Learning Theory

Dr. Víctor Uc Cetina

Facultad de Matemáticas
Universidad Autónoma de Yucatán

cetina@informatik.uni-hamburg.de
https://sites.google.com/view/victoruccetina

Content

Learning Theory

- Learning Theory is a research field devoted to studying the design and analysis of machine learning algorithms.

Learning Theory

- Learning Theory is a research field devoted to studying the design and analysis of machine learning algorithms.
- In particular, such algorithms aim at making accurate predictions or representations based on observations.

Learning Theory

- Learning Theory is a research field devoted to studying the design and analysis of machine learning algorithms.
- In particular, such algorithms aim at making accurate predictions or representations based on observations.
- The Association for Computational Learning (ACL) is in charge of the organization of the Conference on Learning Theory (COLT).

Figure: http://www.learningtheory.org

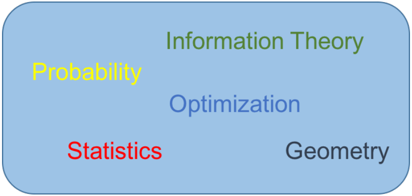
Learning Theory

- The emphasis is on rigorous mathematical analysis using tools from:

Information Theory
Probability
Optimization
Statistics
Geometry

Learning Theory

- The emphasis is on rigorous mathematical analysis using tools from:



- While theoretically rooted, learning theory puts a strong emphasis on efficient computation as well.

Learning Theory

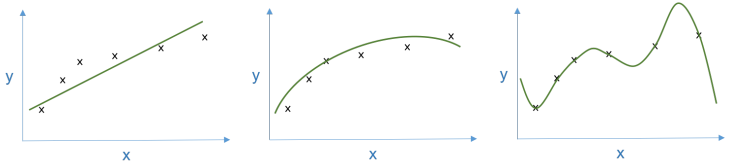
- Learning theory is a whole field of research within the broader field of machine learning.

Learning Theory

- Learning theory is a whole field of research within the broader field of machine learning.
- Important results in this field are usually published in:
  - Conference on Learning Theory (COLT).
  - Journal of Machine Learning Research (JMLR).

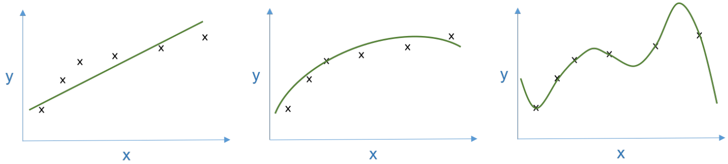
Bias Vs Variance

- When talking about linear regression, we wonder whether to fit a simple model such as the linear $y = \theta_0 + \theta_1 x$, or a more complex model such as the polynomial $y = \theta_0 + \theta_1 x + \ldots + \theta_5 x^5$.



Bias Vs Variance

- When talking about linear regression, we wonder whether to fit a simple model such as the linear $y = \theta_0 + \theta_1 x$, or a more complex model such as the polynomial $y = \theta_0 + \theta_1 x + \ldots + \theta_5 x^5$.



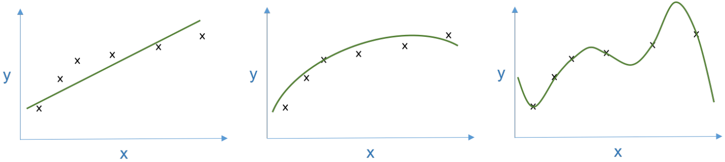
- Given this data set, the 5th order polynomial (rightmost curve) does not result in a good model, because it will not generalize well with other examples.

Bias Vs Variance

- The **generalization error** of a hypothesis $h_\theta(x)$ is its expected error on examples not necessarily in the training set.

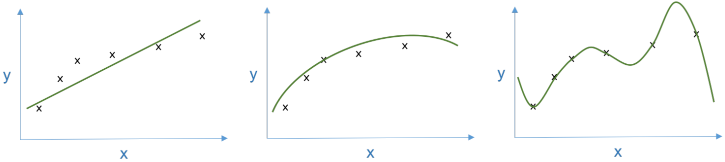
Bias Vs Variance

- The **generalization error** of a hypothesis $h_\theta(x)$ is its expected error on examples not necessarily in the training set.
- The models in the leftmost and the rightmost figures have large generalization errors.



Bias Vs Variance

- The **generalization error** of a hypothesis $h_\theta(x)$ is its expected error on examples not necessarily in the training set.
- The models in the leftmost and the rightmost figures have large generalization errors.



