ESTIMATING THE GENERALIZATION ERROR

- For a fixed model, we are interested in the Generalization Error (GE): $\mathrm{GE}\left(\hat{f}, L\right) := \mathbb{E}\left[L\left(y, \hat{f}(\mathbf{x})\right)\right]$, i.e. the expected error the model makes for data $(\mathbf{x}, y) \sim \mathbb{P}_{xy}$.
- We need an estimator for the GE with m test observations:

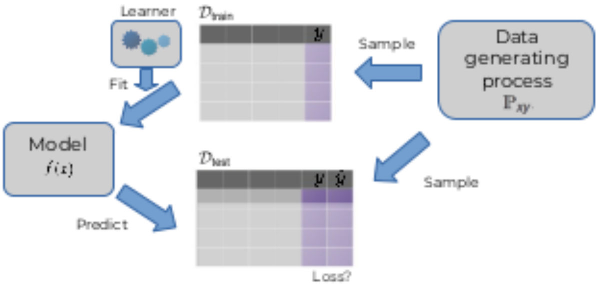
$$\widehat{\mathrm{GE}}(\hat{f}, L) := \frac{1}{m} \sum_{(\mathbf{x},y)} \left[L\left(y, \hat{f}(\mathbf{x})\right)\right]$$

- However, if $(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}$, $\widehat{\mathrm{GE}}(\hat{f}, L)$ will be biased via overfitting the training data.
- Thus, we estimate the GE using unseen data $(\mathbf{x}, y) \in \mathcal{D}_{\text{test}}$:

$$\widehat{\mathrm{GE}}(\hat{f}, L) := \frac{1}{m} \sum_{(\mathbf{x},y) \in \mathcal{D}_{\text{test}}} \left[L\left(y, \hat{f}(\mathbf{x})\right)\right]$$

ESTIMATING THE GENERALIZATION ERROR / 2

- Usually, we have no access to new **unseen** data.
- Thus, we divide our data set manually into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$.
- This process is depicted below.



METRICS FOR CLASSIFICATION / 2

For hard-label classification, the confusion matrix is a useful representation:

		True Class y	
		$+$	$-$
Pred.	$+$	True Positive (TP)	False Positive (FP)
$\hat{y}$	$-$	False Negative (FN)	True Negative (TN)

From this matrix a variety of evaluation metrics, including precision and recall, can be computed.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

ESTIMATING THE GENERALIZATION ERROR (BETTER)

While

$$\widehat{\mathrm{GE}}(\hat{f}, L) := \frac{1}{m} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{test}}} \left[L\left(y, \hat{f}(\mathbf{x})\right)\right]$$

will be unbiased, with a small m it will suffer from high variance. We have two options to decrease the variance:

- Increase m.
- Compute $\widehat{\mathrm{GE}}(\hat{f}, L)$ for multiple test sets and aggregate them.

With a finite amount of data, increasing m would mean to decrease the size of the training data. Thus, we focus on using multiple (B) test sets:

$$\mathcal{J} = \left((J_{\mathrm{train},1}, J_{\mathrm{test},1}), \ldots, (J_{\mathrm{train},B}, J_{\mathrm{test},B})\right).$$

where we compute $\widehat{\mathrm{GE}}(\hat{f}, L)$ for each set and aggregate the estimates. These B sets are generated through **resampling**.