Machine Learning

Learning Theory

Alberto Maria Metelli - Francesco Trovò

Model Evaluation Options

- Validation
- Cross-validation/LOO
- Adjustment techniques

Open questions:

- How much can we trust the value provided by Validation/Cross-validation/LOO?
- Are there other options not requiring retraining/testing on independent data?

Model Evaluation Options

- Validation
- Cross-validation/LOO
- Adjustment techniques

Open questions:

- How much can we trust the value provided by Validation/Cross-validation/LOO?
- Are there other options not requiring retraining/testing on independent data?

Model Evaluation Options

- Validation
- Cross-validation/LOO
- Adjustment techniques

Open questions:

- How much can we trust the value provided by Validation/Cross-validation/LOO?
- Are there other options not requiring retraining/testing on independent data?

Supervised Learning Framework

- Input space $\mathcal{X}$ (typically $\mathcal{X} = \mathbb{R}^M$)
- Output space $\mathcal{Y}$ (e.g., $\mathcal{Y} = \mathbb{R}$ for regression, $\mathcal{Y} = \{C_1, \dots, C_k\}$ for classification)
- (Unknown) joint probability $p(\mathbf{x}, t)$ on $\mathcal{X} \times \mathcal{Y}$
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (e.g. $\ell(y, y') = (y - y')^2$, $\ell(y, y') = \mathbb{1}\{y \neq y'\}$)
- Hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ (e.g. linear models $\mathcal{H} = \{h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} : \mathbf{w} \in \mathbb{R}^d\}$)

Supervised Learning Framework

- Input space $\mathcal{X}$ (typically $\mathcal{X} = \mathbb{R}^M$)
- Output space $\mathcal{Y}$ (e.g., $\mathcal{Y} = \mathbb{R}$ for regression, $\mathcal{Y} = \{C_1, \ldots, C_k\}$ for classification)
- (Unknown) joint probability $p(\mathbf{x}, t)$ on $\mathcal{X} \times \mathcal{Y}$
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (e.g. $\ell(y, y') = (y - y')^2$, $\ell(y, y') = \mathbb{1}\{y \neq y'\}$)
- Hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ (e.g. linear models $\mathcal{H} = \{h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} : \mathbf{w} \in \mathbb{R}^d\}$)

Supervised Learning Framework

- Input space $\mathcal{X}$ (typically $\mathcal{X} = \mathbb{R}^M$)
- Output space $\mathcal{Y}$ (e.g., $\mathcal{Y} = \mathbb{R}$ for regression, $\mathcal{Y} = \{C_1, \dots, C_k\}$ for classification)
- (Unknown) joint probability $p(\mathbf{x}, t)$ on $\mathcal{X} \times \mathcal{Y}$
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (e.g. $\ell(y, y') = (y - y')^2$, $\ell(y, y') = \mathbb{1}\{y \neq y'\}$)
- Hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ (e.g. linear models $\mathcal{H} = \{h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} : \mathbf{w} \in \mathbb{R}^d\}$)

Supervised Learning Framework

- Input space $\mathcal{X}$ (typically $\mathcal{X} = \mathbb{R}^M$)
- Output space $\mathcal{Y}$ (e.g., $\mathcal{Y} = \mathbb{R}$ for regression, $\mathcal{Y} = \{C_1, \dots, C_k\}$ for classification)
- (Unknown) joint probability $p(\mathbf{x}, t)$ on $\mathcal{X} \times \mathcal{Y}$
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (e.g. $\ell(y, y') = (y - y')^2$, $\ell(y, y') = \mathbb{1}\{y \neq y'\}$)
- Hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ (e.g. linear models $\mathcal{H} = \{h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} : \mathbf{w} \in \mathbb{R}^d\}$)