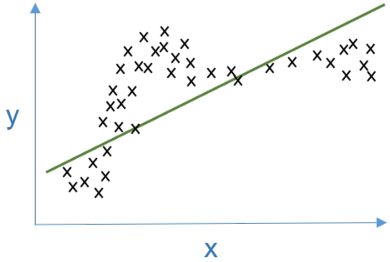
- However, the problems that the two models suffer from are very different.

Bias Vs Variance

- If the relationship between x and y is not linear, then even if we were fitting a linear model to a very large amount of training data, the linear model would still fail to accurately capture the structure in the data.

Bias Vs Variance

- If the relationship between x and y is not linear, then even if we were fitting a linear model to a very large amount of training data, the linear model would still fail to accurately capture the structure in the data.
- A linear model might be too simple to capture the structure in the data.

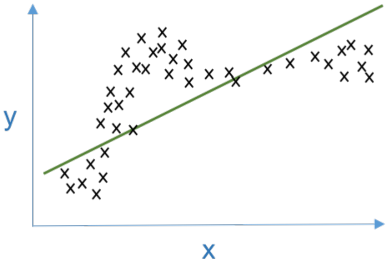


Bias Vs Variance

- Informally, we define the **bias** of a model to be the **expected generalization error** even if we were to fit it to a very large training set.

Bias Vs Variance

- Informally, we define the **bias** of a model to be the **expected generalization error** even if we were to fit it to a very large training set.
- Thus, for the problem above, the linear model suffers from **large bias**, and may underfit the data: it may fail to capture the structure exhibited by the data.

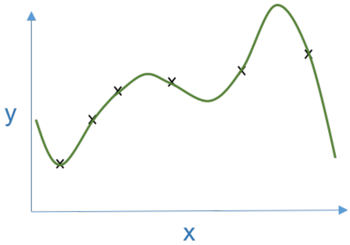


Bias Vs Variance

- The **variance** of a model fitting procedure is a second component of the generalization error.

Bias Vs Variance

- The **variance** of a model fitting procedure is a second component of the generalization error.
- When fitting a 5th order polynomial, there is a large risk that we're fitting patterns in the data that happened to be present in our small, training set, but that do not reflect the wider pattern of the relationship between x and y.



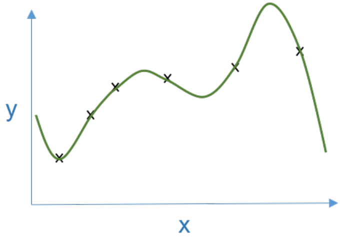
Bias Vs Variance

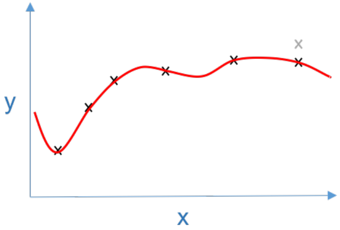
- Perhaps in the training set we happened by chance to get a slightly more higher than average value for some x.

Bias Vs Variance

- Perhaps in the training set we happened by chance to get a slightly more higher than average value for some x.
- By fitting this spurious pattern in the training set, we might again obtain a model with large generalization error.

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5.$$

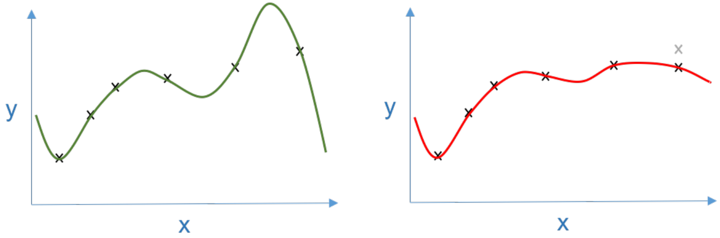




Bias Vs Variance

- Perhaps in the training set we happened by chance to get a slightly more higher than average value for some x.
- By fitting this spurious pattern in the training set, we might again obtain a model with large generalization error.

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5.$$



- In this case we say the model has a **large variance**.

Bias Vs Variance

- Often, there is a tradeoff between **bias** and **variance**.

Bias Vs Variance

- Often, there is a tradeoff between **bias** and **variance**.
- If our model is too simple and has very few parameters, then it may have large bias (but small variance).

  $h_\theta(x) = \theta_0 + \theta_1 x_1$.

