Machine Learning

Bert Huang
Data Intensive Studies Center
Tufts University

Outline

- Most common learning setting: supervised learning
- Empirical risk minimization
- Generalization and generalization error

Supervised Learning

- Draw data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ from distribution $\mathbb{D}$
- Algorithm A learns hypothesis $h \in H$ from set H of possible hypotheses $A(D) = h$
- We measure the quality of h as the expected **loss**: $\underset{(x,y)\in\mathbb{D}}{E}[\ell(y, h(x))]$
    - This quantity is known as the **risk**
    - E.g., loss could be the Hamming loss $\ell_{\text{Hamming}}(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{otherwise} \end{cases}$

Supervised Learning

- Draw data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ from distribution $\mathbb{D}$
- Algorithm A learns hypothesis $h \in H$ from set H of possible hypotheses $A(D) = h$

$$\underset{(x,y) \in \mathbb{D}}{E} [\ell(y, h(x))]$$

Risk

$$\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i)$$

Empirical Risk

Examples

- Maximum likelihood estimation. Loss = negative log likelihood
- Support vector machines. Loss = hinge loss
- Neural networks. Loss = any differentiable loss

Empirical Risk Minimization

- Algorithm A solves

$$\min_{h \in \mathcal{H}} \quad \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i)) \qquad := \min_{h \in \mathcal{H}} \quad \hat{R}(h)$$

- **Generalization error** compares **risk** to **empirical risk**

$$E_{x \sim \mathbb{D}} \left[\ell(y, h(x) \right] - \sum_{i=1}^{n} \ell(y_i, h(x_i)) \quad := \quad R(h) - \hat{R}(h)$$

Example Generalization Error Bound

true risk

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\mathrm{VC}(H)\left(\log \frac{2n}{\mathrm{VC}(H)} + 1\right) - \log \frac{\delta}{4}}{n}}$$

