Model selection (addendum)

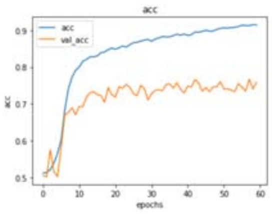
Akito Sakurai

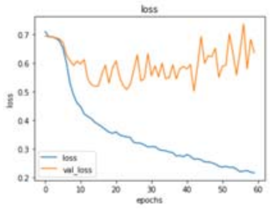
Three types of datasets

- There are three types of datasets used in machine learning.
  training, validation, test
- “training” dataset is clear.
  - “validation” dataset and “test” dataset are ambiguous.
- “validation” dataset is used in training
- “test” dataset is to be used just once

validation

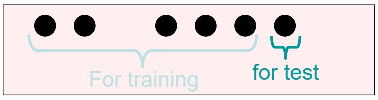
- A validation dataset is repeatedly used in a training process.
- When you are to estimate generalization error in a training process, you apply the model to the validation dataset.
  - Usually you stop the training process, when you find a time or complexity at which the validation error/loss hits a minimum.





k-fold cross validation

Divide the training dataset into k groups, train the model with the $(k-1)$ groups and measure the prediction errors on the remaining group (test set) ; and repeat the process k times by changing the test set.



It is not almighty, but works in many cases.

CV measures goodness of algorithms/model architectures
CV is used to determine the best architecture and/or parameters.

Validation/test dataset

- Because a validation dataset is used in training (to see when the training is to be halted), it is not to be used for estimating its generalization error.
- A test dataset is the one to be used to estimate generalization error.
  - If you see the estimated generalization error and retrain the trained model, the test dataset is never to be used for a test dataset

Validation/test sets are quite often mixed up but should be clearly distinguished when you do experiments.