MAC 0460 / 5832
Introduction to Machine Learning

15 – Validation

IME/USP (09/06/2021)

We have seen that E_{in} is computed over the training set

E_{in} is a (super)optimistic estimate of E_{out}

$$E_{out} = E_{in} + \text{generalization\_error}$$

Minimizing only E_{in} will lead to overfitting

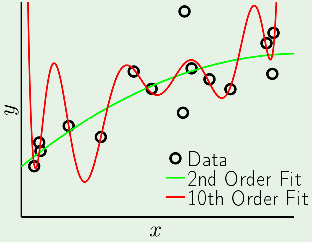
Overfitting

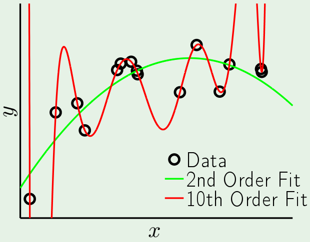
"fitting the data more than is warranted"

Fitting to noisy data (stochastic noise)

Model noise ?? (deterministic noise, according to prof. Abu-Mostafa)

Two fits for each target





Noisy low-order target

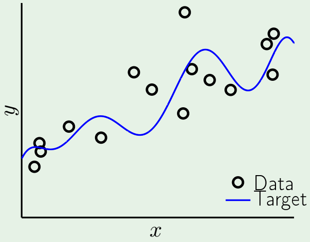
	2nd Order	10th Order
E_{in}	0.050	0.034
E_{out}	0.127	**9.00**

Noiseless high-order target

	2nd Order	10th Order
E_{in}	0.029	10^{-5}
E_{out}	0.120	**7680**

A detailed experiment

Impact of **noise level** and **target complexity**



$$y \;=\; f(x) + \underbrace{\epsilon(x)}_{\sigma^2} = \underbrace{\sum_{q=0}^{Q_f} \alpha_q \, x^q}_{\text{normalized}} + \epsilon(x)$$

noise level: σ^2

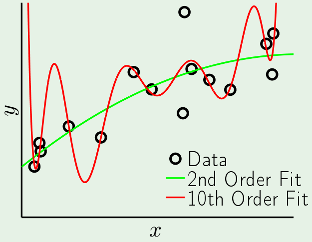
target complexity: Q_f

data set size: N

The overfit measure

We fit the data set $(x_1, y_1), \cdots, (x_N, y_N)$ using our two models:

$\mathcal{H}_2$: 2nd-order polynomials $\qquad$ $\mathcal{H}_{10}$: 10th-order polynomials

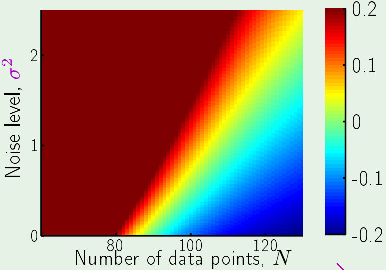


Compare out-of-sample errors of

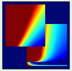
$$g_2 \in \mathcal{H}_2 \quad \text{and} \quad g_{10} \in \mathcal{H}_{10}$$

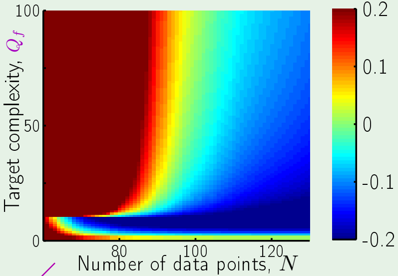
overfit measure: $E_{\text{out}}(g_{10}) - E_{\text{out}}(g_2)$

The results



Impact of σ^2





Impact of Q_f

Overfitting

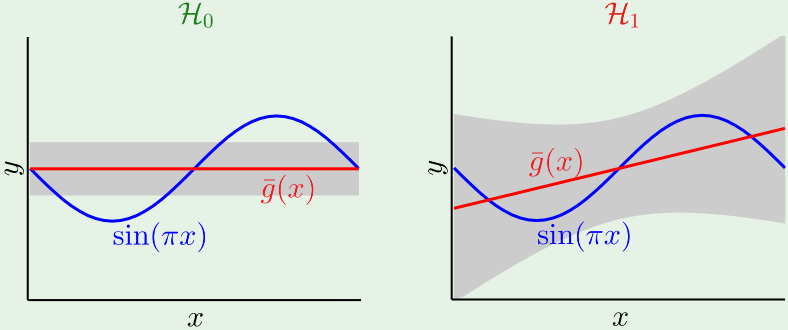
Stochastic noise – Too much effort to reduce E_{in} leads to fitting to the noise in the data

Deterministic noise – This is very subtle (my opinion)

A good example to illustrate this is the senoid example

There is "a right complexity $\mathcal{H}$" for each target and amount N of training data.

and the winner is ...