Supervised Learning Framework

- Input space $\mathcal{X}$ (typically $\mathcal{X} = \mathbb{R}^M$)
- Output space $\mathcal{Y}$ (e.g., $\mathcal{Y} = \mathbb{R}$ for regression, $\mathcal{Y} = \{C_1, \dots, C_k\}$ for classification)
- (Unknown) joint probability $p(\mathbf{x}, t)$ on $\mathcal{X} \times \mathcal{Y}$
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (e.g. $\ell(y, y') = (y - y')^2$, $\ell(y, y') = \mathbb{1}\{y \neq y'\}$)
- Hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ (e.g. linear models $\mathcal{H} = \{h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} : \mathbf{w} \in \mathbb{R}^d\}$)

Supervised Learning Framework

- **Population risk minimization**: we know $p(\mathbf{x}, t)$ and we minimize the **true loss** $\mathcal{L}$

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathcal{L}(h) = \mathbb{E}_{t,\mathbf{x}}[\ell(h(\mathbf{x}), t)]$$

- **Empirical risk minimization**: we have a **training dataset** $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_n, t_n)\}_{n=1}^N$ i.i.d. from p and we minimize the **training loss** $\hat{\mathcal{L}}$

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{\mathcal{L}}(h) = \frac{1}{N} \sum_{n=1}^{N} \ell(h(\mathbf{x}_n), t_n)$$

Supervised Learning Framework

- **Population risk minimization**: we know $p(\mathbf{x}, t)$ and we minimize the **true loss** $\mathcal{L}$

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathcal{L}(h) = \mathbb{E}_{t,\mathbf{x}}[\ell(h(\mathbf{x}), t)]$$

- **Empirical risk minimization**: we have a **training dataset** $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_n, t_n)\}_{n=1}^N$ i.i.d. from p and we minimize the **training loss** $\hat{\mathcal{L}}$

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{\mathcal{L}}(h) = \frac{1}{N} \sum_{n=1}^{N} \ell(h(\mathbf{x}_n), t_n)$$

Central Problem

- We want to evaluate the **true loss** of the **empirical risk minimizer**

$$\mathcal{L}(\hat{h}) = \mathbb{E}_{t,\mathbf{x}}[\ell(\hat{h}(\mathbf{x}), t)|\hat{h}]$$

- The true loss $\mathcal{L}(\hat{h})$ **cannot** be computed exactly without knowing $p(\mathbf{x}, t)$
- The **training loss** $\hat{\mathcal{L}}(\hat{h})$ is a **negatively biased estimator** for the true loss $\mathcal{L}(\hat{h})$!

Central Problem

- We want to evaluate the **true loss** of the **empirical risk minimizer**

$$\mathcal{L}(\hat{h}) = \mathbb{E}_{t,\mathbf{x}}[\ell(\hat{h}(\mathbf{x}), t)|\hat{h}]$$

- The true loss $\mathcal{L}(\hat{h})$ **cannot** be computed exactly without knowing $p(\mathbf{x}, t)$
- The **training loss** $\hat{\mathcal{L}}(\hat{h})$ is a **negatively biased estimator** for the true loss $\mathcal{L}(\hat{h})$!

Central Problem

- We want to evaluate the **true loss** of the **empirical risk minimizer**

$$\mathcal{L}(\hat{h}) = \mathbb{E}_{t,\mathbf{x}}[\ell(\hat{h}(\mathbf{x}), t)|\hat{h}]$$

- The true loss $\mathcal{L}(\hat{h})$ **cannot** be computed exactly without knowing $p(\mathbf{x}, t)$
- The **training loss** $\hat{\mathcal{L}}(\hat{h})$ is a **negatively biased estimator** for the true loss $\mathcal{L}(\hat{h})$!

Central Problem

- Thus, we look for a **Probably Approximately Correct (PAC)** result:

$$\mathcal{L}(\hat{h}) \leq \epsilon \left(\text{quantities that can be computed from data}\right) \quad \text{w.p.} \quad 1-\delta$$

- Two cases:
  - We have a **test dataset** $\mathcal{D}_{\text{test}}$
  - We only have the **training dataset** $\mathcal{D}_{\text{train}}$

Central Problem

- Thus, we look for a **Probably Approximately Correct (PAC)** result:

$$\mathcal{L}(\hat{h}) \leq \epsilon \left(\text{quantities that can be computed from data}\right) \quad \text{w.p.} \quad 1 - \delta$$

