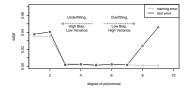
# **Introduction to Machine Learning**

# **Evaluation Test Error**





#### Learning goals

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

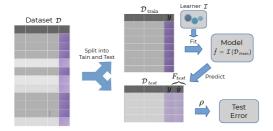
#### TEST ERROR AND HOLD-OUT SPLITTING

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

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

× × ×

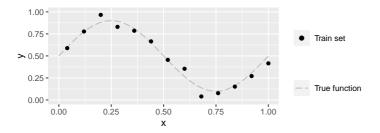
• Partition data, e.g., 2/3 for train and 1/3 for test.



A.k.a. holdout splitting.

## **EXAMPLE: POLYNOMIAL REGRESSION**

#### Previous example:

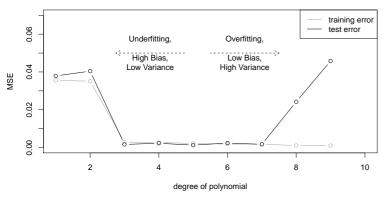


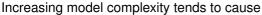


$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \theta_0 + \theta_1 \mathbf{x} + \dots + \theta_d \mathbf{x}^d = \sum_{i=0}^d \theta_i \mathbf{x}^i.$$

#### **TEST ERROR**

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





- a decrease in training error, and
- a U-shape in test error (first underfit, then overfit, sweet-spot in the middle).

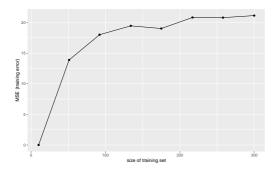


#### TRAINING VS. TEST ERROR

- 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.

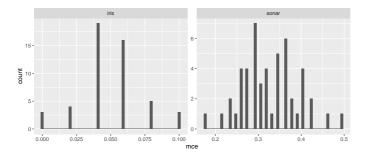




#### **BIAS AND VARIANCE**

- 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).





### BIAS-VARIANCE OF HOLD-OUT – EXPERIMENT

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

- $\bullet$  Smaller training set  $\to$  poor fit, pessimistic bias in  ${\rm GE}.$
- Smaller test set → high variance.



#### Experiment:

- spirals data (sd = 0.1), with CART tree.
- Goal: estimate real performance of a model with  $|\mathcal{D}_{\text{train}}| = 500$ .
- Split rates  $s \in \{0.05, 0.10, ..., 0.95\}$  with  $|\mathcal{D}_{\mathsf{train}}| = s \cdot 500$ .
- 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 10<sup>5</sup> fresh points.