SCIENCE ACADEMY

Break Point of an Hypothesis Set

- Given an hypothesis set $\mathcal{H}$
- A hypothesis set $\mathcal{H}$ **shatters N points** $\iff m_{\mathcal{H}}(N) = 2^N$
  - There is a position of N points and a class assignment that you can classify using $h \in \mathcal{H}$
  - It does not mean all sets of N points can be classified in any way
- k is a **break point** for $\mathcal{H} \iff m_{\mathcal{H}}(k) < 2^k$
  - I.e., no data set of size k can be shattered by $\mathcal{H}$
  - E.g.,
    - For 2D perceptron: a break point is 4
    - For positive rays: a break point is 2
    - For positive intervals: a break point is 3
    - For convex set on a plane: there is no break point

Break Point for an Hypothesis Set and Learning

- If there is a break point for a hypothesis set $\mathcal{H}$, it can be shown that:
  - $m_{\mathcal{H}}(N)$ is polynomial in N
  - Instead of Hoeffding's inequality for learning

    $$\Pr(|E_{in}(g) - E_{out}(g)| > \varepsilon) \leq 2Me^{-2\varepsilon^2 N}$$

    you can use the Vapnik-Chervonenkis inequality:

    $$\Pr(\text{bad generalization}) \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\varepsilon^2 N}$$

  - Since $m_{\mathcal{H}}(N)$ is polynomial in N, it will be dominated by the negative exponential, given enough examples
  - You can have a generalization bound: machine learning works!
- A hypothesis set can be characterized from the learning point of view by the **existence and value of a break point**

SCIENCE ACADEMY

- ***The VC Dimension***
- Overfitting
- Bias Variance Analysis

VC Dimension of an Hypothesis Set

- The **VC dimension of a hypothesis set** $\mathcal{H}$, denoted as $d_{VC}(\mathcal{H})$, is defined as the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$
  - I.e., the VC dimension is the most points $\mathcal{H}$ can shatter
- **Properties** of the VC dimension: if $d_{VC}(\mathcal{H}) = N$ then
  - Exists a constellation of N points that can be shattered by $\mathcal{H}$
    - Not all sets of N points can be shattered
    - If N points were placed randomly, they could not be necessarily shattered
  - $\mathcal{H}$ can *shatter* N points for any $N \leq d_{VC}(\mathcal{H})$
  - The *smallest break point* is $d_{VC} - 1$
  - The *growth function* in terms of the VC dimension is $m_{\mathcal{H}} \leq \sum_{i=0}^{d_{VC}} \binom{N}{i}$
  - The VC dimension is the *order of the polynomial bounding* $m_{\mathcal{H}}$

SCIENCE ACADEMY

VC Dimension: Interpretation

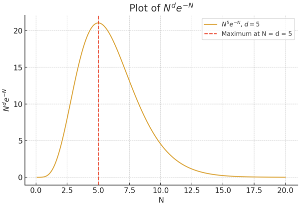
- The VC dimension **measures the complexity** of a hypothesis set in terms of **effective parameters**
- E.g.,
  - A perceptron in a d-dimensional space has $d_{VC} = d + 1$
  - In fact d_{VC} is the number of perceptron parameters!
  - E.g., for a 2D perceptron ($d = 2$), the break point is 2, so $d_{VC} = 3$
- The VC dimension considers the model as a black box in order to estimate effective parameters
  - How many points N a model can shatter, not the number of parameters
- Not all parameters contribute to degrees of freedom
  - E.g., combining N 1D perceptrons gives $2N$ parameters, but the effective degrees of freedom remain 2
- A complex hypothesis $\mathcal{H}$:
  - Has more parameters (higher VC dimension d_{VC})
  - Requires more examples for training

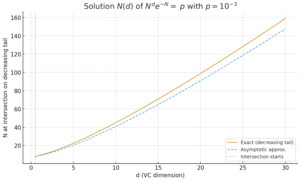
VC Generalization Bounds

- How many data points are needed to obtain $\Pr(|E_{in} - E_{out}| > \varepsilon) \leq \delta$?
- The VC inequality states

$$\Pr(\text{bad generalization}) \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\varepsilon^2 N}$$

- $N^d e^{-N}$ abstracts the upper bound term
  - Plot $N^d e^{-N}$ vs. N: Power dominates for small N, exponential for large N and brings it to 0
  - Vary d (VC dimension) function peaks for larger N, then approaches the region of interest ≤ 1
- Plot intersection of $N^d e^{-N}$ with a probability as a function of d
  - Examples N needed are proportional to d
  - Rule of thumb: $N \geq 10 d_{VC}$ for generalization