- Two cases:
  - We have a **test dataset** $\mathcal{D}_{\text{test}}$
  - We only have the **training dataset** $\mathcal{D}_{\text{train}}$

Central Problem

- Thus, we look for a **Probably Approximately Correct (PAC)** result:

$$\mathcal{L}(\hat{h}) \leq \epsilon\left(\text{quantities that can be computed from data}\right) \quad \text{w.p.} \quad 1-\delta$$

- Two cases:
  - We have a **test dataset** $\mathcal{D}_{\text{test}}$
  - We only have the **training dataset** $\mathcal{D}_{\text{train}}$

Central Problem

- Thus, we look for a **Probably Approximately Correct (PAC)** result:

$$\mathcal{L}(\hat{h}) \leq \epsilon \left(\text{quantities that can be computed from data}\right) \quad \text{w.p.} \quad 1 - \delta$$

- Two cases:
    - We have a **test dataset** $\mathcal{D}_{\text{test}}$
    - We only have the **training dataset** $\mathcal{D}_{\text{train}}$

Using a Test Set

Test Set

- We have a test set $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_j, t_j)\}_{j=1}^{J}$ of i.i.d. samples from p and **independent** from the training dataset $\mathcal{D}_{\text{train}}$
- For an arbitrary hypothesis $h \in \mathcal{H}$, we can evaluate the **test loss**:

$$\tilde{\mathcal{L}}(h) = \frac{1}{J} \sum_{j=1}^{J} \ell(h(\mathbf{x}_j), t_j)$$

- The empirical risk minimizer $\hat{h}$ is **independent** of $\mathcal{D}_{\text{test}}$ (while it is **dependent** on $\mathcal{D}_{\text{train}}$!):

The **test loss** $\tilde{\mathcal{L}}(\hat{h})$ is an **unbiased** estimator for the true loss $\mathcal{L}(\hat{h})$

Test Set

- We have a test set $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_j, t_j)\}_{j=1}^{J}$ of i.i.d. samples from p and **independent** from the training dataset $\mathcal{D}_{\text{train}}$
- For an arbitrary hypothesis $h \in \mathcal{H}$, we can evaluate the **test loss**:

$$\tilde{\mathcal{L}}(h) = \frac{1}{J} \sum_{j=1}^{J} \ell(h(\mathbf{x}_j), t_j)$$

- The empirical risk minimizer $\hat{h}$ is **independent** of $\mathcal{D}_{\text{test}}$ (while it is **dependent** on $\mathcal{D}_{\text{train}}$!):

The **test loss** $\tilde{\mathcal{L}}(\hat{h})$ is an **unbiased** estimator for the true loss $\mathcal{L}(\hat{h})$

Test Set

- We have a test set $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_j, t_j)\}_{j=1}^{J}$ of i.i.d. samples from p and **independent** from the training dataset $\mathcal{D}_{\text{train}}$
- For an arbitrary hypothesis $h \in \mathcal{H}$, we can evaluate the **test loss**:

$$\tilde{\mathcal{L}}(h) = \frac{1}{J} \sum_{j=1}^{J} \ell(h(\mathbf{x}_j), t_j)$$

- The empirical risk minimizer $\hat{h}$ is **independent** of $\mathcal{D}_{\text{test}}$ (while it is **dependent** on $\mathcal{D}_{\text{train}}$!):

The **test loss** $\tilde{\mathcal{L}}(\hat{h})$ is an **unbiased** estimator for the true loss $\mathcal{L}(\hat{h})$