empirical risk

failure probability

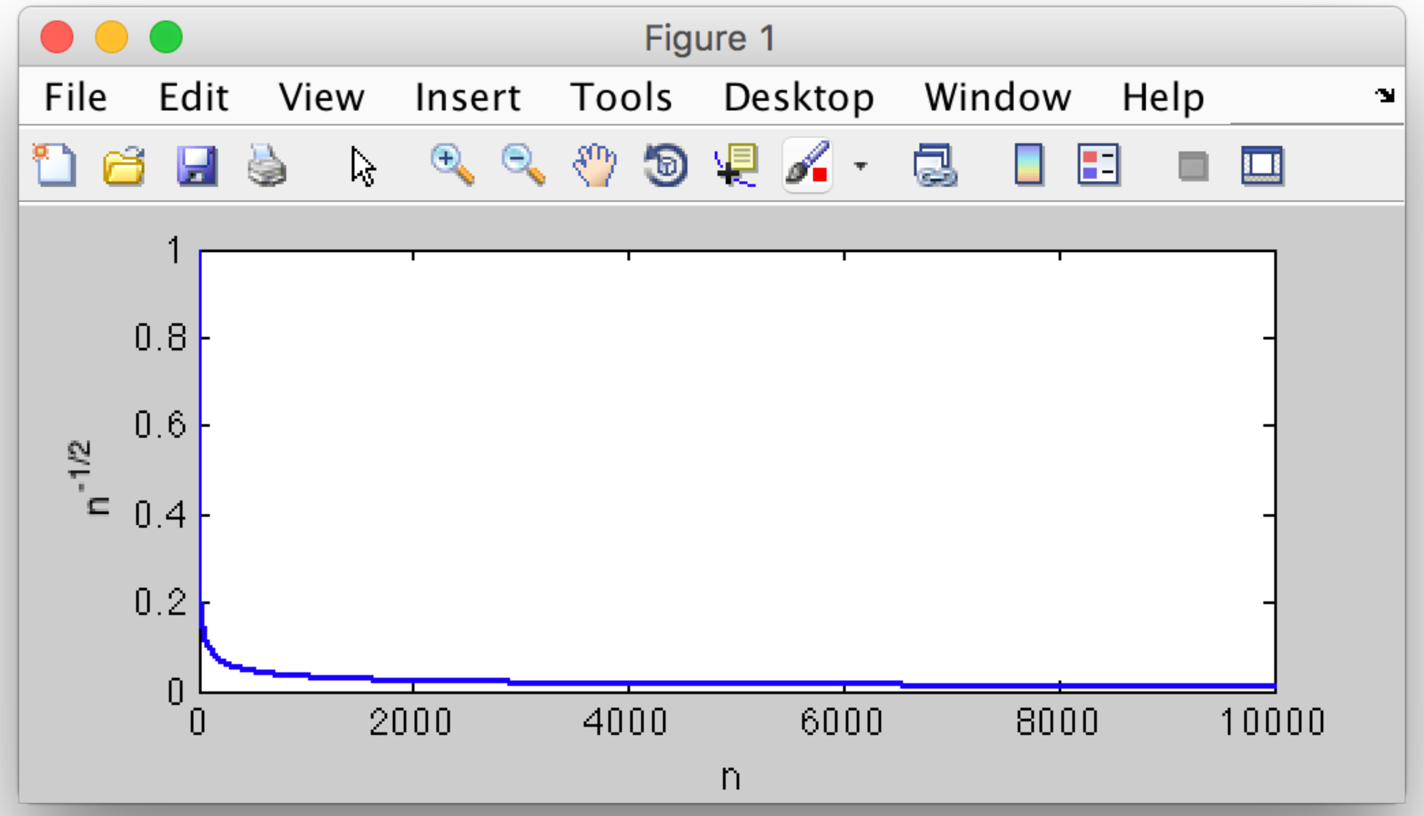
size of training data

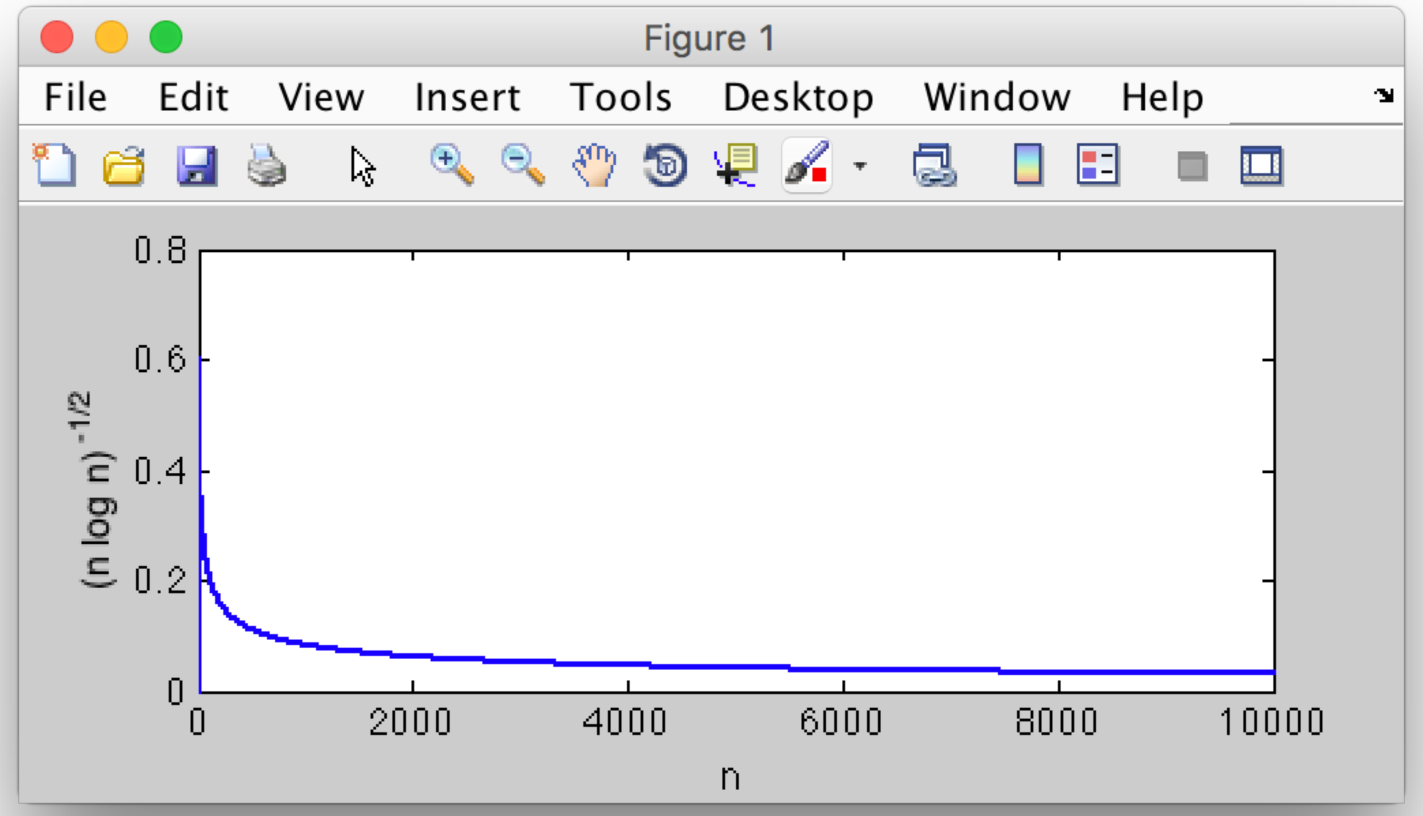
Vapnik-Chervonenkis dimension (model complexity)