SCIENCE ACADEMY

VC Generalization Bounds

- The VC inequality

$$\Pr(|E_{in} - E_{out}| > \varepsilon) \le 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\varepsilon^2 N}$$

can be used in several ways to relate ε, δ, and N, e.g.,

- Examples
  - "Given $\varepsilon = 1\%$ error, how many examples N are needed to get $\delta = 0.05$?"
  - "Given N examples, what's the probability of an error larger than ε?"
- You can equate δ to $4m_{\mathcal{H}}(2N)e^{\frac{1}{8}\varepsilon^2 N}$ and solve for ε, getting

$$\Omega(N, \mathcal{H}, \delta) = \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

- Then you can say $|E_{out} - E_{in}| \le \Omega(N, \mathcal{H}, \delta)$ with probability $\ge 1 - \delta$
  - The generalization bounds are then: $\Pr(E_{out} \le E_{in} + \Omega) \ge 1 - \delta$

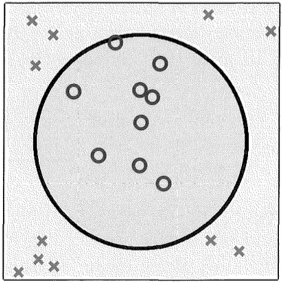
SCIENCE ACADEMY

How to Void the VC Analysis Guarantee

- Consider the case where data is genuinely non-linear
  - E.g., "o" points in the center and "x" in the corners
- Transform to high-dimensional $\mathcal{Z}$ with:

$$\Phi : \underline{\boldsymbol{x}} = (x_0, \ldots, x_d) \rightarrow \underline{\boldsymbol{z}} = (z_0, \ldots, z_{\tilde{d}})$$

- $d_{VC} \leq \tilde{d} + 1$; smaller $\tilde{d}$ improves generalization
  - Use $\underline{\boldsymbol{z}} = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)$
  - Why not $\underline{\boldsymbol{z}} = (1, x_1^2, x_2^2)$?
  - Why not $\underline{\boldsymbol{z}} = (1, x_1^2 + x_2^2)$?
  - Why not $\underline{\boldsymbol{z}} = (x_1^2 + x_2^2 - 0.6)$?

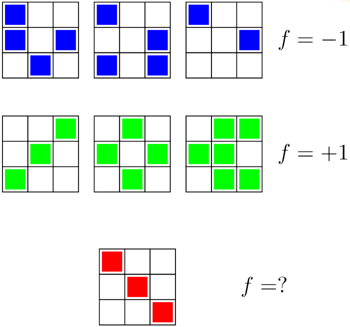


- Some model coefficients were zero and discarded, leaving machine learning the rest
  - VC analysis is a warranty, forfeited if data is examined before model selection (data snooping)
  - From VC analysis, complexity is that of the initial hypothesis set

- The VC Dimension
- ***Overfitting***
- Bias Variance Analysis

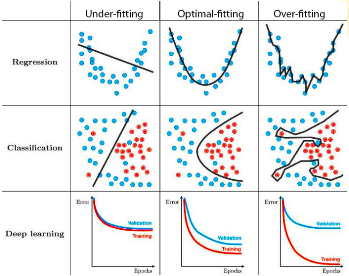
Overfitting: Definition

- **Overfitting** occurs when the model fits the data more than what is warranted
- Surpass point where E_{out} is minimal (optimal fit)
  - Model complexity too high for data/noise
  - Noise in training set mistaken for signal
- **Fitting noise instead of signal** is not useless but harmful
  - Model infers in-sample pattern that, when extrapolated out-of-sample, deviates from target function $\implies$ poor generalization



Optimal Fit

- The opposite of overfitting is **optimal fit**
  - Train a model with the proper complexity for the data
- The optimal fit:
  - Implies that E_{out} is minimal
  - Does not imply that generalization error $E_{out} - E_{in}$ is minimal (e.g., no training at all implies generalization error equal to 0)
- The **generalization error** is the additional error $E_{out} - E_{in}$ you see when you go from in-sample to out-of-sample

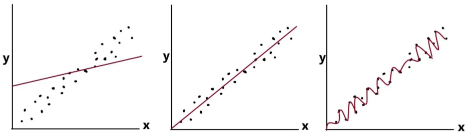


Overfitting: Diamond Price Example

- Predict diamond price as a function of carat size (regression problem)
- True relationship:

$$\text{price} \sim (\text{carat size})^2 + \varepsilon$$

  where:
  - Square function: price increases more with rarity
  - Noise: e.g., market noise, missing features
- **Fit with:**
  - **Line**
    - Underfit
    - High bias (large error)
    - Low variance (stable model)
  - **Polynomial of degree 2**
    - right fit
  - **Polynomial of degree 10**
    - Overfit (wiggly curve)
    - Low bias