Bias Vs Variance

- Often, there is a tradeoff between **bias** and **variance**.
- If our model is too simple and has very few parameters, then it may have large bias (but small variance).

  $h_\theta(x) = \theta_0 + \theta_1 x_1$.
- If our model is too complex and has very many parameters, then it may suffer from large variance (but smaller bias)

  $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5$.

Bias Vs Variance

- Often, there is a tradeoff between **bias** and **variance**.
- If our model is too simple and has very few parameters, then it may have large bias (but small variance).

  $h_\theta(x) = \theta_0 + \theta_1 x_1.$
- If our model is too complex and has very many parameters, then it may suffer from large variance (but smaller bias)

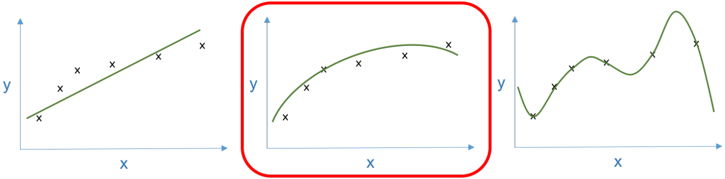
  $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5.$
- Therefore, in general we can say that

Model	Parameters	Bias	Variance
Simple	Few	Large	Small
Complex	Many	Small	Large

Bias Vs Variance

- In our example, fitting a quadratic function does better than either of the extremes of a first or a fifth order polynomial.

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$



Three learning theory questions

1. Can we make formal the bias/variance tradeoff that was just discussed?

Three learning theory questions

1. Can we make formal the bias/variance tradeoff that was just discussed?
2. Can we relate the error on the training set to the generalization error?

Three learning theory questions

1. Can we make formal the bias/variance tradeoff that was just discussed?
2. Can we relate the error on the training set to the generalization error?
3. Are there conditions under which we can actually prove that learning algorithms will work well?

Union Bound Lemma

- Let $A_1, A_2, \ldots, A_k$ be k different events (that may not be independent). Then

  $P(A_1 \cup A_2, \cup \ldots \cup A_k) \leq P(A_1) + \ldots + P(A_k)$.

Union Bound Lemma

- Let $A_1, A_2, \ldots, A_k$ be k different events (that may not be independent). Then

  $P(A_1 \cup A_2, \cup \ldots \cup A_k) \leq P(A_1) + \ldots + P(A_k).$

- In probability theory, the union bound is usually stated as an axiom, but it also makes intuitive sense: The probability of any one of k events happening is at most the sum of the probabilities of the k different events.

Hoeffding Inequality Lemma

- Let $Z_1, \ldots, Z_m$ be m independent and identically distributed (iid) random variables drawn from a Bernoulli(ϕ) distribution. I.e., $P(Z_i = 1) = \phi$, and $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} = (1/m)\sum_{i=1}^{m} Z_i$ be the mean of these random variables, and let any $\gamma > 0$ be fixed. Then

  $P(|\phi - \hat{\phi}| > \gamma) \leq 2\exp(-2\gamma^2 m)$

Hoeffding Inequality Lemma

- Let $Z_1, \ldots, Z_m$ be m independent and identically distributed (iid) random variables drawn from a Bernoulli(ϕ) distribution. I.e., $P(Z_i = 1) = \phi$, and $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} = (1/m) \sum_{i=1}^{m} Z_i$ be the mean of these random variables, and let any $\gamma > 0$ be fixed. Then

  $$P(|\phi - \hat{\phi}| > \gamma) \leq 2 \exp(-2\gamma^2 m)$$

- This lemma, also called the Chernoff bound, say that if we take $\hat{\phi}$ to be our estimate of ϕ, then the probability of our estimate to be far from the true value is small, as long as m is large.

Empirical Risk Minimization

- Using the Union Bound and the Hoeffding Inequality lemmas we will be able to prove some of the deepest and most important results in learning theory.

Empirical Risk Minimization

- Using the Union Bound and the Hoeffding Inequality lemmas we will be able to prove some of the deepest and most important results in learning theory.
- We will restrict our attention to the binary classification problem in which the labels are $y \in \{0, 1\}$. However, everything we will say here generalizes to other, including regression and multi-class classification problems.

Empirical Risk Minimization

- Using the Union Bound and the Hoeffding Inequality lemmas we will be able to prove some of the deepest and most important results in learning theory.
- We will restrict our attention to the binary classification problem in which the labels are $y \in \{0, 1\}$. However, everything we will say here generalizes to other, including regression and multi-class classification problems.
- We assume we are given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$, where the training examples are drawn iid from a probability distribution $\mathcal{D}$.

