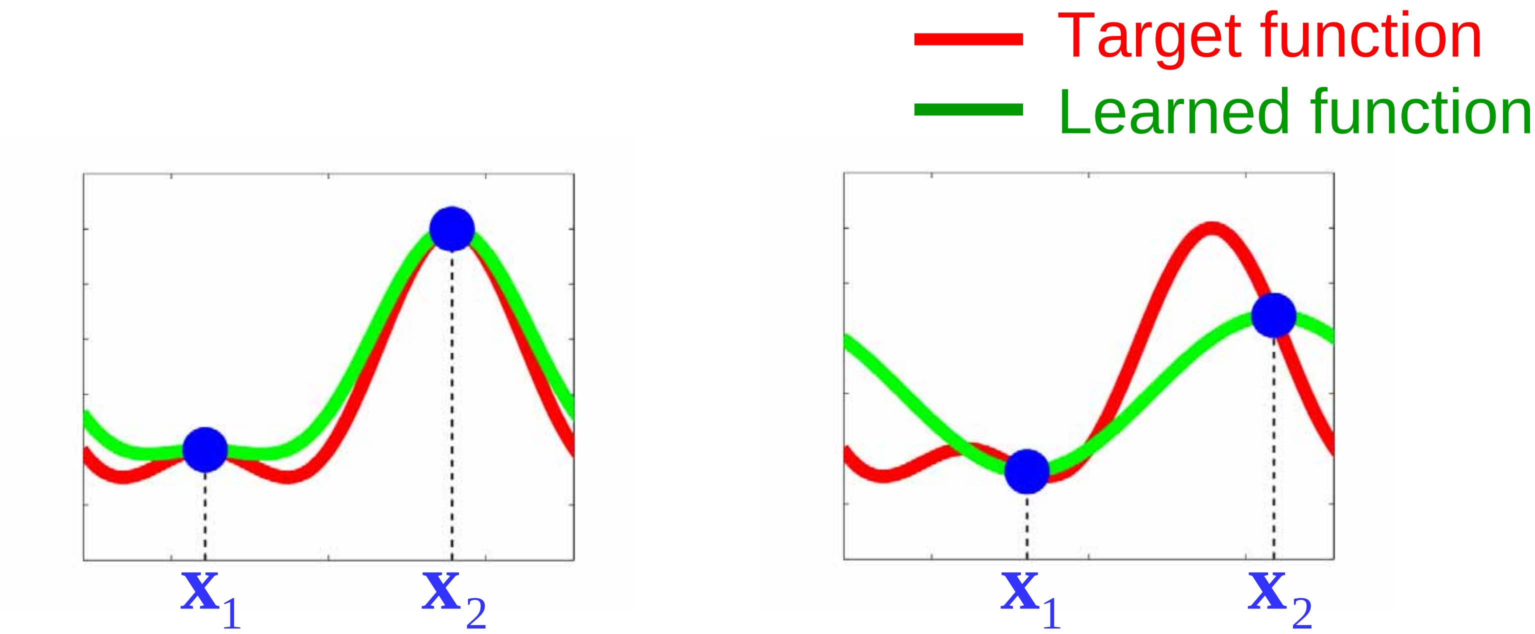
Pattern Information Processing: Active Learning

Masashi Sugiyama
(Department of Computer Science)

Contact: W8E-505
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/~sugi/

Active Learning

For obtaining good learning results, training input points should be determined appropriately.

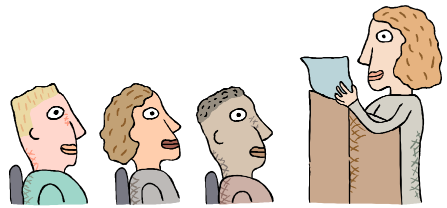


Good questions

Bad questions

Active Learning: Analogy to Real Life

- It is not interesting to passively attend the lecture.



- It is more effective to actively ask questions in the lecture.

Formal Description

$$G = \int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{t}) - f(\boldsymbol{t}) \right)^2 q(\boldsymbol{t}) d\boldsymbol{t}$$

- Determine training input points so that

$$\min_{\{\boldsymbol{x}_i\}_{i=1}^{n}} G$$

Setting

- $q(\boldsymbol{x})$ is known.
- Linear model:

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

- Least-squares learning:

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{L}\boldsymbol{y}$$

$$\boldsymbol{L} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$$

$$\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i)$$

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_b)^\top$$

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$$

Estimating Generalization Error

$$\min_{\{\boldsymbol{x}_i\}_{i=1}^n} G \qquad G = \int_{\mathcal{D}} \left(\hat{f}(\boldsymbol{t}) - f(\boldsymbol{t}) \right)^2 q(\boldsymbol{t}) d\boldsymbol{t}$$