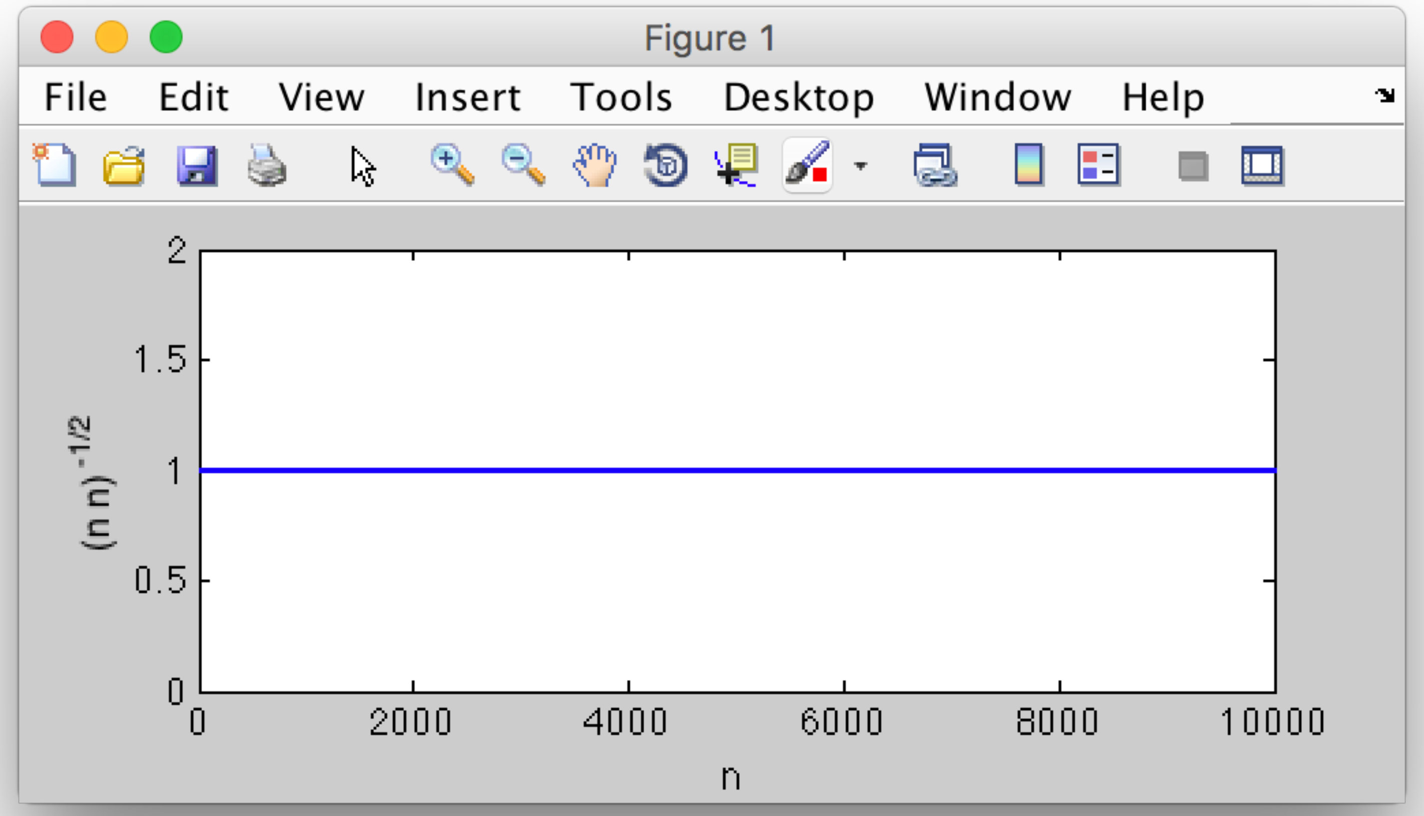
$$R(h) \leq \hat{R}(h) + \sqrt{\frac{VC(H)\left(\log \frac{2n}{VC(H)} + 1\right) - \log \frac{\delta}{4}}{n}} \approx \sqrt{\frac{\text{complexity(H)}}{n}}$$

if complexity is fixed

if complexity is O(n)







Takeaway Points

- Supervised learning trains from labeled examples
- Empirical risk minimization finds **hypothesis** in **hypothesis class** that scores lowest empirical risk
- But usually we care about **true risk**
- Difference between **true risk** and **empirical risk** is the **generalization error**
- **Generaliztion error** shrinks with more data (and simpler models)