Empirical Risk Minimization

- For a hypothesis h, we define the **training error** (also called **empirical risk** or **empirical error** in learning theory) to be

$$\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} 1\left\{h(x^{(i)}) \neq y^{(i)}\right\}.$$

Empirical Risk Minimization

- For a hypothesis h, we define the **training error** (also called **empirical risk** or **empirical error** in learning theory) to be

$$\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} 1\left\{h(x^{(i)}) \neq y^{(i)}\right\}.$$

- This is just the fraction of training examples that h missclasifies.

Empirical Risk Minimization

- We also define the **generalization error** to be

  $\varepsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$.

Empirical Risk Minimization

- We also define the **generalization error** to be

  $\varepsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$.
- This is the probability that, if we now draw a new example (x, y) from the distribution probability $\mathcal{D}$, h will misclassify it.

Empirical Risk Minimization

- We also define the **generalization error** to be

  $\varepsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$.
- This is the probability that, if we now draw a new example (x, y) from the distribution probability $\mathcal{D}$, h will misclassify it.
- Note that we have assumed that the training data was drawn from the same distribution $\mathcal{D}$ with which we are going to evaluate our hypothesis h. This is sometimes also referred to as one of the **PAC** assumptions.

Empirical Risk Minimization

PAC stands for "probably approximately correct", which is a framework and set of assumptions under which numerous results on learning theory were proved.

Empirical Risk Minimization

PAC stands for "probably approximately correct", which is a framework and set of assumptions under which numerous results on learning theory were proved.

Of these, the assumption of training and testing on the same distribution, and the assumption of the independently drawn training examples, were the most important.

Empirical Risk Minimization

- Consider the setting of linear classification, and let $h_\theta(\mathbf{x}) = 1\{\theta^\top \mathbf{x}\}$. What is a reasonable way of fitting the parameters θ?

Empirical Risk Minimization

- Consider the setting of linear classification, and let $h_\theta(\mathbf{x}) = 1\ \{\theta^\top \mathbf{x}\}$. What is a reasonable way of fitting the parameters θ?
- One approach is to try to minimize the training error, and pick $\hat{\theta} = \arg\min_\theta \hat{\varepsilon}(h_\theta)$.

Empirical Risk Minimization

- Consider the setting of linear classification, and let $h_\theta(\mathbf{x}) = 1\{\theta^\top \mathbf{x}\}$. What is a reasonable way of fitting the parameters θ?
- One approach is to try to minimize the training error, and pick
  $$\hat{\theta} = \arg\min_{\theta} \hat{\varepsilon}(h_\theta).$$
- We call this process **empirical risk minimization (ERM)**, and the resulting hypothesis output by the learning algorithm is $\hat{h} = h_{\hat{\theta}}$.

Empirical Risk Minimization

- Consider the setting of linear classification, and let $h_\theta(\mathbf{x}) = 1\,\{\theta^\top \mathbf{x}\}$. What is a reasonable way of fitting the parameters θ?
- One approach is to try to minimize the training error, and pick
  $$\hat{\theta} = \arg\min_{\theta} \hat{\varepsilon}(h_\theta).$$
- We call this process **empirical risk minimization (ERM)**, and the resulting hypothesis output by the learning algorithm is $\hat{h} = h_{\hat{\theta}}$.
- We think of ERM as the most basic learning algorithm. Algorithms such as logistic regression can also be viewed as approximations to ERM.

Hypothesis Class $\mathcal{H}$

- We define the **hypothesis class** $\mathcal{H}$ used by a learning algorithm to be the set of all classifiers considered by it.

Hypothesis Class $\mathcal{H}$

- We define the **hypothesis class** $\mathcal{H}$ used by a learning algorithm to be the set of all classifiers considered by it.
- For linear classification, $\mathcal{H} = \{h_\theta : h_\theta(x) = 1\ \{\theta^\top \mathbf{x} \geq 0\}, \theta \in \mathbb{R}^{n+1}\}$ is thus the set of all classifiers $\mathcal{X}$ (the domain of the inputs) where the decision boundary is linear.