Hoeffding Inequality Bound

Let $X_1, \ldots, X_t$ be i.i.d. random variables with support in $[0, L]$ and identical mean $\mathbb{E}[X_i] =: X$ and let $\bar{X}_t = \frac{\sum_{i=1}^{t} X_i}{t}$ be the sample mean. Then:

$$\mathbb{P}\left(X \leq \bar{X}_t + u\right) \geq 1 - e^{-\frac{2tu^2}{L^2}}$$

Meaning that we can built an upper bound with at least $1 - \delta$ confidence setting $\delta = e^{-\frac{2tu^2}{L^2}}$ The bound becomes:

$$X \leq \bar{X}_t + u = \bar{X}_t + L\sqrt{\frac{\log(1/\delta)}{2n}}$$

Test Set

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_j), t_j)\}_{j=1}^{J}$ are i.i.d. conditioned to $\hat{h}$
  - $\tilde{\mathcal{L}}(\hat{h})$ can be regarded as a sample mean of i.i.d. samples estimating the true mean $\mathcal{L}(\hat{h})$
- Under the assumption of **bounded loss** $\ell(y, y') \in [0, L]$, we can apply **Höeffding's inequality**:

$$\mathcal{L}(\hat{h}) \leq \tilde{\mathcal{L}}(\hat{h}) + L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2J}} \quad \text{w.p.} \quad 1 - \delta$$

  - The larger the test set (J), the more precise the estimate $\tilde{\mathcal{L}}(\hat{h})$ is
  - No dependence on the hypothesis space $\mathcal{H}$ complexity

Test Set

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_j), t_j)\}_{j=1}^{J}$ are i.i.d. conditioned to $\hat{h}$
  - $\tilde{\mathcal{L}}(\hat{h})$ can be regarded as a sample mean of i.i.d. samples estimating the true mean $\mathcal{L}(\hat{h})$
- Under the assumption of **bounded loss** $\ell(y, y') \in [0, L]$, we can apply **Höeffding's inequality**:

$$\mathcal{L}(\hat{h}) \leq \tilde{\mathcal{L}}(\hat{h}) + L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2J}} \quad \text{w.p.} \quad 1-\delta$$

  - The larger the test set (J), the more precise the estimate $\tilde{\mathcal{L}}(\hat{h})$ is
  - No dependence on the hypothesis space $\mathcal{H}$ complexity

Test Set

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_j), t_j)\}_{j=1}^{J}$ are i.i.d. conditioned to $\hat{h}$
  - $\tilde{\mathcal{L}}(\hat{h})$ can be regarded as a sample mean of i.i.d. samples estimating the true mean $\mathcal{L}(\hat{h})$
- Under the assumption of **bounded loss** $\ell(y, y') \in [0, L]$, we can apply **Höeffding's inequality**:

$$\mathcal{L}(\hat{h}) \leq \tilde{\mathcal{L}}(\hat{h}) + L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2J}} \quad \text{w.p.} \quad 1-\delta$$

  - The larger the test set (J), the more precise the estimate $\tilde{\mathcal{L}}(\hat{h})$ is
  - No dependence on the hypothesis space $\mathcal{H}$ complexity

Test Set

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_j), t_j)\}_{j=1}^{J}$ are i.i.d. conditioned to $\hat{h}$
  - $\tilde{\mathcal{L}}(\hat{h})$ can be regarded as a sample mean of i.i.d. samples estimating the true mean $\mathcal{L}(\hat{h})$
- Under the assumption of **bounded loss** $\ell(y, y') \in [0, L]$, we can apply **Höeffding's inequality**:

$$\mathcal{L}(\hat{h}) \leq \tilde{\mathcal{L}}(\hat{h}) + L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2J}} \quad \text{w.p.} \quad 1-\delta$$

  - The larger the test set (J), the more precise the estimate $\tilde{\mathcal{L}}(\hat{h})$ is
  - No dependence on the hypothesis space $\mathcal{H}$ complexity

Test Set

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_j), t_j)\}_{j=1}^{J}$ are i.i.d. conditioned to $\hat{h}$
  - $\tilde{\mathcal{L}}(\hat{h})$ can be regarded as a sample mean of i.i.d. samples estimating the true mean $\mathcal{L}(\hat{h})$
- Under the assumption of **bounded loss** $\ell(y, y') \in [0, L]$, we can apply **Höeffding's inequality**:

$$\mathcal{L}(\hat{h}) \leq \tilde{\mathcal{L}}(\hat{h}) + L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2J}} \quad \text{w.p.} \quad 1 - \delta$$

