq y_2$. But to an outsider, x_1, x_2, y_1, y_2 are hidden. They can only see which applicants were taken and which were discarded. In the above diagram, the blue rectangle is the true decision boundary. Can we find an algorithm so that given any training sample, it can find a hypothesis that can perform arbitrarily well depending on the sample size with a very high probability?

Axis-aligned Rectangles contd.



An obvious choice of H would be the set of all axis-aligned rectangles. If our learning algorithm chooses the rectangle h with the minimum area that contains all the blue points, then definitely $R_S(h) = 0$. If we test this rectangle with a test set, the only way it can make mistakes is by classifying the points that fall outside $h = R'$ but inside R as red. We want to find an upper bound for the likelihood of this kind of error. We want that $Pr[R(h) > \epsilon] \leq \delta$ for arbitrarily small but positive ϵ and δ.

Axis-aligned Rectangles contd.



Let's connect the corners of the outer and inner rectangles with red lines. Let's say, the probability that a point falls into region A is a. Similarly define b, c, d. So, the total probability of making a mistake is $a + b + c + d$.

$Pr[R(h) > \epsilon] = Pr[a + b + c + d > \epsilon]$

$\leq Pr[a \geq \frac{\epsilon}{4} \vee b \geq \frac{\epsilon}{4} \vee c \geq \frac{\epsilon}{4} \vee d \geq \frac{\epsilon}{4}]$

$\leq Pr[a \geq \frac{\epsilon}{4}] + Pr[b \geq \frac{\epsilon}{4}] + Pr[c \geq \frac{\epsilon}{4}] + Pr[d \geq \frac{\epsilon}{4}]$ (union bound)

$\leq 4Pr[a \geq \frac{\epsilon}{4}] \leq 4(1 - \frac{\epsilon}{4})^m \leq 4e^{\frac{-m\epsilon}{4}} = \delta$ ($e^x \geq 1 + x, \quad \forall x \in \mathbb{R}$)

Solving for m gives, $m \geq \frac{4}{\epsilon} \log \frac{4}{\delta}$. So, if we want 99% accuracy with 99% confidence, then we only need $m = \frac{4}{0.01} \log \frac{4}{0.01} = 1040$ points.

ERM with Finite Hypothesis Classes

The most obvious way to restrict a hypothesis class is to put restrictions on its size. For now, we focus on hypothesis classes that are finite in size.

Consistency Assumption

A hypothesis class H is said to be consistent if there exists a hypothesis h so that $R(h) = 0$. This is also called the realizability assumption. Obviously $R_S(h) = 0$ for any sample S.

I.I.D. Assumption

We will also assume that the instances of the dataset are independent of each other and have an identical probability distribution.

Learning Guarantees for Consistent and Finite Hypothesis Classes

Learning Guarantee: Finite and Consistent Scenario

Let H be a finite and consistent hypothesis set that contains the target concept c. Let A be an algorithm such that for any sample S, it returns a hypothesis $h \in H$ so that $R_S(h) = 0$. Then we can say that for any positive $0 < \epsilon, \delta < 1$, there exists a minimal sample size $m \geq \frac{1}{\epsilon}(\log |H| + \log \frac{1}{\delta})$ so that if we take any sample of size m, then the algorithm will return such a hypothesis h so that $Pr[R(h) \leq \epsilon]$ can hold with at least $1 - \delta$ probability.

Proof: $Pr[\exists h \in H : R_S(h) = 0 \wedge R(h) > \epsilon]$
$\leq \sum_{h \in H} Pr[R_S(h) = 0 \wedge R(h) > \epsilon]$ (Union bound)
$\leq \sum_{h \in H} Pr[R_S(h) = 0 | R(h) > \epsilon]$ $(P(A \wedge B) \leq P(A|B))$
$\leq |H|(1 - \epsilon)^m \leq |H| e^{-\epsilon m} = \delta.$ $(e^x \geq 1 + x, \quad \forall x \in \mathbb{R})$
Solving for m yields the desired bound.

When is a problem PAC-learnable?

The PAC Learning Framework

A concept class C is said to be PAC-learnable if there exists an algorithm $\mathcal{A}$ that can return a hypothesis h for any concept $c \in C$, any sample S, any $\epsilon > 0$ and any $\delta > 0$, so that if we make the sample size sufficiently large ($m \geq poly(\frac{1}{\epsilon}, \frac{1}{\delta}, n, |C|)$), then the following holds:

$$Pr[R(h) \leq \epsilon] \geq 1 - \delta$$

where m is the sample size and n is the number of features. *poly* is a polynomial function. If $\mathcal{A}$ runs in $\mathcal{O}(poly(\frac{1}{\epsilon}, \frac{1}{\delta}, n, |C|))$ run-time complexity, then C is said to be efficiently PAC-learnable.

