Machine Learning

Anders Jonsson & Vicenç Gómez

Master in Intelligence Interactive Systems
2020-21

Lecture 3
Bias-Variance Tradeoff and Overfitting

Content

Content

True loss vs. training loss

- True loss or risk $L_{\mathcal{D},f}(h)$ measures the mistakes of h on the entire domain set $\mathcal{X}$ (with distribution $\mathcal{D}$ and labelling function f)
- Training loss or empirical risk $L_S(h)$ measures the mistakes of h on the training set $S = ((x_1, y_1), \dots, (x_m, y_m))$
- Want h with small $L_{\mathcal{D},f}(h)$, but can only measure $L_S(h)$

$$L_{\mathcal{D},f}(h) = L_S(h) + (L_{\mathcal{D},f}(h) - L_S(h))$$

- Generalization: minimize $L_{\mathcal{D},f}(h) - L_S(h)$

Generalization properties

- How well does $L_S(h)$ approximate $L_{\mathcal{D},f}(h)$?
- Hoeffding's inequality for a single, fixed hypothesis h:

$$\mathbb{P}\left[|L_S(h) - L_{\mathcal{D},f}(h)| > \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

- Hypothesis h_S that minimizes the empirical risk:

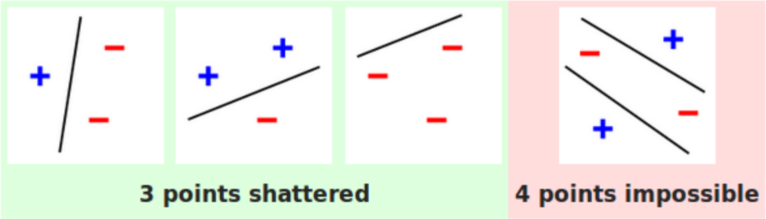
$$\mathbb{P}\left[|L_S(h_S) - L_{\mathcal{D},f}(h_S)| > \epsilon\right] \leq 2|\mathcal{H}|e^{-2m\epsilon^2}$$

VC dimension

- Problem: $\mathcal{H}$ is often an infinite set $\Rightarrow |\mathcal{H}|$ is unbounded
- Vapnik-Chervonenkis (VC) dimension D_{VC}: effective size of $\mathcal{H}$
- Hypothesis h_S that minimizes the empirical risk:

$$\mathbb{P}\left[|L_S(h_S) - L_{\mathcal{D},f}(h_S)| > \epsilon\right] \leq 2D_{VC}e^{-2m\epsilon^2}$$

VC dimension

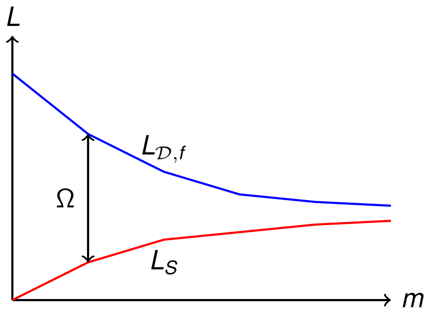


Model complexity

- For linear models, $|\mathcal{H}| = \infty$ but $D_{VC} = d + 1$!
- Model complexity: number of model parameters (e.g. weights)
- D_{VC} is often proportional to the model complexity
- A more complex model is less likely to generalize well!
- Alternative formulation of Hoeffding's inequality:

$$L_{\mathcal{D},f}(h_S) \leq L_S(h_S) + \Omega(m, D_{VC})$$

Learning curves



- The training loss usually increases as a function of m
- The true loss usually decreases as a function of m
- Equivalently, $\Omega(m, D_{VC})$ decreases as a function of m

No Free Lunch theorem

- Let $\mathcal{A}$ be any binary classification algorithm on domain set $\mathcal{X}$
- Let $m \leq |\mathcal{X}|/2$ be the size of the training set S

No Free Lunch theorem

- Let $\mathcal{A}$ be any binary classification algorithm on domain set $\mathcal{X}$
- Let $m \leq |\mathcal{X}|/2$ be the size of the training set S