Overfitting: 2-Features Classification Example

- Assume:
  - We want to separate 2 classes using 2 features x_1, x_2
  - The class boundary of sample points has a parabola shape
- We can use logistic regression and a decision boundary equal to:
  - A line $\text{logit}(w_0 + w_1 x + w_2 y) \to$ underfit
    - High bias, low variance
  - A parabola $\text{logit}(w_0 + w_1 x + w_2 x^2 + w_3 xy + w_4 y^2) \to$ right fit
  - A wiggly decision boundary $\text{logit}(w_0 +$ high powers of $x_1, x_2) \to$ overfit
    - Low bias, high variance

Margin in Classification

- Classification margin is the difference between the chosen class and the next predicted class
- Even if the error on training data gets to 0, one can improve out-of-sample performance by increasing the margin
    - More robust to noise

- The VC Dimension
- Overfitting
- ***Bias Variance Analysis***

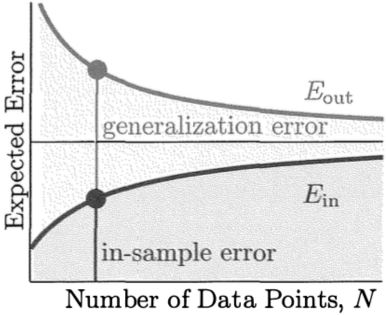
VC Analysis vs Bias-Variance Analysis

- Both VC analysis and bias-variance analysis are concerned with the hypothesis set $\mathcal{H}$
  - VC analysis:

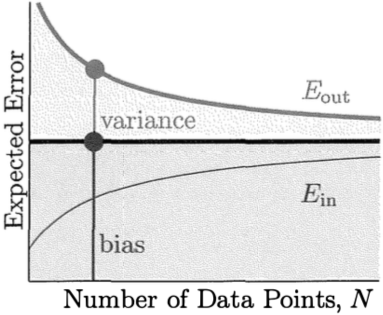
$$E_{out} \leq E_{in} + \Omega(\mathcal{H})$$

  - Bias-variance analysis:

$$E_{out} = \text{bias and variance}$$



VC Analysis



Bias-Variance Analysis

Hypothesis Set and Bias-Variance Analysis

- Learning consists in finding $g \in \mathcal{H}$ such that $g \approx f$ where f is an unknown function
- The tradeoff in learning is between:
  - Bias vs variance
  - Overfitting vs underfitting
  - More complex vs less complex $\mathcal{H}$ / h
  - Approximation (in-sample) vs generalization (out-of-sample)

Decomposing Error in Bias-Variance

- Consider machine learning problem
  - Regression set-up: target is a real-valued function
  - Hypothesis set $\mathcal{H} = \{h_1(\underline{\boldsymbol{x}}), h_2(\underline{\boldsymbol{x}}), \dots h_n(\underline{\boldsymbol{x}})\}$
  - Training data D with N examples
  - Error is squared error $E_o ut = \mathbb{E}[(g(\underline{\boldsymbol{x}}) - f(\underline{\boldsymbol{x}}))^2]$
  - Choose the best function g from $\mathcal{H}$ that approximates f
- What is the out-of-sample error $E_{out}(g)$ as function of $\mathcal{H}$ for a training set of N examples?

Decomposing Error in Bias-Variance

- The final hypothesis g depends on the training set D, so we make the dependency explicit $g^{(D)}$:

$$E_{out}(g^{(D)}) \triangleq \mathbb{E}_{\underline{\boldsymbol{x}}}[(g^{(D)}(\underline{\boldsymbol{x}}) - f(\underline{\boldsymbol{x}}))^2]$$

- We are interested in:
  - The hypothesis set $\mathcal{H}$ rather than the specific h; and
  - In a training set D of N examples, rather than the specific D
- Therefore we Remove the dependency from D by averaging over all the possible training sets D with N examples:

$$E_{out}(\mathcal{H}) \triangleq \mathbb{E}_D[E_{out}(g^{(D)})] = \mathbb{E}_D[\mathbb{E}_{\underline{\boldsymbol{x}}}[(g^{(D)}(\underline{\boldsymbol{x}}) - f(\underline{\boldsymbol{x}}))^2]]$$

Decomposing Error in Bias-Variance

- Switch the order of the expectations since the quantity is non-negative:

$$E_{out}(\mathcal{H}) = \mathbb{E}_{\underline{\boldsymbol{x}}}[\mathbb{E}_D[(g^{(D)}(\underline{\boldsymbol{x}}) - f(\underline{\boldsymbol{x}}))^2]$$