Hypothesis Class $\mathcal{H}$

- We define the **hypothesis class** $\mathcal{H}$ used by a learning algorithm to be the set of all classifiers considered by it.
- For linear classification, $\mathcal{H} = \{h_\theta : h_\theta(x) = 1\ \{\theta^\top \mathbf{x} \geq 0\}, \theta \in \mathbb{R}^{n+1}\}$ is thus the set of all classifiers $\mathcal{X}$ (the domain of the inputs) where the decision boundary is linear.
- If we were studying neural networks, then we could let $\mathcal{H}$ be the set of all classifiers representable by some neural network architecture.

Hypothesis Class $\mathcal{H}$

- Empirical risk minimization can now be thought of as a minimization over the class of functions $\mathcal{H}$, in which the learning algorithm picks the hypothesis:

  $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$.

The Finite Class $\mathcal{H}$

- Lets start by considering a learning problem in which we have a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_k\}$ consisting of k hypothesis.

The Finite Class $\mathcal{H}$

- Lets start by considering a learning problem in which we have a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_k\}$ consisting of k hypothesis.
- Thus, $\mathcal{H}$ is just a set of k functions mapping from $\mathcal{X}$ to $\{0, 1\}$, and empirical risk minimization selects $\hat{h}$ to be whichever of these k functions has smallest training error.

The Finite Class $\mathcal{H}$

- Lets start by considering a learning problem in which we have a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_k\}$ consisting of k hypothesis.
- Thus, $\mathcal{H}$ is just a set of k functions mapping from $\mathcal{X}$ to $\{0, 1\}$, and empirical risk minimization selects $\hat{h}$ to be whichever of these k functions has smallest training error.
- To give some guarantees on the generalization error of $\hat{h}$, we will:

The Finite Class $\mathcal{H}$

- Lets start by considering a learning problem in which we have a finite hypothesis class $\mathcal{H} = \{h_1, \dots, h_k\}$ consisting of k hypothesis.
- Thus, $\mathcal{H}$ is just a set of k functions mapping from $\mathcal{X}$ to $\{0, 1\}$, and empirical risk minimization selects $\hat{h}$ to be whichever of these k functions has smallest training error.
- To give some guarantees on the generalization error of $\hat{h}$, we will:
  1. Show that $\hat{\varepsilon}(h)$ is a reliable estimate of $\varepsilon(h)$ for all h.

The Finite Class $\mathcal{H}$

- Lets start by considering a learning problem in which we have a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_k\}$ consisting of k hypothesis.
- Thus, $\mathcal{H}$ is just a set of k functions mapping from $\mathcal{X}$ to $\{0, 1\}$, and empirical risk minimization selects $\hat{h}$ to be whichever of these k functions has smallest training error.
- To give some guarantees on the generalization error of $\hat{h}$, we will:
  1. Show that $\hat{\varepsilon}(h)$ is a reliable estimate of $\varepsilon(h)$ for all h.
  2. Show that this implies an upper-bound on the generalization error of $\hat{h}$.

The Finite Class $\mathcal{H}$

- Take any one fixed $h_i \in \mathcal{H}$.

The Finite Class $\mathcal{H}$

- Take any one fixed $h_i \in \mathcal{H}$.
- Consider a Bernoulli random variable Z whose distribution is defined as follows:

The Finite Class $\mathcal{H}$

- Take any one fixed $h_i \in \mathcal{H}$.
- Consider a Bernoulli random variable Z whose distribution is defined as follows:
- We are going to sample $(x, y) \sim \mathcal{D}$. Then we set $Z = 1\{h_i(x) \neq y\}$.

The Finite Class $\mathcal{H}$

- Take any one fixed $h_i \in \mathcal{H}$.
- Consider a Bernoulli random variable Z whose distribution is defined as follows:
- We are going to sample $(x, y) \sim \mathcal{D}$. Then we set $Z = 1\left\{h_i(x) \neq y\right\}$.
- Similarly, we also define $Z_j = 1\left\{h_i(x^{(j)}) \neq y^{(j)}\right\}$.

The Finite Class $\mathcal{H}$

- Take any one fixed $h_i \in \mathcal{H}$.
- Consider a Bernoulli random variable Z whose distribution is defined as follows:
- We are going to sample $(x, y) \sim \mathcal{D}$. Then we set $Z = 1\left\{h_i(x) \neq y\right\}$.
- Similarly, we also define $Z_j = 1\left\{h_i(x^{(j)}) \neq y^{(j)}\right\}$.
- Since, our training set was drawn iid from $\mathcal{D}$, then Z and the Z_j's have the same distribution.

The Finite Class $\mathcal{H}$

- We see that the misclassification probability on a randomly drawn example, that is $\varepsilon(h)$, is exactly the expected value of Z (and Z_j's).