In the last lecture, we saw that the problem of learning axis-aligned rectangles is PAC-learnable. The concept class and the hypothesis class in that example were identical. PAC stands for Probably $(\geq 1 - \delta)$*Approximately*$(\leq \epsilon)$*Correct*.

Example: Conjunction of Booleans

0	1	1	0	1	1	+
0	1	1	1	1	1	+
0	0	1	1	0	1	-
0	1	1	1	1	1	+
1	0	0	1	1	0	-
0	1	0	0	1	1	+
0	1	?	?	1	1	

In the above picture, the first six columns represent six variables $x_1, x_2, x_3, x_4, x_5, x_6$. The rightmost column contains the label. $+$ means 1 and - means 0. The concept we are trying to learn is the conjunction of these variables or their negations or their absence, for example: $x_1 \wedge \bar{x}_2 \wedge x_4 \wedge x_5 \wedge \bar{x}_6$ or just $\bar{x}_1 \wedge \bar{x}_3$. It's a boolean expression. But we don't know which one it is. One the expressions that satisfy the given dataset is: $\bar{x}_1 \wedge x_2 \wedge x_5 \wedge x_6$. In the test example, we have $x_1 = 0, x_2 = 1$, $x_5 = 1, x_6 = 1$. So, our prediction will be $\bar{0} \wedge 1 \wedge 1 \wedge 1 = 1$. Is this concept class PAC-learnable if it has n boolean variables?

Conjunction of Booleans: Proof of Learnability

Obviously the concept class is large. $|H| = |C| = 3^n$. Because each of the variables can be in 3 forms $\{x, \bar{x}, nothing\}$. Can we find such an algorithm that can always return a consistent hypothesis? Yes. Look at the rows with a positive label. If a variable has a value of 0, then it can't be in the x form. If a variable has a value of 1, then it can't be in the $\bar{x}$ form. So, by checking all the positive rows, we can discard the possible forms of each variable. In the end, if a variable has both x and $\bar{x}$ forms discarded, then the variable is discarded from the expression. Otherwise, it can remain through one of its remaining forms. If it has both forms intact, then it doesn't appear in the expression. Now, is the sample size still valid for PAC-learning? From the last lecture $m = \frac{1}{\epsilon}(\log|H| + \log\frac{1}{\delta}) = \frac{1}{\epsilon}(n\log 3 + \log\frac{1}{\delta}) = poly(\frac{1}{\epsilon}, \frac{1}{\delta}, n)$. So, this concept class is definitely PAC-learnable.

Learning Guarantees for Finite and Inconsistent Hypothesis Classes

In most cases, our hypothesis class may not contain the target concept. It may not even contain a hypothesis that is error-free in the training sample. That isn't too bad. Since it admits some error, we can also be relieved that our hypothesis isn't overfitting. But can we derive any learning guarantees? If the data instances are independent and identically distributed, yes. In that case, the true error being arbitrarily distant from the empirical error is very unlikely. In fact, the following holds:

Learning Guarantee with IID samples

If we are talking about binary classifiers tested with a sample S of size m, then,

$$Pr[|R(h) - R_S(h)| \geq \epsilon] \leq 2e^{-2m\epsilon^2}$$

Learning Guarantee: Proof

Proof.

It's obvious that, $R_S(h)$ is an unbiased estimate of $R(h)$ because $E_{S\sim\mathcal{D}}[R_S(h)] = R(h)$. Since we are dealing with binary classifiers, the error lies in the set $\{0, 1\}$. Let $X_1, X_2, X_3, \cdots, X_m$ be the error of the m data points. Define $S = X_1 + X_2 + \cdots + X_m$. Since the data points are I.I.D., we can assume that the errors are random variables taking values in the range $[0, 1]$. By using *Hoeffding's Inequality*, we can say that,

$$Pr[|S - E[S]| \geq m\epsilon] \leq 2e^{-2m\epsilon^2}$$
$$Pr[|\frac{S}{m} - \frac{E[S]}{m}| \geq \epsilon] \leq 2e^{-2m\epsilon^2}$$
$$Pr[|\frac{S}{m} - E[\frac{S}{m}]| \geq \epsilon] \leq 2e^{-2m\epsilon^2}$$
$$Pr[|R_S(h) - R(h)| \geq \epsilon] \leq 2e^{-2m\epsilon^2}$$

Learning Bounds for Finite Inconsistent Hypothesis Classes

Generalization Bound - Fixed Hypothesis

$$Pr[|R_S(h) - R(h)| \geq \epsilon] \leq 2e^{-2m\epsilon^2} = \delta$$
$$Pr[|R_S(h) - R(h)| \leq \epsilon] \geq 1 - 2e^{-2m\epsilon^2} = 1 - \delta$$

Solving for ϵ yields that the following holds with probability at least $1 - \delta$:

$$R(h) \leq R_S(h) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

But the above bound holds for a fixed hypothesis h. But what about h_S returned by the learning algorithm? Can we say the same about that hypothesis? No. Because when h is fixed, $E_S[R_S(h)] = R(h)$. But $E_S[R_S(h_S)] \neq R(h_S)$ because on the left side, every time we take a different sample S for the expectation, we get a different h_S returned by the algorithm which might be different from the h_S of the right-hand side. So, we need to derive a wider bound that holds for all hypotheses in H.