- Focus on $\mathbb{E}_D[(g^{(D)}(\underline{\boldsymbol{x}}) - f(\underline{\boldsymbol{x}}))^2]$ which is a function of $\underline{\boldsymbol{x}}$
- Define the *average hypothesis* over all training sets as:

$$\overline{g}(\underline{\boldsymbol{x}}) \triangleq \mathbb{E}_D[g^{(D)}(\underline{\boldsymbol{x}})]$$

- Add and subtract it inside the $\mathbb{E}_D$ expression:

$$\begin{aligned}
E_{out}(\mathcal{H}) =& \mathbb{E}_{\underline{\boldsymbol{x}}}\left[\mathbb{E}_D\left[\left(g^{(D)}(\underline{\boldsymbol{x}}) - f(\underline{\boldsymbol{x}})\right)^2\right]\right] \\
=& \mathbb{E}_{\underline{\boldsymbol{x}}}\mathbb{E}_D[(g^{(D)} - \overline{g} + \overline{g} - f)^2] \\
=& \mathbb{E}_{\underline{\boldsymbol{x}}}\mathbb{E}_D[(g^{(D)} - \overline{g})^2 + (\overline{g} - f)^2 + 2(g^{(D)} - \overline{g})(\overline{g} - f)] \\
& (\mathbb{E}_D \text{ is linear and } (\overline{g} - f) \text{ doesn't depend on } D) \\
=& \mathbb{E}_{\underline{\boldsymbol{x}}}\left[\mathbb{E}_D[(g^{(D)} - \overline{g})^2] + (\overline{g} - f)^2 + 2\mathbb{E}_D[(g^{(D)} - \overline{g})](\overline{g} - f)\right]
\end{aligned}$$

Decomposing Error in Bias-Variance

- The cross term:

$$\mathbb{E}_D[(g^{(D)} - \overline{g})](\overline{g} - f)$$

disappears since applying the expectation on D, it is equal to:

$$(g^{(D)} - \mathbb{E}_D[\overline{g}])(\overline{g} - f) = 0 \cdot (\overline{g} - f) = 0 \cdot \text{constant}$$

- Finally:

$$\begin{aligned} E_{out}(\mathcal{H}) &= \mathbb{E}_{\underline{\boldsymbol{x}}}[\mathbb{E}_D[(g^{(D)} - \overline{g})^2] + (\overline{g}(\underline{\boldsymbol{x}}) - f(\underline{\boldsymbol{x}}))^2] \\ &= \mathbb{E}_{\underline{\boldsymbol{x}}}[\mathbb{E}_D[(g^{(D)} - \overline{g})^2]] + \mathbb{E}_{\underline{\boldsymbol{x}}}[(\overline{g} - f)^2] \quad (\mathbb{E}_{\underline{\boldsymbol{x}}} \text{ is linear}) \\ &= \mathbb{E}_{\underline{\boldsymbol{x}}}[\text{var}(\underline{\boldsymbol{x}})] + \mathbb{E}_{\underline{\boldsymbol{x}}}[\text{bias}(\underline{\boldsymbol{x}})^2] \\ &= \text{variance} + \text{bias} \end{aligned}$$

Interpretation of Average Hypothesis

- The average hypothesis over all training sets

$$\overline{g}(\underline{\boldsymbol{x}}) \triangleq \mathbb{E}_D[g^{(D)}(\underline{\boldsymbol{x}})]$$

  can be interpreted as the "best" hypothesis from $\mathcal{H}$ training on N samples
    - Note: $\overline{g}$ is not necessarily $\in \mathcal{H}$
- In fact it's like ensemble learning:
    - Consider all the possible data sets D with N samples
    - Learn g from each D
    - Average the hypotheses

Interpretation of Variance and Bias Terms

- The out-of-sample error can be decomposed as:

$$E_{out}(\mathcal{H}) = \text{bias}^2 + \text{variance}$$

::: columns :::: {.column width=60%}

- **Bias term**

$$\text{bias}^2 = \mathbb{E}_{\underline{\boldsymbol{x}}}[(\overline{g}(\underline{\boldsymbol{x}}) - f(\underline{\boldsymbol{x}}))^2]$$

  - Does not depend on learning as it is not a function of the data set D
  - Measures how limited $\mathcal{H}$ is
    - I.e., the ability of $\mathcal{H}$ to approximate the target with infinite training sets

- **Variance term**

$$\text{variance} = \mathbb{E}_{\underline{\boldsymbol{x}}}\mathbb{E}_D[(g^{(D)}(\underline{\boldsymbol{x}}) - \overline{g}(\underline{\boldsymbol{x}}))^2]$$

  - Measures variability of the learned hypothesis from D for any $\underline{\boldsymbol{x}}$
    - With infinite training sets, we could focus on the "best" g, which is $\overline{g}$
    - But we have only one data set D at a time, incurring a cost :::: :::: {.column width=35%}