Theorem

There exist $\mathcal{D}$ and f such that with probability at least $1/7$ on the choice of S, it holds that $L_{\mathcal{D},f}(\mathcal{A}(S)) \geq 1/8$

No Free Lunch theorem

- Let $\mathcal{A}$ be any binary classification algorithm on domain set $\mathcal{X}$
- Let $m \leq |\mathcal{X}|/2$ be the size of the training set S

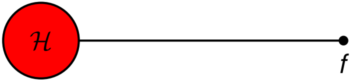
Theorem

There exist $\mathcal{D}$ and f such that with probability at least $1/7$ on the choice of S, it holds that $L_{\mathcal{D},f}(\mathcal{A}(S)) \geq 1/8$

No algorithm does well on all learning problems!

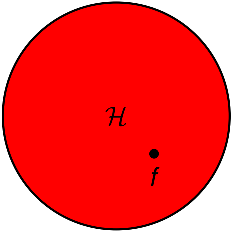
Content

Bias



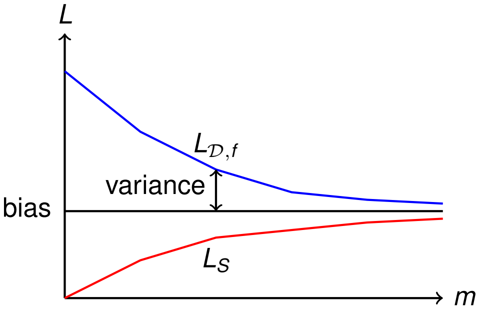
- It is essential to restrict the class $\mathcal{H}$ of hypothesis functions
- However, too much restriction prevents us from approximating f!
- Bias: how “far” the labelling function f is from the class $\mathcal{H}$

Variance



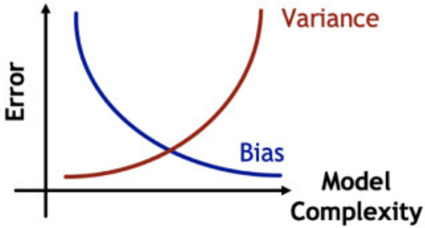
- The larger the hypothesis class, the more likely it is to include f
- However, this makes it more difficult to zoom in on the correct f
- Variance: how far the ERM hypothesis h_S is from f on average

Learning curves



- Bias determines the theoretical limit of $L_{\mathcal{D},f}$
- Variance determines how far $L_{\mathcal{D},f}$ is from this limit
- Variance decreases as a function of m

Bias-variance tradeoff



- Less complex model $\Rightarrow$ more bias
- More complex model $\Rightarrow$ more variance
- Tradeoff: impossible to achieve 0 bias and 0 variance

Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$\begin{cases} \mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} = \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x) - f(x))^2\}\} \end{cases}$$

Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$
\begin{cases}
\mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} = \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x) - f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x) - f(x))^2\}\}
\end{cases}
$$

Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$\begin{cases} \mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} = \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x) - f(x))^2\}\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x) - f(x))^2\}\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x) - \overline{h}(x) + \overline{h}(x) - f(x))^2\}\} \end{cases}$$

Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$
\begin{cases}
\mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} = \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x) - f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x) - f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x) - \overline{h}(x) + \overline{h}(x) - f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x) - \overline{h}(x))^2 + (\overline{h}(x) - f(x))^2
\end{cases}
$$

Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$\begin{cases} \mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} = \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x) - f(x))^2\}\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x) - f(x))^2\}\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x) - \overline{h}(x) + \overline{h}(x) - f(x))^2\}\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x) - \overline{h}(x))^2 + (\overline{h}(x) - f(x))^2 \\ \qquad\qquad + 2(h_S(x) - \overline{h}(x))(\overline{h}(x) - f(x))\}\} \end{cases}$$

Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$
\begin{cases}
\mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} = \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x)-f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\overline{h}(x)+\overline{h}(x)-f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\overline{h}(x))^2+(\overline{h}(x)-f(x))^2 \\
\qquad\qquad + 2(h_S(x)-\overline{h}(x))(\overline{h}(x)-f(x))\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\overline{h}(x))^2\}+(\overline{h}(x)-f(x))^2+0\}
\end{cases}
$$

Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$
\begin{cases}
\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} &= \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x)-f(x))^2\}\} \\
&= \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-f(x))^2\}\} \\
&= \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\bar{h}(x)+\bar{h}(x)-f(x))^2\}\} \\
&= \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\bar{h}(x))^2+(\bar{h}(x)-f(x))^2 \\
&\qquad\qquad +2(h_S(x)-\bar{h}(x))(\bar{h}(x)-f(x))\}\} \\
&= \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\bar{h}(x))^2\}+(\bar{h}(x)-f(x))^2+0\} \\
&= \mathbb{E}_{x\sim\mathcal{D}}\{\quad \mathit{variance}(x) \quad + \quad \mathit{bias}(x) \quad\}
\end{aligned}
\end{cases}
$$

Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$
\begin{cases}
\mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} = \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x)-f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\overline{h}(x)+\overline{h}(x)-f(x))^2\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\overline{h}(x))^2+(\overline{h}(x)-f(x))^2 \\
\qquad\qquad + 2(h_S(x)-\overline{h}(x))(\overline{h}(x)-f(x))\}\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\overline{h}(x))^2\}+(\overline{h}(x)-f(x))^2+0\} \\
\qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\quad \mathit{variance}(x) \quad + \quad \mathit{bias}(x) \quad\} \\
\qquad = \mathit{variance}+\mathit{bias}
\end{cases}
$$

Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$\begin{cases} \mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} = \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x)-f(x))^2\}\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-f(x))^2\}\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\bar{h}(x)+\bar{h}(x)-f(x))^2\}\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\bar{h}(x))^2+(\bar{h}(x)-f(x))^2 \\ \qquad\qquad +2(h_S(x)-\bar{h}(x))(\bar{h}(x)-f(x))\}\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\bar{h}(x))^2\}+(\bar{h}(x)-f(x))^2+0\} \\ \qquad = \mathbb{E}_{x\sim\mathcal{D}}\{ \quad \mathit{variance}(x) \quad + \quad \mathit{bias}(x) \quad \} \\ \qquad = \mathit{variance}+\mathit{bias} \end{cases}$$

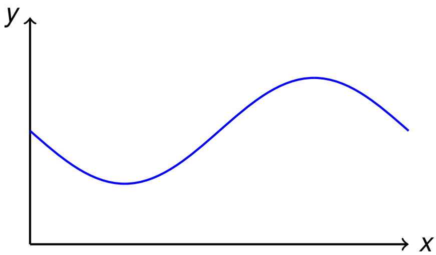
Bias-variance characterization

Regression task, squared error, ERM hypothesis h_S:

$$
\begin{cases}
\mathbb{E}_{S\sim\mathcal{D},f}\{L_{\mathcal{D},f}(h_S)\} = \mathbb{E}_{S\sim\mathcal{D},f}\{\mathbb{E}_{x\sim\mathcal{D}}\{(h_S(x)-f(x))^2\}\} \\
\quad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-f(x))^2\}\} \\
\quad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\bar{h}(x)+\bar{h}(x)-f(x))^2\}\} \\
\quad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\bar{h}(x))^2+(\bar{h}(x)-f(x))^2 \\
\qquad\qquad +2(h_S(x)-\bar{h}(x))(\bar{h}(x)-f(x))\}\} \\
\quad = \mathbb{E}_{x\sim\mathcal{D}}\{\mathbb{E}_{S\sim\mathcal{D},f}\{(h_S(x)-\bar{h}(x))^2\}+(\bar{h}(x)-f(x))^2+0\} \\
\quad = \mathbb{E}_{x\sim\mathcal{D}}\{\quad \mathit{variance}(x) \quad + \quad \mathit{bias}(x) \quad\} \\
\quad = \mathit{variance} + \mathit{bias}
\end{cases}
$$