The Finite Class $\mathcal{H}$

- We see that the misclassification probability on a randomly drawn example, that is $\varepsilon(h)$, is exactly the expected value of Z (and Z_j's).
- Moreover the training error can be written

$$\hat{\varepsilon}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j.$$

The Finite Class $\mathcal{H}$

- We see that the misclassification probability on a randomly drawn example, that is $\varepsilon(h)$, is exactly the expected value of Z (and Z_j's).
- Moreover the training error can be written

$$\hat{\varepsilon}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j.$$

- Thus, $\hat{\varepsilon}(h_i)$ is exactly the mean of the m random variables Z_j that are drawn iid from a Bernoulli distribution $\mathcal{D}$ with mean $\varepsilon(h_i)$.

The Finite Class $\mathcal{H}$

- We see that the misclassification probability on a randomly drawn example, that is $\varepsilon(h)$, is exactly the expected value of Z (and Z_j's).
- Moreover the training error can be written

$$\hat{\varepsilon}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j.$$

- Thus, $\hat{\varepsilon}(h_i)$ is exactly the mean of the m random variables Z_j that are drawn iid from a Bernoulli distribution $\mathcal{D}$ with mean $\varepsilon(h_i)$.
- We can apply the Hoeffding inequality and obtain

$$P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

The Finite Class $\mathcal{H}$

- The expression $P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2\exp(-2\gamma^2 m)$

  shows that for a particular h_i the training error will be close to the generalization error with high probability, assuming that m is large.

The Finite Class $\mathcal{H}$

- The expression $P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2\exp(-2\gamma^2 m)$

  shows that for a particular h_i the training error will be close to the generalization error with high probability, assuming that m is large.
- Let A_i denote the event that $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$.

The Finite Class $\mathcal{H}$

- The expression $P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2\exp(-2\gamma^2 m)$

  shows that for a particular h_i the training error will be close to the generalization error with high probability, assuming that m is large.
- Let A_i denote the event that $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$.
- We have seen that for any particular A_i it holds true that $P(A_i) \leq 2\exp(-2\gamma^2 m)$.

The Finite Class $\mathcal{H}$

Thus, using the union bound, we have that

$$P(\exists h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \quad = P(A_1 \cup \ldots \cup A_k)$$

The Finite Class $\mathcal{H}$

Thus, using the union bound, we have that

$$\begin{aligned} P(\exists h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \ldots \cup A_k) \\ &\leq \sum_{i=1}^{k} P(A_i) \end{aligned}$$

The Finite Class $\mathcal{H}$

Thus, using the union bound, we have that

$$\begin{aligned} P(\exists h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \ldots \cup A_k) \\ &\leq \sum_{i=1}^{k} P(A_i) \\ &\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 m) \end{aligned}$$

The Finite Class $\mathcal{H}$

Thus, using the union bound, we have that

$$
\begin{aligned}
P(\exists h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \ldots \cup A_k) \\
&\leq \sum_{i=1}^{k} P(A_i) \\
&\leq \sum_{i=1}^{k} 2 \exp(-2\gamma^2 m) \\
&= 2k \exp(-2\gamma^2 m)
\end{aligned}
$$

The Finite Class $\mathcal{H}$

Thus, using the union bound, we have that

$$\begin{aligned} P(\exists h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \ldots \cup A_k) \\ &\leq \sum_{i=1}^{k} P(A_i) \\ &\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 m) \\ &= 2k\exp(-2\gamma^2 m) \end{aligned}$$

If we substract both sides from 1, we find that

$$P(\neg\exists h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) = P(\forall h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma)$$

The Finite Class $\mathcal{H}$

Thus, using the union bound, we have that

$$
\begin{aligned}
P(\exists h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \ldots \cup A_k) \\
&\leq \sum_{i=1}^{k} P(A_i) \\
&\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 m) \\
&= 2k\exp(-2\gamma^2 m)
\end{aligned}
$$

If we substract both sides from 1, we find that

$$
\begin{aligned}
P(\neg\exists h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(\forall h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \\
& 1 - 2k\exp(-2\gamma^2 m)
\end{aligned}
$$

The Finite Class $\mathcal{H}$

Thus, the expression

$$P(\forall h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \geq 1 - 2k\ \exp(-2\gamma^2 m)$$

The Finite Class $\mathcal{H}$

Thus, the expression

$$P(\forall h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \geq 1 - 2k\,\exp(-2\gamma^2 m)$$

means that, with probability at least $1 - 2k\,\exp(-2\gamma^2 m)$, we have that $\varepsilon(h)$ will be within γ of $\hat{\varepsilon}(h)$ for all $h \in \mathcal{H}$.