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

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

Validation versus regularization

In one form or another, $E_{\text{out}}(h) \;=\; E_{\text{in}}(h) \;+\;$ overfit penalty

Regularization:

$$E_{\text{out}}(h) \;=\; E_{\text{in}}(h) + \underbrace{\text{overfit penalty}}_{\text{regularization estimates this quantity}}$$

Validation:

$$\underbrace{E_{\text{out}}(h)}_{\text{validation estimates this quantity}} = \; E_{\text{in}}(h) \;+\; \text{overfit penalty}$$

How to compute a better estimate of E_{out} ?

Validation error

Partition the existing dataset into two subsets:

$$D = D_{train} \cup D_{val}$$

D_{train} is used for training and for computing E_{in}

D_{val} is used to compute E_{val}

E_{val} is an unbiased estimate of E_{out}

$$E[E_{val}(g)] = E_{out}(g)$$

Let $K = |D_{val}|$. Then

$$E_{val}(g) = E_{out}(g) \pm O(\frac{1}{K})$$

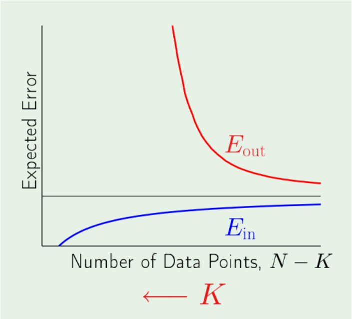
Chapter 5 of the book “Machine Learning”, by Tom Mitchell shows how to compute a confidence interval for E_{out}

That is, an interval $E_{val} \pm \Delta$ that contains E_{out} with high probability ($\Delta = O(\frac{1}{\sqrt{K}})$)

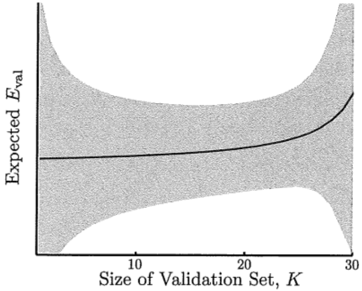
Large $K = |D_{val}|$ yields a good estimate of E_{out} (small variance) but at the same time, with less training data, E_{out} tend to be larger than when using the whole dataset

$K = |D_{val}|$

E_{in}, E_{out} $\times$ training samples



Mean and variance of E_{val}



$$E_{out} \leq E_{val} + O\left(\frac{1}{\sqrt{K}}\right)$$

Sample set D is finite. There is a trade-off.

- large $|D_{val}|$ $\rightsquigarrow$ small $|D_{train}|$ (small amount of training data) $\rightsquigarrow$ large E_{out} and $E_{val} \approx E_{out}$
- small $|D_{val}|$ $\rightsquigarrow$ large $|D_{train}|$ $\rightsquigarrow$ it is possible that E_{out} is small, but E_{val} has large variance

We could train a hypothesis g on D and report E_{val} of the hypothesis g^- trained on D_{train}
(but $E_{val}(g^-)$ is not an estimate of $E_{out}(g)$)

In practice, $K = N/5$ is a good choice

The dilemma about K

The following chain of reasoning:

$$\underbrace{E_{\text{out}}(g) \approx E_{\text{out}}(g^-)}_{\text{(small } K\text{)}} \approx \underbrace{E_{\text{val}}(g^-)}_{\text{(large } K\text{)}}$$

highlights the dilemma in selecting K:

Can we have K both small and large? ☺

Leave one out

$N - 1$ points for training, and 1 point for validation!

$$\mathcal{D}_n = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n-1}, y_{n-1}), \cancel{(\mathbf{x}_n, y_n)}, (\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_N, y_N)$$

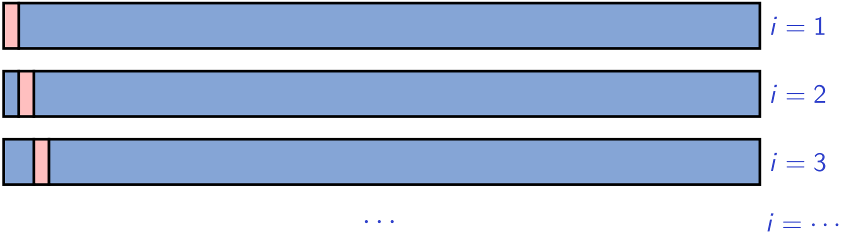
Final hypothesis learned from $\mathcal{D}_n$ is g_n^-

$$\mathbf{e}_n = E_{\text{val}}(g_n^-) = \mathbf{e}\left(g_n^-(\mathbf{x}_n), y_n\right)$$

cross validation error: $$E_{\text{cv}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{e}_n$$