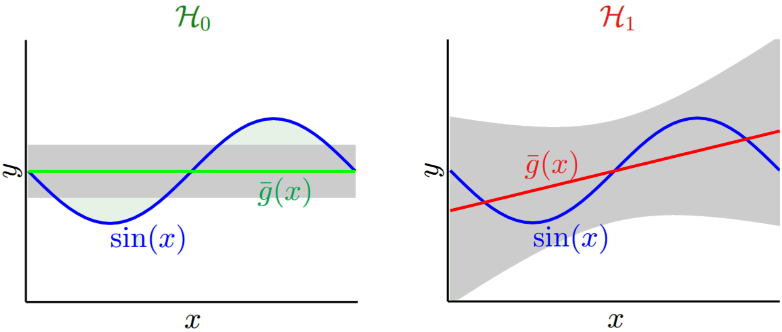
$\bar{h}(x) = \mathbb{E}_{S\sim\mathcal{D},f}\{h_S(x)\}$: average ERM hypothesis on input x

Example



- Assume that f is a sine curve
- $\mathcal{H}_0$: constant hypotheses
- $\mathcal{H}_1$: linear hypotheses
- $m = 2$: only sample 2 data points
- Which hypothesis class is better?

Comparison

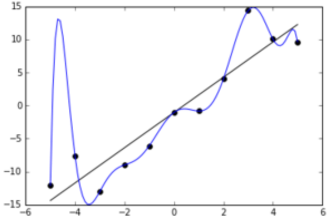


bias $= \mathbf{0.50}$ var $= \mathbf{0.25}$ bias $= \mathbf{0.21}$ var $= \mathbf{1.69}$

$\overline{g}(x) = \overline{h}(x)$: average ERM hypothesis on input x

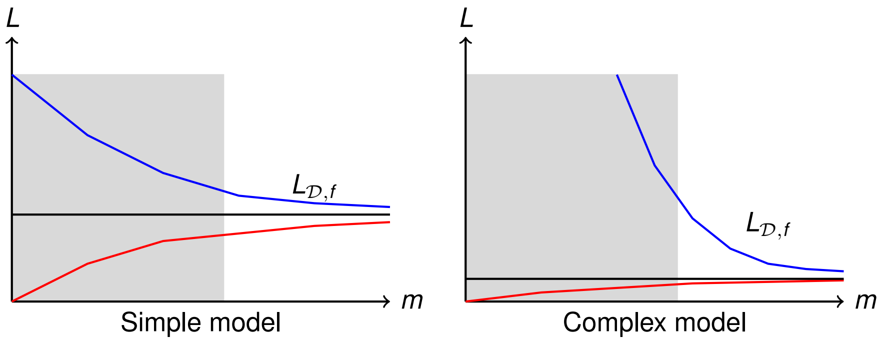
Content

Overfitting



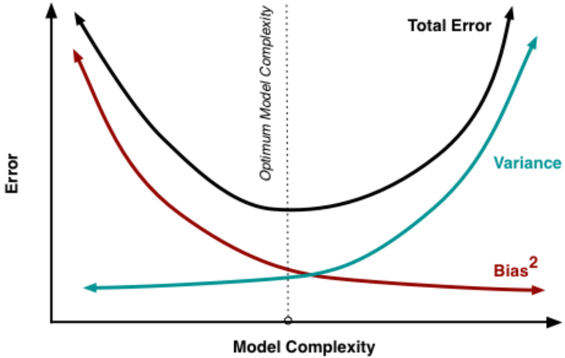
- We can often make the training loss smaller using a more complex model
- Overfitting: sacrifice true loss for smaller training loss

Learning curves



- Higher model complexity $\Rightarrow$ smaller training loss $L_S(h)$
- Poor generalization properties $\Rightarrow$ larger true loss $L_{\mathcal{D},f}(h)$

Overfitting and bias-variance tradeoff



- There exists a theoretical optimum model complexity
- Increasing the model complexity more causes the loss to blow up
- In practice: better to start with simpler models!