The Finite Class $\mathcal{H}$

Thus, the expression

$$P(\forall h \in \mathcal{H},\ |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \geq 1 - 2k\ \exp(-2\gamma^2 m)$$

means that, with probability at least $1 - 2k\ \exp(-2\gamma^2 m)$, we have that $\varepsilon(h)$ will be within γ of $\hat{\varepsilon}(h)$ for all $h \in \mathcal{H}$.

This is called a **uniform convergence** result, because this is a bound that holds simultaneously for all $h \in \mathcal{H}$.

The Finite Class $\mathcal{H}$

- Given γ and some $\delta > 0$, how large must m be before we can guarantee that with probability at least $1 - \delta$, the training error will be within γ of the generalization error?

The Finite Class $\mathcal{H}$

- Given γ and some $\delta > 0$, how large must m be before we can guarantee that with probability at least $1 - \delta$, the training error will be within γ of the generalization error?
- By setting $\delta = 2k \exp(-2\gamma^2 m)$ and solving for m, we find that

  $m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$,

  then with probability at least $1 - \delta$, we have that $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma$ for all $h \in \mathcal{H}$.

The Finite Class $\mathcal{H}$

- Given γ and some $\delta > 0$, how large must m be before we can guarantee that with probability at least $1 - \delta$, the training error will be within γ of the generalization error?
- By setting $\delta = 2k \exp(-2\gamma^2 m)$ and solving for m, we find that

  $m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$,

  then with probability at least $1 - \delta$, we have that $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma$ for all $h \in \mathcal{H}$.
- This bound tells us how many training examples we need in order to make a guarantee. The training set size m that a certain method or algorithm requires in order to achieve a certain level of performance is also called the algorithm's sample complexity.

The Finite Class $\mathcal{H}$

- The key property of the bound above is that the number of training examples needed to make this guarantee is only logarithmic in k, the number of hypotheses in $\mathcal{H}$.

The Finite Class $\mathcal{H}$

- The key property of the bound above is that the number of training examples needed to make this guarantee is only logarithmic in k, the number of hypotheses in $\mathcal{H}$.
- Similarly, if we fix m and δ and solve for γ, we get that with probability $1-\delta$, we have that for all $h \in \mathcal{H}$,

  $$|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}.$$

The Finite Class $\mathcal{H}$

- Define $h^* = \arg\min_{h \in \mathcal{H}} \varepsilon(h)$ to be the best possible hypothesis in $\mathcal{H}$.

The Finite Class $\mathcal{H}$

- Define $h^* = \arg\min_{h \in \mathcal{H}} \varepsilon(h)$ to be the best possible hypothesis in $\mathcal{H}$.
- We have:

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \end{aligned}$$

The Finite Class $\mathcal{H}$

- Define $h^* = \arg\min_{h \in \mathcal{H}} \varepsilon(h)$ to be the best possible hypothesis in $\mathcal{H}$.
- We have:

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma \end{aligned}$$

- In the first and third step we used $|\varepsilon(\hat{h}) - \hat{\varepsilon}(\hat{h})| \leq \gamma$ and $\hat{\varepsilon}(h^*) \leq \varepsilon(h^*) + \gamma$, respectively.

The Finite Class $\mathcal{H}$

- **Theorem.** Let $|\mathcal{H}| = k$, and let any m, δ be fixed. Then with probability at least $1 - \delta$, we have that

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h) \right) + 2\sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$

The Finite Class $\mathcal{H}$

- **Theorem.** Let $|\mathcal{H}| = k$, and let any m, δ be fixed. Then with probability at least $1 - \delta$, we have that

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h) \right) + 2\sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$

- When we switch from $\mathcal{H}$ to $\mathcal{H}'$ where $\mathcal{H}' \supseteq \mathcal{H}$, the first term can only decrease, and therefore our bias can only decrease.

The Finite Class $\mathcal{H}$

- **Theorem.** Let $|\mathcal{H}| = k$, and let any m, δ be fixed. Then with probability at least $1 - \delta$, we have that

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h) \right) + 2\sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$

- When we switch from $\mathcal{H}$ to $\mathcal{H}'$ where $\mathcal{H}' \supseteq \mathcal{H}$, the first term can only decrease, and therefore our bias can only decrease.
- However, if k increases, then the second term will would also increase, and therefore the variance would increase as well.