  - The larger the test set (J), the more precise the estimate $\tilde{\mathcal{L}}(\hat{h})$ is
  - No dependence on the hypothesis space $\mathcal{H}$ complexity

Using the Training Set

Training Set Only

- We can use **only** the same training set $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_n, t_n)\}_{n=1}^{N}$ of i.i.d. samples from p used to learn the empirical risk minimizer $\hat{h}$
- For an arbitrary hypothesis $h \in \mathcal{H}$, we can evaluate the **training loss**:

$$\hat{\mathcal{L}}(h) = \frac{1}{N} \sum_{n=1}^{N} \ell(h(\mathbf{x}_n), t_n)$$

- The empirical risk minimizer $\hat{h}$ is **dependent** on $\mathcal{D}_{\text{train}}$ ($\hat{h}$ is obtained from the very same $\mathcal{D}_{\text{train}}$!): The **training loss** $\hat{\mathcal{L}}(\hat{h})$ a **negatively biased** estimator for the true loss $\mathcal{L}(\hat{h})$, indeed:

$$\mathbb{E}[\hat{\mathcal{L}}(\hat{h})|\hat{h}] \leq \mathbb{E}[\hat{\mathcal{L}}(h^*)|\hat{h}] \leq \mathcal{L}(h^*) \leq \mathcal{L}(\hat{h})$$

Training Set Only

- We can use **only** the same training set $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_n, t_n)\}_{n=1}^N$ of i.i.d. samples from p used to learn the empirical risk minimizer $\hat{h}$
- For an arbitrary hypothesis $h \in \mathcal{H}$, we can evaluate the **training loss**:

$$\hat{\mathcal{L}}(h) = \frac{1}{N} \sum_{n=1}^{N} \ell(h(\mathbf{x}_n), t_n)$$

- The empirical risk minimizer $\hat{h}$ is **dependent** on $\mathcal{D}_{\text{train}}$ ($\hat{h}$ is obtained from the very same $\mathcal{D}_{\text{train}}$!): The **training loss** $\hat{\mathcal{L}}(\hat{h})$ a **negatively biased** estimator for the true loss $\mathcal{L}(\hat{h})$, indeed:

$$\mathbb{E}[\hat{\mathcal{L}}(\hat{h})|\hat{h}] \leq \mathbb{E}[\hat{\mathcal{L}}(h^*)|\hat{h}] \leq \mathcal{L}(h^*) \leq \mathcal{L}(\hat{h})$$

Training Set Only

- We can use **only** the same training set $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_n, t_n)\}_{n=1}^N$ of i.i.d. samples from p used to learn the empirical risk minimizer $\hat{h}$
- For an arbitrary hypothesis $h \in \mathcal{H}$, we can evaluate the **training loss**:

$$\hat{\mathcal{L}}(h) = \frac{1}{N} \sum_{n=1}^{N} \ell(h(\mathbf{x}_n), t_n)$$

- The empirical risk minimizer $\hat{h}$ is **dependent** on $\mathcal{D}_{\text{train}}$ ($\hat{h}$ is obtained from the very same $\mathcal{D}_{\text{train}}$!): The **training loss** $\hat{\mathcal{L}}(\hat{h})$ a **negatively biased** estimator for the true loss $\mathcal{L}(\hat{h})$, indeed:

$$\mathbb{E}[\hat{\mathcal{L}}(\hat{h})|\hat{h}] \leq \mathbb{E}[\hat{\mathcal{L}}(h^*)|\hat{h}] \leq \mathcal{L}(h^*) \leq \mathcal{L}(\hat{h})$$

Training Set Only

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_n), t_n)\}_{n=1}^{N}$ are **not i**.i.d. conditioned to $\hat{h}$
  - **Cannot** apply Höeffding's inequality!
- **Statistical Learning Theory** approach (Vapnik):

$$\mathcal{L}(\hat{h}) = \hat{\mathcal{L}}(\hat{h}) + \mathcal{L}(\hat{h}) - \hat{\mathcal{L}}(\hat{h}) \leq \hat{\mathcal{L}}(\hat{h}) + \sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$$

- Now, the problem becomes providing bounds on $\sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$, called **uniform bounds**.