Learning Bounds for Finite Inconsistent Hypothesis Classes

Generalization Bound for a Whole Hypothesis Class

Let H be a finite hypothesis class of binary classifiers. Then for any $\delta > 0$, with probability at least $1 - \delta$ the following holds:

$$\forall h \in H, \quad R(h) \leq R_S(h) + \sqrt{\frac{\log |H| + \log \frac{2}{\delta}}{2m}}$$

Proof:

$$\begin{gathered}
Pr[\forall h \in H, \quad |R(h) - R_S(h)| \leq \epsilon] \\
= 1 - Pr[\exists h \in H, \quad |R(h) - R_S(h)| \geq \epsilon] \\
= 1 - Pr[(|R(h_1) - R_S(h_1)| \geq \epsilon) \vee \cdots \vee (|R(h_{|H|}) - R_S(h_{|H|})| \geq \epsilon)] \\
\geq 1 - \sum_{h \in H} Pr[|R(h) - R_S(h)| \geq \epsilon] \\
\geq 1 - 2|H|e^{-2m\epsilon^2} = 1 - \delta
\end{gathered}$$

Solving for ϵ yields the aforementioned result.

Extensions to PAC: Bayes Error

In the deterministic scenario, given a feature vector x, there is a unique label $y = f(x)$. So, the hypothesis we are trying to learn is a function f. But, let's say, we are trying to find a correlation between the height and weight of a person with their propensity to have diabetes. Let's say, the height of a person is 171cm and the weight of a person is 65 kg, we want to predict whether they have a risk for diabetes. Now, the dataset may contain the record of 125 people who have the same height and weight combination. And 100 of them didn't have diabetes but 25 did. In this case, we can't find a deterministic function to describe the dataset. The best we can do is we can calculate a distribution $\bar{P}(y = NO|x) = \frac{100}{125}$ and $\bar{P}(y = YES|x) = \frac{25}{125}$. And an obvious classifier would be $f(x) = \arg\max_{y \in \mathcal{Y}} \bar{P}(y|x)$. This is hypothesis is called the Bayes hypothesis. The Bayes hypothesis gives the minimum possible error which is $R^* = \min\{P(y = YES|x), P(y = NO|x)\}$.

Extensions to PAC: Nonrealizability

The Bayes error R* is unknown because the true distribution $P(y|x)$ is unknown. We only know $\bar{P}(y|x)$ from the dataset. In this stochastic scenario, there's no hypothesis that can yield a generalization error of 0. So, we can't arbitrarily reduce the error by increasing the sample size. But we can compare the generalization error of a hypothesis with the Bayes error.

$$R(h) - R* = \underbrace{R(h) - \min_{h \in H} R(h)}_{\text{estimation}} + \underbrace{\min_{h \in H} R(h) - R*}_{\text{approximation}}$$

The approximation error is the comparison of the performance of the best hypothesis in class with that of the absolute best hypothesis. It's a property of the hypothesis set. It's unknown because the Bayes hypothesis is unknown. And the estimation error is the property of an individual hypothesis of set H. If we apply ERM, the estimation error of the hypothesis returned can be bound.

Estimation Error Bound for ERM

Let's say we are applying ERM on a dataset S of size m and the scenario is stochastic. Let's say $h*$ is the best hypothesis in class. But we might not be able to tell which one it is because we don't know the true distribution of the data.

$$\begin{aligned} R(h^{ERM}) - R(h*) &= R(h^{ERM}) - R_S(h^{ERM}) + R_S(h^{ERM}) - R(h*) \\ &\leq R(h^{ERM}) - R_S(h^{ERM}) + R_S(h*) - R(h*) \\ &\leq 2 \sup_{h \in H} |R(h) - R_S(h)| \\ &\leq 2\sqrt{\frac{\log |H| + \log \frac{2}{\delta}}{2m}} \end{aligned}$$

So, even though we cannot arbitrarily reduce the true error of a hypothesis in the stochastic scenario, we can at least make the true error converge to that of the best hypothesis in class. That's what we will aim to do in agnostic PAC learning.

Agnostic PAC-learning

In order to generalize the concept of learning:

1. We will replace the deterministic functions $h : \mathcal{X} \to \mathcal{Y}$ with a joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. So the concept class C disappears.
2. Instead of having a function $h(x) = y$, we will compute distributions $P(y|x)$.
3. $R(h_S)$ will be compared against $R(h*)$ where $h* = \arg\min_{h \in H} R(h)$.