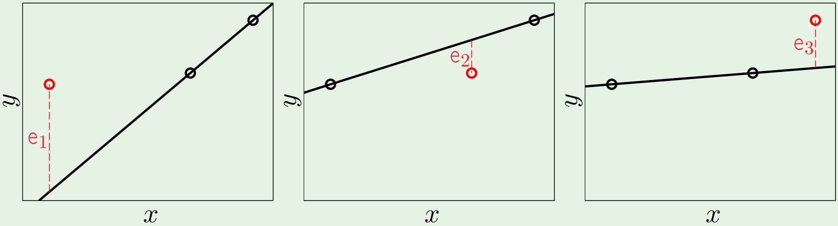
Training is repeated $N = |D|$ times

At training round i, $D_{train}^{(i)} = D \setminus \{\mathbf{x}^{(i)}\}$ and $D_{val}^{(i)} = \{\mathbf{x}^{(i)}\}$



Cross-validation error: $E_{cv} = \dfrac{1}{N} \sum_{i=1}^{N} E_{val}^{(i)}$

Illustration of cross validation



$$E_{\text{cv}} = \frac{1}{3}\left(\mathrm{e}_1 + \mathrm{e}_2 + \mathrm{e}_3 \right)$$

g^-: hypothesis trained on $N-1$ examples

We can show that E_{cv} is an unbiased estimator of $E[E_{out}(g^-)]$

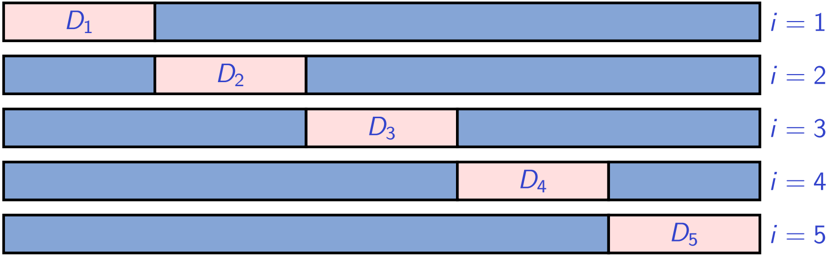
$$E_{\text{out}}(g) \underset{\text{(small } K)}{\approx} E_{\text{out}}(g^-) \underset{\text{(large } K)}{\approx} E_{\text{val}}(g^-)$$

Divide D into k parts $D_1, D_2, \ldots, D_k$ of approximately equal sizes

Repeat the training k times

At training round i, $D_{train}^{(i)} = D \setminus D_i$ and $D_{val}^{(i)} = D_i$

Example with $k = 5$ (five folds):



Cross-validation error: $E_{cv} = \frac{1}{k} \sum_{i=1}^{k} E_{val}^{(i)}$

Is E_{cv} a good estimator of E_{out} ?

k models $\rightsquigarrow$ k values for E_{out} $\rightsquigarrow$ average $\overline{E}_{out}$

It can be demonstrated that E_{cv} is an unbiased estimator of $\overline{E}_{out}$

The variance of E_{cv} can not be easily computed

Empirically, it has been observed that E_{cv} is a good estimator of $\overline{E}_{out}$

Further reading:

- Dietterich, Thomas G., Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput., 10(7), p.1895-1923, 1998
- Chapter 7 of "The Elements of Statistical learning", by Hastie *et al.*

Can we do better ?

There are a lot of discussions in literature about how to get unbiased estimates of E_{out} with small variance, etc

Further reading:

- Dietterich, Thomas G., Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput., 10(7), p.1895-1923, 1998
- Chapter 7 of "The Elements of Statistical learning", by Hastie *et al.*

- for the holdout method ($D = D_{train} \cup D_{val}$) a common proportion is 70%$\sim$80% for training and 20%$\sim$30% for validation

- for k-fold cross-validation, usual value of k is 5 or 10

- leave-one-out is just k-fold cross-validation, with $k = |D|$

  Requires $|D|$ training rounds $\rightsquigarrow$ computationally intense

  For small $|D|$ it could be the best option

- holdout should be sufficient if both D_{train} and D_{val} are large and representative enough of the true distribution
  (this usually is not the case in practice)

- k-fold cross-validation is largely used for model selection

Model selection

We already know

- how to train an ML algorithms (get hypothesis)
- how to evaluate a hypothesis (E_{val}, E_{cv}, other metrics)

How do we choose ONE ?

Model selection

Here we call as **model** any specific hypothesis g in the hypothesis space $\mathcal{H}$ that resulted after training

For example, after doing logistic regression we have a weight vector $\mathbf{w}$ which characterizes the learned classifier (the model)

As we have seen, we can compute $E_{val}(g)$ over a validation set

Model selection

Suppose you have two models, g_1 and g_2, as well as $E_{val}(g_1)$ and $E_{val}(g_2)$

If $E_{val}(g_1) < E_{val}(g_2)$, would you choose g_1 without hesitation ?

