Resampling methods and Model selection

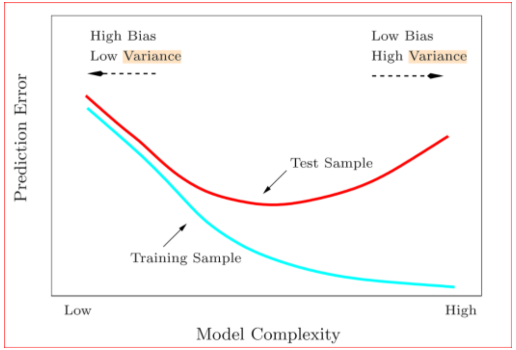
Rao Muhammad Umer, PostDoc,

Institute of AI for Health (AIH),

Helmholtz Muenchen, Germany.

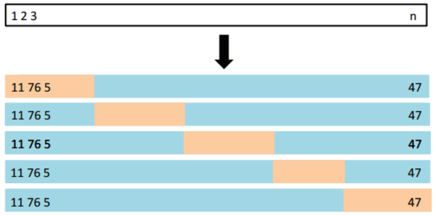
Resampling Methods

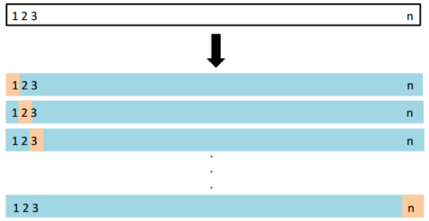
- **Measurement of Generalization Performance**
- Typically we do not have access to real world test examples
- Use the given "**training**" set for **approximating / estimating** the **generalization performance**.
- **Guidelines**
  - There should be "**enough**" training examples left.
  - **Test labels** should **not** be used, directly or indirectly, during **training**.
    - Test data (without labels) can be used.
  - You should be clear about the intended use and application of the system.
  - You should be clear about the objective of performance evaluation.
- **Issues:**
  - The **variance** of our estimate **increases** as the **size of the testset decreases**.
  - A small **increase** in the pessimistic **bias** when we **decrease** the size of the **training set**.
  - **bias–variance problem**



Resampling Methods

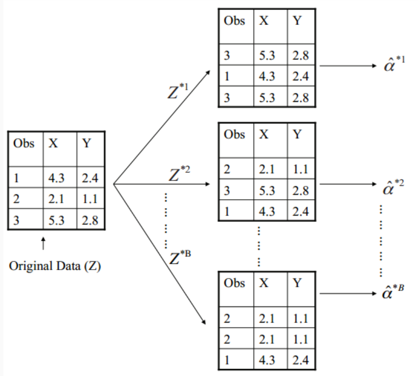
- **Cross-Validation: K-fold**
  - For estimation of variation
  - Divide the data into **K folds**
    - For k = 1…K
      - Train on K-1 sets leaving the k^{th} set out for validation
      - Validate on the k^{th} set and obtain the performance metrics
    - Report the average and the variation in the performance.
- If **K = n (Number of examples),** then this extreme case is called **Leave One Out CV (LOOCV)**
  - Useful if the amount of **data** is **small**.
- **Stratification** (for **Imbalanced data**)
  - Make sure that each fold contains the **same number of examples** as the overall data
    - If a class has 20 percent examples in the whole dataset, in all samples drawn from the dataset, it should also have approximately 20 percent examples.
- What will be the impact on approximated performance with increase in K? what is the smallest/largest value of K?





Resampling Methods

- **Bootstrapping**
  - More **overlap** between **samples**
  - Useful for very **small datasets**
  - For i = 1…b
    - Pick N examples at random from the data set of N examples with replacement.
    - Train the classifier on these examples.
    - Evaluate the classifier on the original data set and obtain the performance metric.
    - Average the performance metric to obtain "**resubstitution accuracy**": **acc_s**
  - However, the **resubstitution accuracy** is an **over-estimate** of the **true accuracy** due to the inclusion of the training examples in testing.
  - To avoid over-estimate problem, solution is **.632** or the **.632+ bootstrap**.



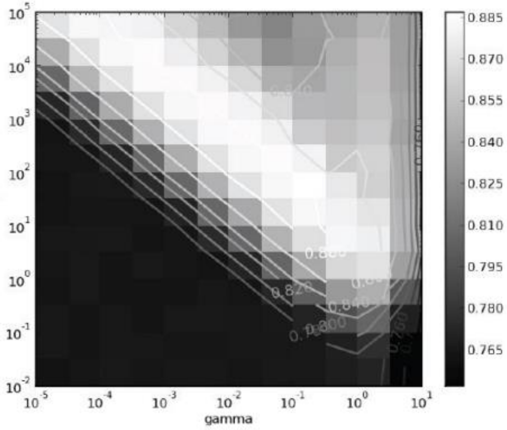
- **Jackknifing:**
  - In cross-validation you compute a statistic on the left out sample(s) using a model built on the kept samples.
  - In jackknifing, you compute a statistic from the kept samples only.

Resampling Methods

- **So what to use?**
  - **5/10 Fold Cross-Validation** is good.
  - However, for **small sample sizes**, it can have a **large variance** in which case you can use **LOOCV** or the **.632** or the **.632+ bootstrap**.

Model Hyperparameter Selection

- **Grid Search**
  - Exhaustive Search through Cross-Validation
    - Recommended: Nested Cross Validation or separate test set
- There can be a range of parameter values that yield optimal values and these equivalent points in the parameter space fall along a ridge.
- **Searching for optimal parameters selection**
  - Regularization Path Finding
  - Gradient Based Approaches
  - Evolutionary approaches
- **Some links:**
  - http://scikit-learn.org/stable/modules/grid_search.html
  - http://hyperopt.github.io/
  - http://hyperopt.github.io/hyperopt-sklearn/



Thanks!

Q&A?