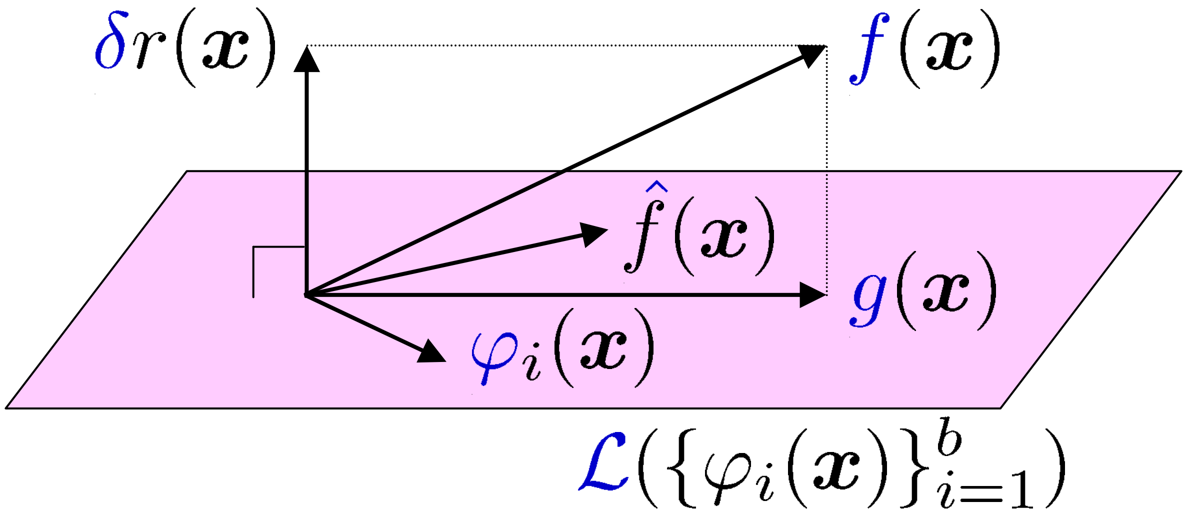
- We have to estimate unknown generalization error.
- This is similar to model selection.
- We do not have training output values $\{y_i\}_{i=1}^n$ in active learning!

Decomposition of Target Function

$$f(\boldsymbol{x}) = g(\boldsymbol{x}) + \delta r(\boldsymbol{x}) \qquad \int r(\boldsymbol{x})^2 q(\boldsymbol{x}) d\boldsymbol{x} = 1$$

$$g(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i^* \varphi_i(\boldsymbol{x}) \qquad \int \varphi_i(\boldsymbol{x}) r(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} = 0$$

$$\delta \geq 0$$



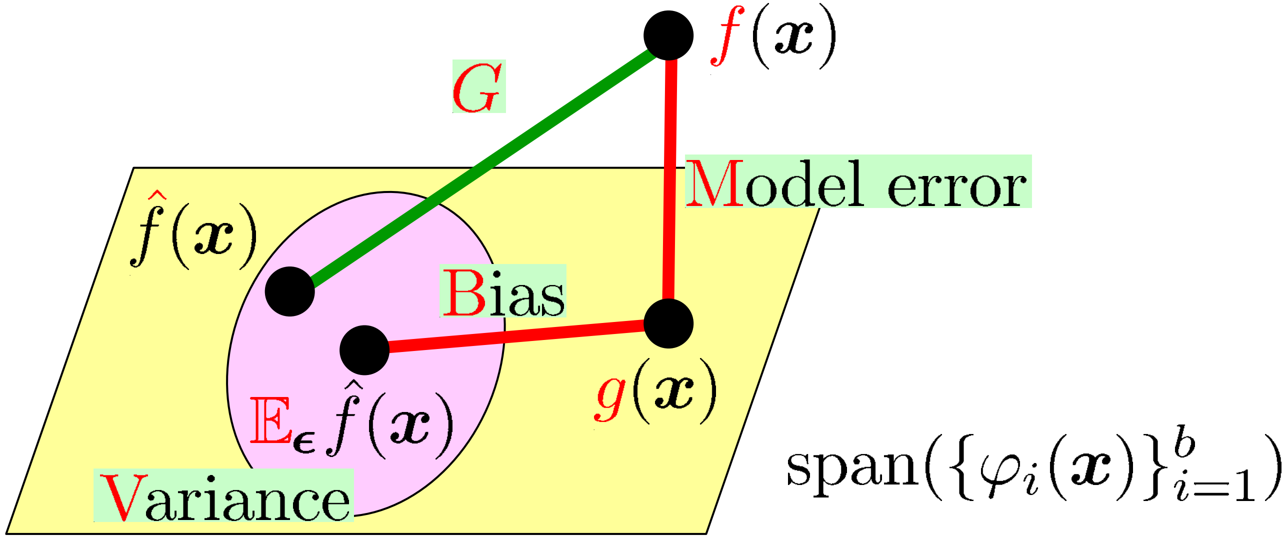
Bias/Variance Decomposition

$$\mathbb{E}_{\epsilon} G = \mathbb{E}_{\epsilon} \int \left(\hat{f}(\boldsymbol{x}) - g(\boldsymbol{x}) - \delta r(\boldsymbol{x})\right)^2 q(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \delta^2 \int r(\boldsymbol{x})^2 q(\boldsymbol{x}) d\boldsymbol{x} \qquad \text{(model error)}$$

$$+ \int \left(\mathbb{E}_{\epsilon} \hat{f}(\boldsymbol{x}) - g(\boldsymbol{x})\right)^2 q(\boldsymbol{x}) d\boldsymbol{x} \qquad \text{(bias)}$$

$$+ \mathbb{E}_{\epsilon} \int \left(\hat{f}(\boldsymbol{x}) - \mathbb{E}_{\epsilon} \hat{f}(\boldsymbol{x})\right)^2 q(\boldsymbol{x}) d\boldsymbol{x} \qquad \text{(variance)}$$



Assumption

- We assume that model is correct
  - $\delta = 0$: model error vanishes
  - Least squares is unbiased: bias vanishes
- Only variance remains!