  They will depend on:
  - The size of the training set N
  - The complexity of the hypothesis space $\mathcal{H}$

Training Set Only

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_n), t_n)\}_{n=1}^{N}$ are **not i**.i.d. conditioned to $\hat{h}$
  - **Cannot** apply Höeffding's inequality!
- **Statistical Learning Theory** approach (Vapnik):

$$\mathcal{L}(\hat{h}) = \hat{\mathcal{L}}(\hat{h}) + \mathcal{L}(\hat{h}) - \hat{\mathcal{L}}(\hat{h}) \leq \hat{\mathcal{L}}(\hat{h}) + \sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$$

- Now, the problem becomes providing bounds on $\sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$, called **uniform bounds**.

  They will depend on:
  - The size of the training set N
  - The complexity of the hypothesis space $\mathcal{H}$

Training Set Only

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_n), t_n)\}_{n=1}^N$ are **not i**.i.d. conditioned to $\hat{h}$
  - **Cannot** apply Höeffding's inequality!
- **Statistical Learning Theory** approach (Vapnik):

$$\mathcal{L}(\hat{h}) = \hat{\mathcal{L}}(\hat{h}) + \mathcal{L}(\hat{h}) - \hat{\mathcal{L}}(\hat{h}) \leq \hat{\mathcal{L}}(\hat{h}) + \sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$$

- Now, the problem becomes providing bounds on $\sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$, called **uniform bounds**.

  They will depend on:
  - The size of the training set N
  - The complexity of the hypothesis space $\mathcal{H}$

Training Set Only

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_n), t_n)\}_{n=1}^{N}$ are **not i**.i.d. conditioned to $\hat{h}$
  - **Cannot** apply Höeffding's inequality!
- **Statistical Learning Theory** approach (Vapnik):

$$\mathcal{L}(\hat{h}) = \hat{\mathcal{L}}(\hat{h}) + \mathcal{L}(\hat{h}) - \hat{\mathcal{L}}(\hat{h}) \leq \hat{\mathcal{L}}(\hat{h}) + \sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$$

- Now, the problem becomes providing bounds on $\sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$, called **uniform bounds**.

  They will depend on:
  - The size of the training set N
  - The complexity of the hypothesis space $\mathcal{H}$

Training Set Only

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_n), t_n)\}_{n=1}^{N}$ are **not i**.i.d. conditioned to $\hat{h}$
  - **Cannot** apply Höeffding's inequality!
- **Statistical Learning Theory** approach (Vapnik):

$$\mathcal{L}(\hat{h}) = \hat{\mathcal{L}}(\hat{h}) + \mathcal{L}(\hat{h}) - \hat{\mathcal{L}}(\hat{h}) \leq \hat{\mathcal{L}}(\hat{h}) + \sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$$

- Now, the problem becomes providing bounds on $\sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$, called **uniform bounds**.

  They will depend on:
  - The size of the training set N
  - The complexity of the hypothesis space $\mathcal{H}$

Training Set Only

- **Crucial observation**: all losses $\{\ell(\hat{h}(\mathbf{x}_n), t_n)\}_{n=1}^{N}$ are **not i**.i.d. conditioned to $\hat{h}$
  - **Cannot** apply Höeffding's inequality!
- **Statistical Learning Theory** approach (Vapnik):

$$\mathcal{L}(\hat{h}) = \hat{\mathcal{L}}(\hat{h}) + \mathcal{L}(\hat{h}) - \hat{\mathcal{L}}(\hat{h}) \leq \hat{\mathcal{L}}(\hat{h}) + \sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$$

- Now, the problem becomes providing bounds on $\sup_{h \in \mathcal{H}} \left| \mathcal{L}(h) - \hat{\mathcal{L}}(h) \right|$, called **uniform bounds**.

