Q & A: Model Evaluation

Lecture series „Machine Learning“

Niels Landwehr

Research Group „Data Science“
Institute of Computer Science
University of Hildesheim

Quiz: Train-Test Splits

- Let's say we want to train a model to solve a simple emotion recognition problem: given an image of the face of a person, say if the person is happy or sad
  - Input to model: color image
  - Output: binary class happy/sad

$\mapsto$ "happy"

- To train a model, three friends (Anne, Brian, Charlie) collect training images:

Example	Person in photo	Facial Expression (Label)
$\mathbf{x}_1$	Anne	happy
$\mathbf{x}_2$	Brian	happy
$\mathbf{x}_3$	Charlie	happy
$\mathbf{x}_4$	Anne	sad
$\mathbf{x}_5$	Brian	sad
$\mathbf{x}_6$	Charlie	sad

- They want to train a model and also estimate the accuracy of the model on novel data using cross-validation

Quiz: Train-Test Splits

Example	Person in photo	Facial Expression (Label)
$\mathbf{x}_1$	Anne	happy
$\mathbf{x}_2$	Brian	happy
$\mathbf{x}_3$	Charlie	happy
$\mathbf{x}_4$	Anne	sad
$\mathbf{x}_5$	Brian	sad
$\mathbf{x}_6$	Charlie	sad

- For cross-validation, which of the following ways of performing cross-validation would work best (colors indicate partitioning of data into folds):

a) Two-fold: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$

b) Two-fold: $\mathbf{x}_5, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_6, \mathbf{x}_3$

c) Three-fold: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$

d) Three-fold: $\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_2, \mathbf{x}_5, \mathbf{x}_3, \mathbf{x}_6$

e) Leave-one-out: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$

f) All work equally well

Quiz: Train-Test Splits

- **Solution**: let's discuss the different splits in turn

a) Two-fold: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$

This will not work because in the training sets per fold, the model sees only one class. It will predict that class also for all instances in the test set of that fold, so that the estimated accuracy is zero.

Example	Person in photo	Facial Expression (Label)
$\mathbf{x}_1$	Anne	happy
$\mathbf{x}_2$	Brian	happy
$\mathbf{x}_3$	Charlie	happy
$\mathbf{x}_4$	Anne	sad
$\mathbf{x}_5$	Brian	sad
$\mathbf{x}_6$	Charlie	sad

b) Two-fold: $\mathbf{x}_5, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_6, \mathbf{x}_3$

This will work somewhat better because at least one example from each class is in each split.

Quiz: Train-Test Splits

- **Solution**: let's discuss the different splits in turn

$c)$ Three-fold: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$

Similar to b): there is at least one example for each class in each fold.

$e)$ Leave-one-out: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$

Slightly better than other solutions so far because there are at least two examples per class in training sets

Example	Person in photo	Facial Expression (Label)
$\mathbf{x}_1$	Anne	happy
$\mathbf{x}_2$	Brian	happy
$\mathbf{x}_3$	Charlie	happy
$\mathbf{x}_4$	Anne	sad
$\mathbf{x}_5$	Brian	sad
$\mathbf{x}_6$	Charlie	sad

$d)$ Three-fold: $\mathbf{x}_1, \mathbf{x}_4, \mathbf{x}_2, \mathbf{x}_5, \mathbf{x}_3, \mathbf{x}_6$

Best solution. The training data has been collected by three persons, but presumably the model should at deployment time make predictions for **new persons**.
To estimate the future error rate reliably, should test on new persons during cross-validation as well.
Additionally, there are also two examples per class in training sets.

Quiz: Holdout With Validation

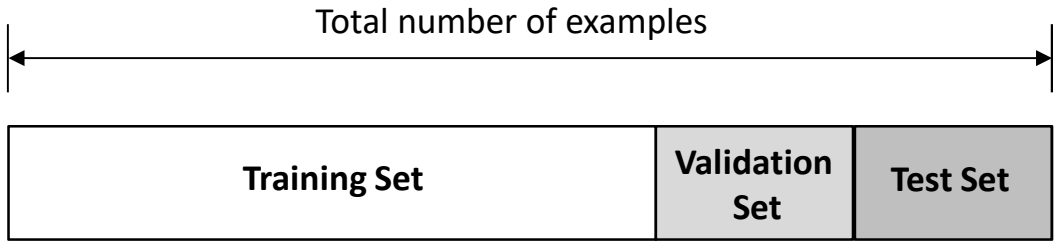
- Let's say we want to train a model with two hyperparameters, namely a learning rate $\eta \in \mathbb{R}_{>0}$ and a regularization weight $\lambda \in \mathbb{R}_{\geq 0}$
- The overall set of data that we have available is $\mathcal{L} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_L, y_L)\}$
- We want to train a model with optimal hyperparameters and also get an estimate of the performance of that model on novel instances
- For hyperparameter tuning, we want to use grid-search with the candidate values $\eta \in \{0.001,\ 0.01,\ 0.1\}$ and $\lambda \in \{0.01,\ 0.03,\ 0.1,\ 0.3,\ 1.0\}$

- **Question**: how many models do we have to train overall when using holdout testing with training, validation and test set?

a) 3 models
b) 8 models
c) 15 models
d) 16 models
e) 17 models
f) 45 models

Quiz: Holdout With Validation

- **Solution**: Overall, we have to train 17 models
  - There are 15 combinations of hyperparameters η and λ
  - For each of the 15 combination we train a model on the training set and evaluate the model on the validation set => 15 model trainings
  - Based on this, we then pick the best hyperparameter combination and retrain the model on the training + validation sets => 1 model training
  - We evaluate the model on the test set for the final error estimate
  - We then retrain the final model on all of the data => 1 model training



Quiz: Confidence Interval

- Assume we perform holdout testing where we split the overall data $\mathcal{L}$ into a training set $\mathcal{D}$ and a test set $\mathcal{T}$. Which of the following statements about the confidence interval are correct?

1. The size of the confidence interval will typically decrease with increasing $|\mathcal{D}|$
2. The size of the confidence interval will typically decrease with increasing $|\mathcal{T}|$
3. The error estimator $\hat{R}_{\mathcal{T}}(f_{\boldsymbol{\theta}^{\mathcal{D}}})$ is an unbiased estimator for the performance of model $f_{\boldsymbol{\theta}^{\mathcal{D}}}$
4. The error estimator $\hat{R}_{\mathcal{T}}(f_{\boldsymbol{\theta}^{\mathcal{D}}})$ is a pessimistic estimator for the performance of model $f_{\boldsymbol{\theta}^{\mathcal{D}}}$
5. We have obtained a confidence interval ϵ for a confidence level of $1-\delta = 0.95$. This means that with probability 95%, the true risk is within the confidence interval

Quiz: Confidence Interval

- **Solution**:

1. Not correct. Increasing training set size will lead to a better model, but not a lower variance (= smaller confidence interval) for the estimate on the test set
2. Correct. Increasing test set size will lead to a lower variance and therefore a smaller confidence interval
3. Correct: For the model $f_{\boldsymbol{\theta}^{\mathcal{D}}}$, the error estimator $\hat{R}_{\mathcal{T}}(f_{\boldsymbol{\theta}^{\mathcal{D}}})$ is unbiased. It is only pessimistic for the model $f_{\boldsymbol{\theta}^{\mathcal{L}}}$ that we train in the end
4. Not correct, see point 3
5. Not correct: once the confidence interval is known, the true risk is in the confidence interval or not (no distribution over true risk). Correct would be: approximately 95% of the time we will get a confidence interval that contains the true risk.