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} G &= \mathbb{E}_{\boldsymbol{\epsilon}} \int \left(\hat{f}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\epsilon}} \hat{f}(\boldsymbol{x}) \right)^2 q(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sigma^2 \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{L}^\top) \\
&= \sigma^2 \mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}) \\
&\propto \mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}^\top \boldsymbol{X})^{-1})
\end{aligned}
$$

$$
\boldsymbol{U}_{i,j} = \int \varphi_i(\boldsymbol{x}) \varphi_j(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x}
$$

Active Learning with LS

- Determine $\{\boldsymbol{x}_i\}_{i=1}^n$ so that

$$\underset{\{\boldsymbol{x}_i\}_{i=1}^n}{\mathrm{argmin}} \left[\mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}) \right]$$

- In active learning, we can not use training output values $\{y_i\}_{i=1}^n$ for estimating generalization error.
- We considered zero-bias cases and evaluated the variance!

How to Optimize

- Determine $\{\boldsymbol{x}_i\}_{i=1}^n$ so that

$$\mathop{\mathrm{argmin}}_{\{\boldsymbol{x}_i\}_{i=1}^n} \left[\mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}) \right]$$

- For trigonometric polynomial models, the solution can be analytically obtained.
- However, in general, simultaneously optimizing n points is not tractable.

How to Optimize (cont.)

- Major approaches to avoid intractability:
  - Optimize points one by one in a greedy manner
  - Optimize probability distribution from which training input points are drawn.

$$\{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n} \overset{i.i.d.}{\sim} p^{(k)}(\boldsymbol{x})$$

$$\underset{k}{\operatorname{argmin}} \left[\operatorname{tr}(\boldsymbol{U}(\boldsymbol{X}^{(k)\top}\boldsymbol{X}^{(k)})^{-1}) \right]$$

$$\boldsymbol{X}_{i,j}^{(k)} = \varphi_j(\boldsymbol{x}_i^{(k)})$$

When Model Is Not Correct

- When model is not correct, least-squares is no longer unbiased (even asymptotically).
- Instead, the following importance-weighted LS is asymptotically unbiased.

$$\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \frac{q(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)} \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 \right]$$

$$\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} p(\boldsymbol{x})$$

$$\boldsymbol{t} \sim q(\boldsymbol{x})$$

IWLS

- IWLS learning result is given by

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i \varphi_i(\boldsymbol{x})$$

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{L}_W \boldsymbol{y}$$

$$\boldsymbol{L}_W = (\boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D}$$

$$\boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i)$$

$$\boldsymbol{D} = \mathrm{diag}\left(\frac{q(\boldsymbol{x}_1)}{p(\boldsymbol{x}_1)}, \frac{q(\boldsymbol{x}_2)}{p(\boldsymbol{x}_2)}, \ldots, \frac{q(\boldsymbol{x}_n)}{p(\boldsymbol{x}_n)} \right)$$

Asymptotic Unbiasedness of IWLS

- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha}^* + \delta \boldsymbol{z}_r + \boldsymbol{\epsilon}$

$$\boldsymbol{z}_r = (r(\boldsymbol{x}_1), r(\boldsymbol{x}_2), \ldots, r(\boldsymbol{x}_n))^\top$$

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$$

- $\boldsymbol{L}_W \boldsymbol{X}\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*$

- $[\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{z}_r]_k = \frac{1}{n}\sum_{i=1}^{n} \varphi_k(\boldsymbol{x}_i) r(\boldsymbol{x}_i) \frac{q(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}$

$$\to \int_{\mathcal{D}} \varphi_k(\boldsymbol{x}) r(\boldsymbol{x}) \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} p(\boldsymbol{x}) d\boldsymbol{x}$$

(Law of large numbers)

$$= \int_{\mathcal{D}} \varphi_k(\boldsymbol{x}) r(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} = 0$$

Active Learning with IWLS

- Variance of IWLS is

$$\sigma^2 \mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_W\boldsymbol{L}_W^\top)$$

- Determine $p(\boldsymbol{x})$ so that

$$\mathop{\mathrm{argmin}}_{p(\boldsymbol{x})}\left[\mathrm{tr}(\boldsymbol{U}\boldsymbol{L}_W\boldsymbol{L}_W^\top)\right]$$

Notification of Final Assignment

1. Apply supervised learning techniques to your data set and analyze it.
2. Write your opinion about this course

- Final report deadline: Aug 11th (Fri.)
- Only e-mail submission is accepted!

sugi@cs.titech.ac.jp