

Lecture 5. Learning (II) Sampling

Outline

- Complexity of Model
- Sampling Issue
- Competitive Learning

Model Complexity

Issue: How complicate the model $F(W,x)$ should be?

- Complicated model or models with too many free parameters may cause over-fitting – small bias but large variance.
- Over-simplified model may lead to large bias even variance may be small.
- If all being equal, heuristics advocates simpler model:

Ockham's razor (Occam's razor): *The most likely hypothesis is the simplest one that is consistent with all observations.*

Minimum Description Length (MDL)

More about Complexities ...

- In practice, the bias can not be measured since $g(x)$ is unknown. Only variance can be measured. So, what should we do?
- We can only try to minimize $J(W)$!
- However, we do not have $J(W)$ either. We only have $J_D(W)$!

$J(W) = E\{|d(i) - F(W, x(i))|^2\}$ -- averaged over entire population

$J_D(W) = \sum_{i \in D} |d(i) - F(W, x(i))|^2$ -- A single realization at $\{x(i); i \in D\}$

- QUESTION: Can $J_D(W)$ predicts $J(W)$ well?

Effect of Finite Training Data

- What we settled for is to approximate $J(W)$ by averaging over several different choices of D . That is,

$$\hat{J}(W) = \langle J_D(W) \rangle$$

Methods –

- Partition of training data (interpolation) and test data (extrapolation).
- M-way cross validation
- Ensemble average (Boot-strapping) rather than single realization.

Training and Testing Error

- Training set D is the data used to develop $F(W,x)$. Hence $W = W(D)$
- Testing set T is the data used to estimate $J(W)$ of a given model $F(W,x)$.
- T should be sampled from the same population as D but T and D must be independently sampled and have no overlap.
- $F(W,x)$ interpolates D at each $x(i)$, and extrapolates to samples in T .
- The fitness of the model $F(W,x)$, measured by $J(W)$ must be measured on T rather than on D .
- During the development of $F(W,x)$, only data in D should be used. Data in T must be reserved for testing.

Sampling Strategy

- Partition available samples into dis-joint training set D and testing set T.
 - Disadvantage: Available data in T are wasted as they are not used for developing $F(W,x)$.
- M-Way Cross-Validation: Partition available data into M equal-sized disjoint partitions. Perform model fitting M times. Each time use M-1 partitions as D and the remaining partition as T. Obtain an estimate $J(W^{(m)})$. Then estimate $J(W) = \langle J(W^{(m)}) \rangle$.
 - Can be used to help deciding model orders, model selections.
 - But can not tell which of the M model should be used.

Ensemble Averages

- Instead of partition training set into disjoint partitions, we sample repeatedly with replacement the subset D .
- Then we estimate $J(W)$ by

$$\hat{J}(W) = \frac{1}{M} \sum_{i=1}^M J_{D(i)}(W)$$

- We may also estimate $F(W, x)$ by

$$\hat{F}(W, x) = \frac{1}{M} \sum_{i=1}^M F_{D(i)}(W, x)$$

provided the average can be taken for the given model.