Introduction to Machine Learning

Evaluation Test Error

× × 0 × × ×



Learning goals

- Understand the definition of test error
- Understand that test error is more reliable than train error
- Bias-Variance analysis of holdout splitting

TEST ERROR AND HOLD-OUT SPLITTING

• Simulate prediction on unseen data, to avoid optimistic bias:

$$ho(\mathbf{y}_{ ext{test}}, oldsymbol{\mathcal{F}}_{ ext{test}})$$
 where $oldsymbol{\mathcal{F}}_{ ext{test}} = egin{bmatrix} \hat{f}_{\mathcal{D}_{ ext{train}}}(\mathbf{x}_{ ext{test}}^{(1)}) \ \dots \ \hat{f}_{\mathcal{D}_{ ext{train}}}(\mathbf{x}_{ ext{test}}^{(m)}) \end{bmatrix}$

0 0 X X 0 X X

• Partition data, e.g., 2/3 for train and 1/3 for test.



A.k.a. holdout splitting.

EXAMPLE: POLYNOMIAL REGRESSION

Previous example:



$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \theta_0 + \theta_1 \mathbf{x} + \cdots + \theta_d \mathbf{x}^d = \sum_{j=0}^d \theta_j \mathbf{x}^j.$$

EXAMPLE: POLYNOMIAL REGRESSION / 2





- d = 1: MSE = 0.038: clearly underfitting
- *d* = 3: MSE = 0.002: pretty OK
- d = 9: MSE = 0.046: clearly overfitting

While train error monotonically decreases in *d*, test error shows that high-d polynomials overfit.

TEST ERROR

Let's plot train and test MSE for all d:



× 0 0 × 0 × ×

degree of polynomial

Increasing model complexity tends to cause

- a decrease in training error, and
- a U-shape in test error

(first underfit, then overfit, sweet-spot in the middle).

- Boston Housing data
- Polynomial regression (without interactions)

The training error...

 decreases with smaller training set size as it becomes easier for the model to learn all observed patterns perfectly.



0 0 X X 0 X X

The training error...

• decreases with increasing model complexity as the model gets better at learning more complex structures.



× 0 0 × 0 × ×

The test error...

• will typically decrease with larger training set size as the model generalizes better with more data to learn from.





The test error...

• will have higher variance with smaller test set size.





The test error...

• will have higher variance with increasing model complexity.





BIAS AND VARIANCE

- Test error is a good estimator of GE, given a) we have enough data b) test data is representative i.i.d.
- Estimates for smaller test sets can fluctuate considerably this is why we use resampling in such situations. Repeated $\frac{2}{3} / \frac{1}{3}$ holdout splits: iris (n = 150) and sonar (n = 208).



× × 0 × × ×

BIAS-VARIANCE OF HOLD-OUT – EXPERIMENT

Hold-out sampling produces a trade-off between **bias** and **variance** that is controlled by split ratio.

- Smaller training set \rightarrow poor fit, pessimistic bias in $\widehat{\operatorname{GE}}$.
- Smaller test set \rightarrow high variance.

Experiment:

- spirals data (sd = 0.1), with CART tree.
- Goal: estimate real performance of a model with $|\mathcal{D}_{train}| = 500$.
- Split rates $s \in \{0.05, 0.10, ..., 0.95\}$ with $|\mathcal{D}_{train}| = s \cdot 500$.
- Estimate error on $\mathcal{D}_{\text{test}}$ with $|\mathcal{D}_{\text{test}}| = (1 s) \cdot 500$.
- 50 repeats for each split rate.
- Get "true" performance by often sampling 500 points, fit learner, then eval on 10⁵ fresh points.

× × ×

BIAS-VARIANCE OF HOLD-OUT – EXPERIMENT / 2



× 0 0 × × ×

- Clear pessimistic bias for small training sets we learn a much worse model than with 500 observations.
- But increase in variance when test sets become smaller.

BIAS-VARIANCE OF HOLD-OUT – EXPERIMENT / 3

- Let's now plot the MSE of the holdout estimator.
- NB: Not MSE of model, but squared difference between estimated holdout values and true performance (horiz. line in prev. plot).
- Best estimator is ca. train set ratio of 2/3.
- NB: This is a single experiment and not a scientific study, but this rule-of-thumb has also been validated in larger studies.



× < 0 × × ×