What if $E_{val}(g_1) = E_{val}(g_2)$?

Model selection in practice

Based on validation or cross-validation error

Usually the one with smallest validation error is chosen

Statistical tests can be applied to test whether $E_{val}(g_1) = E_{val}(g_2)$ or not

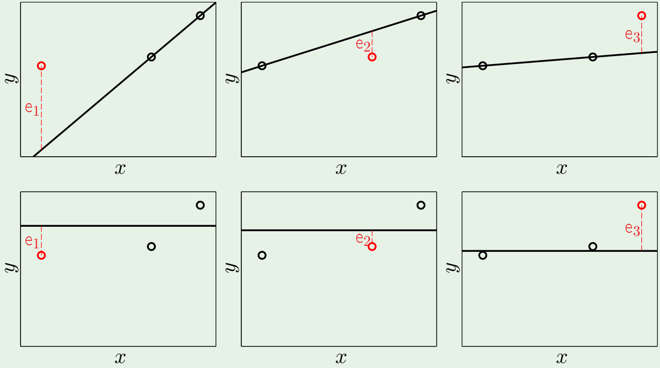
Holdout error: Hypothesis test (see for instance Chapter 5 of the book "Machine Learning", by Tom Mitchell)

Cross-validation error: paired t-test (see Dietterich, Thomas G., Approximate statistical tests for comparing supervised classification learning algorithms)

Model selection using CV

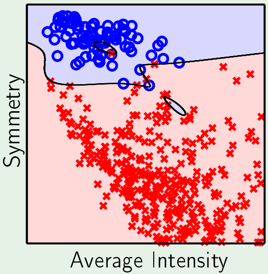
Linear:

Constant:



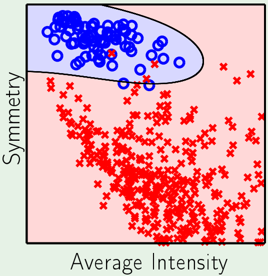
The result

without validation



$E_{\text{in}} = 0\% \quad E_{\text{out}} = 2.5\%$

with validation



$E_{\text{in}} = 0.8\% \quad E_{\text{out}} = 1.5\%$

If we use E_{val} for model selection, E_{val} no longer in an unbiased estimate of E_{out}

We can see model selection also as a kind of training (given a set of hypothesis M, we will choose ONE with smallest E_{val})

In this sense, we can apply the Hoeffding inequality in a similar way

The larger K, the smaller the bound (i.e., $E_{val} \approx E_{out}$)

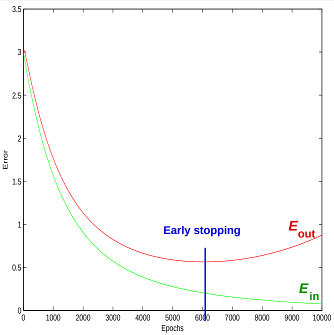
Why 'validation'

$\mathcal{D}_{\text{val}}$ is used to make learning choices

If an estimate of E_{out} affects learning:

the set is no longer a **test** set!

It becomes a validation set



Discussion: with early stopping, we can think that we have a large number of choices (any iteration number is a hypothesis). But in terms of VC dimension, it is just one parameter. This is why methods such as early stopping work in practice.

The process of model selection and performance evaluation

1. Divide the dataset D into $D_{train+val}$ and D_{test}

2. Isolate D_{test} (put it under quarantine ...)

3. Use $D_{train+val}$ for training and choosing a model
   Depending on the selection technique different partitions of $D_{train+val}$ will be used for training and for error estimation

4. the chosen model can be retrained using the whole dataset $D_{train+val}$
   (advantage is that we have more training data)

5. Having the final model, compute E_{test} over D_{test}
   E_{test} would be a less biased estimator of E_{out} than E_{val} and E_{cv} (since these last two would be an optimistic estimate)

In many situations, we just want to choose the best model

We do not need to have an estimate of E_{out}

In such situation, it is common to not consider D_{test}
(the whole set D is used for training and model choice only)

Obviously, the validation error of the chosen model is biased
(because we chose the model with minimum E_{val} value)

The same observation holds with respect to any of the metrics computed on D_{val}, after a model is chosen based on its E_{val} value

References

Section 4.3 of "Learning from data" by Mostafa *et al.*

There are references at:

https://stats.stackexchange.com/questions/18348/differences-between-cross-validation-and-bootstrapping-to-estimate-the-predictio

Chapter 7 of "The Elements of Statistical learning", by Hastie *et al.*

Chapter 5 of "Machine Learning" , by Tom Mitchell

Steven L. Salzberg, On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach, Data Min. Knowl. Discov., 3, pp.317-328, 1993

Dietterich, Thomas G., Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput., 10(7), p.1895-1923, 1998.