Regularization

- Technique that helps overcome the problem of overfitting
- Linear models: introduce constraints on the weight vector w
- Constrained optimization:

$$\min L_S(w) \quad \text{s.t.} \sum_{i=0}^{d} w_i^2 \leq C$$

- Difficult (NP-hard) to optimize

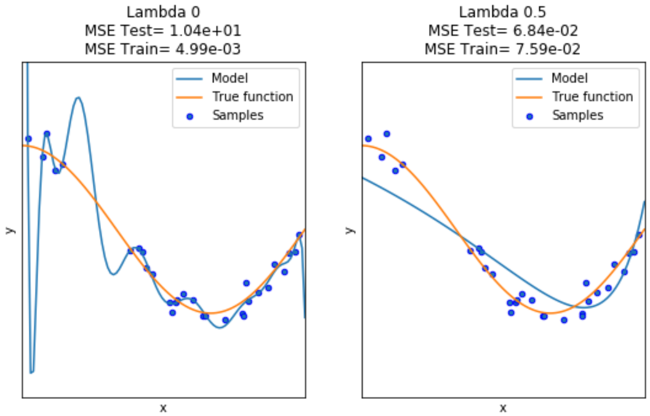
Regularization

- Alternative definition: add extra term to loss function:

$$L_{aug}(w) = L_S(w) + \frac{\lambda}{m} w^\top w$$

- $\sum w_i^2$: L2-norm, weighted decay
- $\sum |w_i|$: L1-norm, sparsity
- Difficulty: no analytical way to select λ
- Linear regression: $w_{reg} = (X^\top X + \lambda I)^{-1} X^\top y$

Regularization



Validation

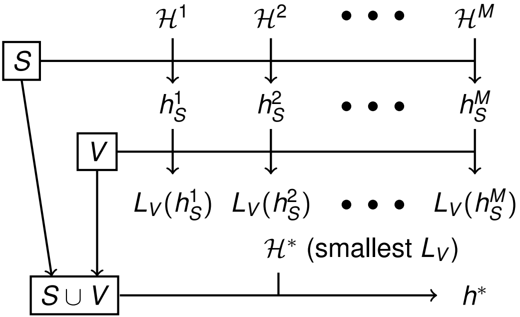
- Alternative to overcome overfitting
- Used for model selection: learning algorithm, non-linear transform, regularizer, parameters, etc.
- Due to overfitting, selecting by $L_S(h)$ is not always a good idea!
- Validation: approximate $L_{\mathcal{D},f}(h)$ better (but still optimistic!)

Validation

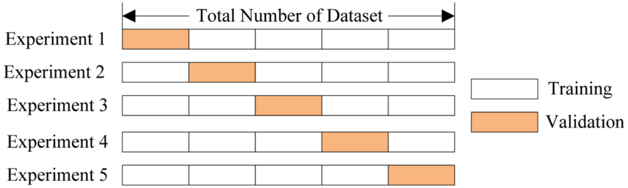
- In addition to S, assume validation set $V = ((x_1, y_1), \ldots, (x_n, y_n))$
- Also assume that V is sampled independently of S
- Validation loss $L_V(h)$ is a much better estimate of $L_{\mathcal{D},f}(h)$!
- In practice: divide dataset into training set and validation set

Model selection

- Train M alternative models on training set S
- Compute validation error $L_V(h_S)$ on each resulting hypothesis
- Select model with smallest validation error, retrain on entire $S \cup V$



Cross-validation



- Partition S into k subsets $S_1, \ldots, S_k$, each of size m/k
- In each experiment, train on $S \setminus S_i$ and validate on S_i
- Cross-validation loss is the average across experiments:

$$L_{cv}(\theta) = \frac{1}{k}\sum_{i=1}^{k} L_{S_i}(h_i)$$

- In practice: $k = 5$ or $k = 10$ are usually good choices

Cross-validation