The Finite Class $\mathcal{H}$

- **Corollary.** Let $|\mathcal{H}| = k$, and let any δ, γ be fixed. Then for $\varepsilon(\hat{h}) \leq \min_{h \in \mathcal{H}} \varepsilon(h) + 2\gamma$ to hold with probability at least $1 - \delta$, it suffices that

$$m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$

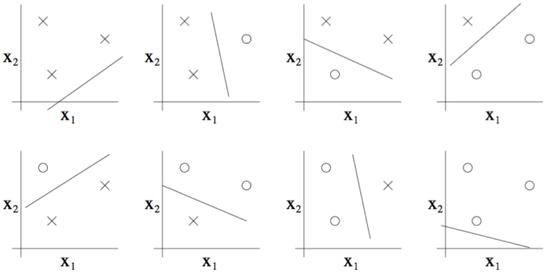
The Finite Class $\mathcal{H}$

- **Corollary.** Let $|\mathcal{H}| = k$, and let any δ, γ be fixed. Then for $\varepsilon(\hat{h}) \leq \min_{h\in\mathcal{H}} \varepsilon(h) + 2\gamma$ to hold with probability at least $1 - \delta$, it suffices that

$$\begin{aligned} m &\geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta} \\ &= O\left(\frac{1}{\gamma^2} \log \frac{k}{\delta}\right) \end{aligned}$$

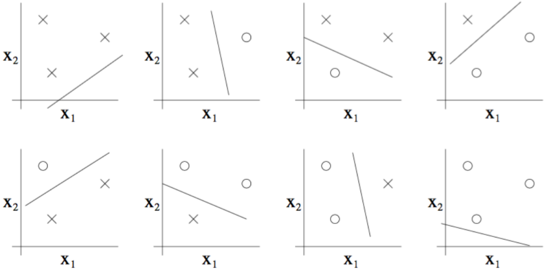
Vapnik-Chervonenkis Dimension

- **Definition.** Given a hypothesis class $\mathcal{H}$, we then define its Vapnik-Chervonenkis dimension, written VC($\mathcal{H}$), to be the size of the largest set of points that is shattered by $\mathcal{H}$.



Vapnik-Chervonenkis Dimension

- **Definition.** Given a hypothesis class $\mathcal{H}$, we then define its Vapnik-Chervonenkis dimension, written $\mathrm{VC}(\mathcal{H})$, to be the size of the largest set of points that is shattered by $\mathcal{H}$.



- For a linear classifier in two dimensions, $\mathrm{VC}(\mathcal{H}) = 3$.

The Infinite Class $\mathcal{H}$

- **Theorem.** Let $\mathcal{H}$ be given, and let $d = \mathrm{VC}(\mathcal{H})$. Then with probability at least $1 - \delta$, we have that for all $h \in \mathcal{H}$,

  $$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq O\left(\sqrt{\frac{d}{m} \log \frac{m}{d} + \frac{1}{m} \log \frac{1}{\delta}}\right).$$

The Infinite Class $\mathcal{H}$

- **Theorem.** Let $\mathcal{H}$ be given, and let $d = \mathrm{VC}(\mathcal{H})$. Then with probability at least $1 - \delta$, we have that for all $h \in \mathcal{H}$,

  $$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right).$$

- Thus, with probability at least $1 - \delta$, we also have that:

  $$\varepsilon(\hat{h}) \leq \varepsilon(h^*)| + O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right).$$

The Infinite Class $\mathcal{H}$

- **Theorem.** Let $\mathcal{H}$ be given, and let $d = \mathrm{VC}(\mathcal{H})$. Then with probability at least $1-\delta$, we have that for all $h \in \mathcal{H}$,

  $$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right).$$

- Thus, with probability at least $1-\delta$, we also have that:

  $$\varepsilon(\hat{h}) \leq \varepsilon(h^*)| + O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right).$$

- In other words, if a hypothesis class has finite VC dimension, then uniform convergence occurs as m becomes large. This allows us to give a bound on $\varepsilon(h)$ in terms of $\varepsilon(h^*)$.

Reference

- Andrew Ng. **Machine Learning Course Notes**. 2003.
- Christopher Bishop. **Pattern Recognition and Machine Learning**. Springer. 2006.

Thank you!

Dr. Victor Uc Cetina
cetina@informatik.uni-hamburg.de