  They will depend on:
  - The size of the training set N
  - The complexity of the hypothesis space $\mathcal{H}$

Training Set Only

Uniform Bounds

We limit to **binary classification** and $\mathcal{L}$ = accuracy:

- Finite hypothesis space ($|\mathcal{H}| < +\infty$) and consistent learning ($\hat{\mathcal{L}}(\hat{h}) = 0$ always):

$$\mathcal{L}(\hat{h}) \leq \frac{\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{N} \quad \text{w.p.} \quad 1-\delta$$

- Finite hypothesis space ($|\mathcal{H}| < +\infty$) and agnostic learning ($\hat{\mathcal{L}}(\hat{h}) > 0$ possibly):

$$\mathcal{L}(\hat{h}) \leq \hat{\mathcal{L}}(\hat{h}) + \sqrt{\frac{\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{2N}} \quad \text{w.p.} \quad 1-\delta$$

Training Set Only

Uniform Bounds

We limit to **binary classification** and $\mathcal{L}$ = accuracy:

- Finite hypothesis space ($|\mathcal{H}| < +\infty$) and consistent learning ($\hat{\mathcal{L}}(\hat{h}) = 0$ always):

$$\mathcal{L}(\hat{h}) \leq \frac{\log |\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{N} \quad \text{w.p.} \quad 1-\delta$$

- Finite hypothesis space ($|\mathcal{H}| < +\infty$) and agnostic learning ($\hat{\mathcal{L}}(\hat{h}) > 0$ possibly):

$$\mathcal{L}(\hat{h}) \leq \hat{\mathcal{L}}(\hat{h}) + \sqrt{\frac{\log |\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{2N}} \quad \text{w.p.} \quad 1-\delta$$

Example

Let assume that we are using a training set composed of $N = 200$ samples: Three different hypothesis spaces:

- $\mathcal{H}_1$ with cardinality e^{22}
- $\mathcal{H}_2$ with cardinality e^{46}
- $\mathcal{H}_3$ with cardinality e^{78}

If we want a confidence of $\delta = e^{-3}$

- Assuming the three estimated models $\hat{h}_1, \hat{h}_2, \hat{h}_3$ are in the version space
- Assuming the three estimated models have error on the training set of $\hat{\mathcal{L}}(\hat{h}_1) = 0.3, \hat{\mathcal{L}}(\hat{h}_1) = 0.15, \hat{\mathcal{L}}(\hat{h}_1) = 0.1$

Training Set Only

Uniform Bounds

- Infinite hypothesis space ($|\mathcal{H}| = \infty$) and agnostic learning ($\hat{\mathcal{L}}(\hat{h}) > 0$ possibly):

$$\mathcal{L}(\hat{h}) \leq \hat{\mathcal{L}}(\hat{h}) + \sqrt{\frac{\mathrm{VC}(\mathcal{H}) \log\left(\frac{2eN}{\mathrm{VC}(\mathcal{H})}\right) + \log\left(\frac{4}{\delta}\right)}{N}} \quad \text{w.p.} \quad 1 - \delta$$

Example

Let assume that we are using a training set composed of $N = 2400$ samples: Three different hypothesis spaces:

- $\mathcal{H}_1$ with cardinality $+\infty$ and $VC(\mathcal{H}_1) = e^2$
- $\mathcal{H}_2$ with cardinality $+\infty$ and $VC(\mathcal{H}_2) = e^4$
- $\mathcal{H}_3$ with cardinality $+\infty$ and $VC(\mathcal{H}_3) = e^6$

If we want a confidence of $\delta = e^{-3}/4$

- Assuming the three estimated models $\hat{h}_1, \hat{h}_2, \hat{h}_3$ are in the version space
- Assuming the three estimated models have error on the training set of $\hat{\mathcal{L}}(\hat{h}_1) = 0.3, \hat{\mathcal{L}}(\hat{h}_1) = 0.15, \hat{\mathcal{L}}(\hat{h}_1) = 0.1$
- Hint: use the fact that $9 \approx e^2, 81 \approx e^4, 400 \approx e^6$