Agnostic PAC-learning

Let H be a hypothesis set and $\mathcal{D}$ be a distribution over $\mathrm{X} \times \mathcal{Y}$. If for any sample S with size $m \geq poly(\frac{1}{\epsilon}, \frac{1}{\delta}, n, |H|)$, any $\epsilon > 0$ and any $\delta > 0$, there is an algorithm $\mathcal{A}$ so that the following holds:

$$Pr_{S \sim \mathcal{D}^m}[R(h_S) - \min_{h \in H} R(h) \leq \epsilon] \geq 1 - \delta$$

then $\mathcal{A}$ is said to be an agnostic PAC-learning algorithm.

Occam's Razor

$$R(h) \leq R_S(h) + \sqrt{\frac{\log |H| + \log \frac{2}{\delta}}{2m}}$$

In the above formula, we can see that the generalization error is bounded from above by the empirical error $R_S(h)$ and a function of H, δ, m. The larger the hypothesis set, the larger the scope of error. But we don't want to make the hypothesis set very small either because that will increase the empirical error. Also, if the sample size m is small, then it's highly likely that we will end up with a dubious and complex hypothesis. And that will lead to overfitting. If two hypothesis sets yield the same empirical error on a dataset, then the one with the smaller size gives a better generalization guarantee. An explicit restriction on the size of the hypothesis class puts an implicit restriction on the dimension and complexity of the hypothesis set as well.

Uniform Convergence

During ERM, we select a hypothesis that minimizes the empirical error. We only hope that this particular hypothesis will minimize the generalization as well. From this, we can derive a sufficient condition for learnability and that's called uniform convergence.

ϵ-representativeness

A training set S is said to be ϵ-representative with respect to a domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, a distribution $\mathcal{D}$ over $\mathcal{Z}$, a hypothesis class $\mathcal{H}$, and a loss function ℓ if:

$$\forall h \in \mathcal{H}, \qquad |R_S(h) - R_{\mathcal{D}}(h)| \leq \epsilon$$

This is a property of the sample S. If it gives an empirical error that's very close to the generalization error, then this sample gives a good glimpse of the true distribution. That's why any hypothesis chosen by ERM will perform well in the distribution $\mathcal{D}$.

Uniform Convergence contd.

Lemma 1

Assume that a training set S is $\frac{\epsilon}{2}-$representative with respect to domain $\mathcal{Z}$, distribution $\mathcal{D}$, hypothesis class $\mathcal{H}$, loss function ℓ. Then the output hypothesis of ERM namely $h_S = ERM_{\mathcal{H}}(S)$ $\in \arg\min_{h\in\mathcal{H}} R_S(h)$ satisfies:

$$R_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} R_{\mathcal{D}}(h) + \epsilon$$

Proof: Let's say, $h*$ be one of the best in class hypotheses. In other words, $h* \in \arg\min_{h\in\mathcal{H}} R_{\mathcal{D}}(h)$. So, we have to prove that $R_{\mathcal{D}}(h_S) - R_{\mathcal{D}}(h*) \leq \epsilon$.

$$\begin{aligned}
&R_{\mathcal{D}}(h_S) - R_{\mathcal{D}}(h*) \\
&= R_{\mathcal{D}}(h_S) - R_S(h_S) + R_S(h_S) - R_{\mathcal{D}}(h*) \\
&\leq R_{\mathcal{D}}(h_S) - R_S(h_S) + R_S(h*) - R_{\mathcal{D}}(h*) \\
&\leq |R_{\mathcal{D}}(h_S) - R_S(h_S)| + |R_S(h*) - R_{\mathcal{D}}(h*)| \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon
\end{aligned}$$

Agnostic PAC-learnability with Uniform Convergence

Uniform Convergence

A hypothesis class $\mathcal{H}$ is said to have the uniform convergence property with respect to a domain $\mathcal{Z}$ and a loss function ℓ, if there exists a sample size $m(\epsilon, \delta)$ so that for any distribution, any ϵ, and any δ, if we take a sample of size $m(\epsilon, \delta)$, then with probability at least $1 - \delta$, S is $\epsilon-$representative. Finite hypothesis classes have uniform convergence. We proved it in the last lecture. Remember?

Lemma 2

If a class has the uniform convergence property, then it's agnostic PAC-learnable.

Proof: Since H has the uniform convergence property, then with probability at least $1 - \delta$, any sample drawn from $\mathcal{Z}$ according to $\mathcal{D}$ is $\epsilon-$representative. And since, any S is $\epsilon -$ *representative*, if we apply ERM on $\mathcal{H}$ with sample S, then according to lemma 1,

$$R_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} R_{\mathcal{D}}(h) + 2\epsilon$$