Pattern Information Processing: Active Learning

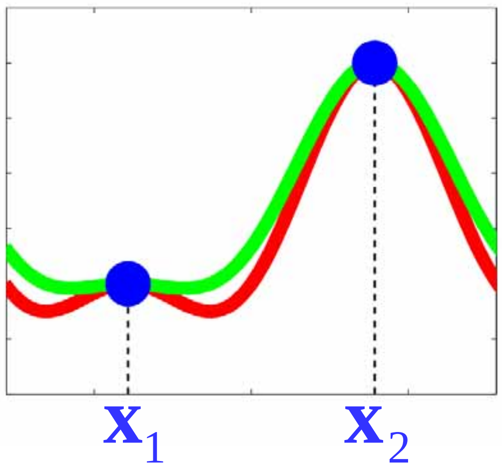
Masashi Sugiyama
(Department of Computer Science)

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

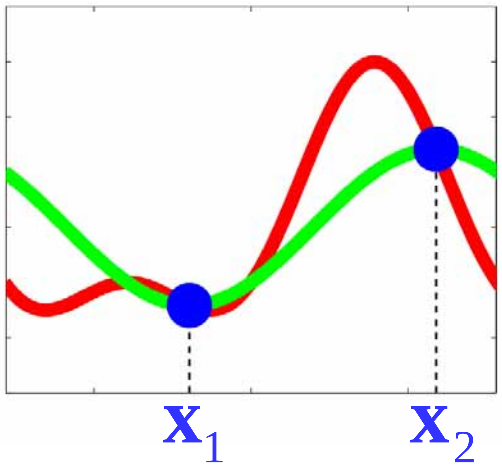
Active Learning

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

Target function
Learned function



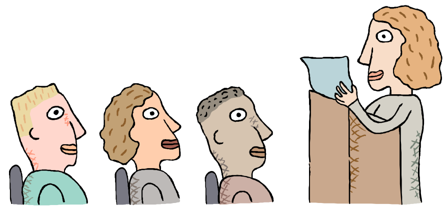
Good questions



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

- Test input point: $\boldsymbol{t}$
- Test input density: $q(\boldsymbol{t})$
- Generalization error (expected test error):

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

- Determine training input points so that

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

Setting

- Training examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$
  - Training outputs y_i : additive noise contained

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

  - Output noise ϵ_i : i.i.d. with mean zero

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\epsilon_i] = 0 \qquad \mathbb{E}_{\boldsymbol{\epsilon}}[\epsilon_i \epsilon_j] = \begin{cases} \sigma^2 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

Setting (cont.)

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

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

- Least-squares learning:

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

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

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

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

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

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!

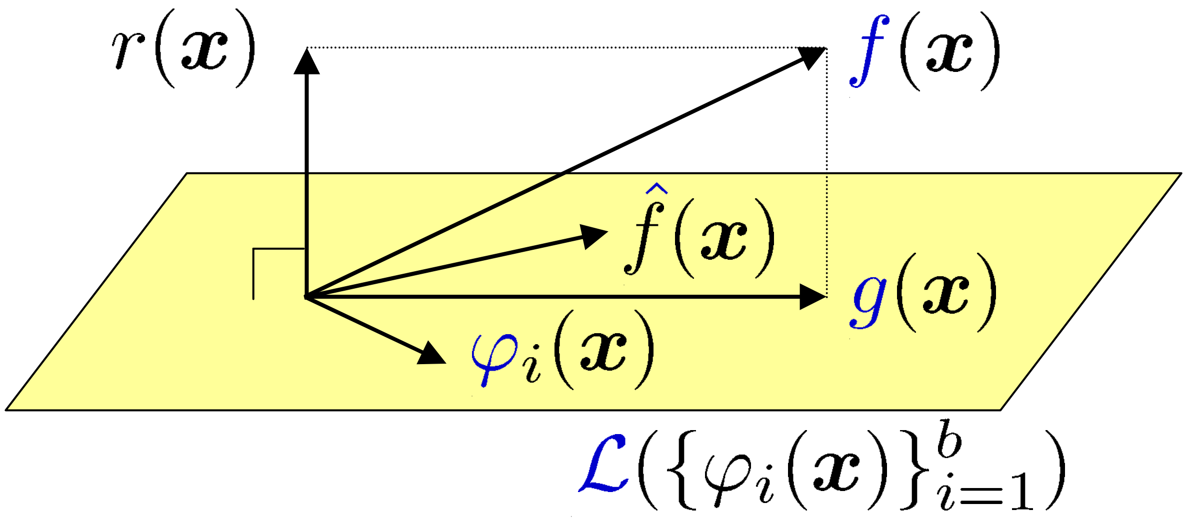
Decomposition of Target Function

$$f(\boldsymbol{x}) = g(\boldsymbol{x}) + r(\boldsymbol{x})$$

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

$$\int \varphi_i(\boldsymbol{x}) r(\boldsymbol{x}) q(\boldsymbol{x}) d\boldsymbol{x} = 0$$

($\varphi_i(\boldsymbol{x})$ and $r(\boldsymbol{x})$ are orthogonal)



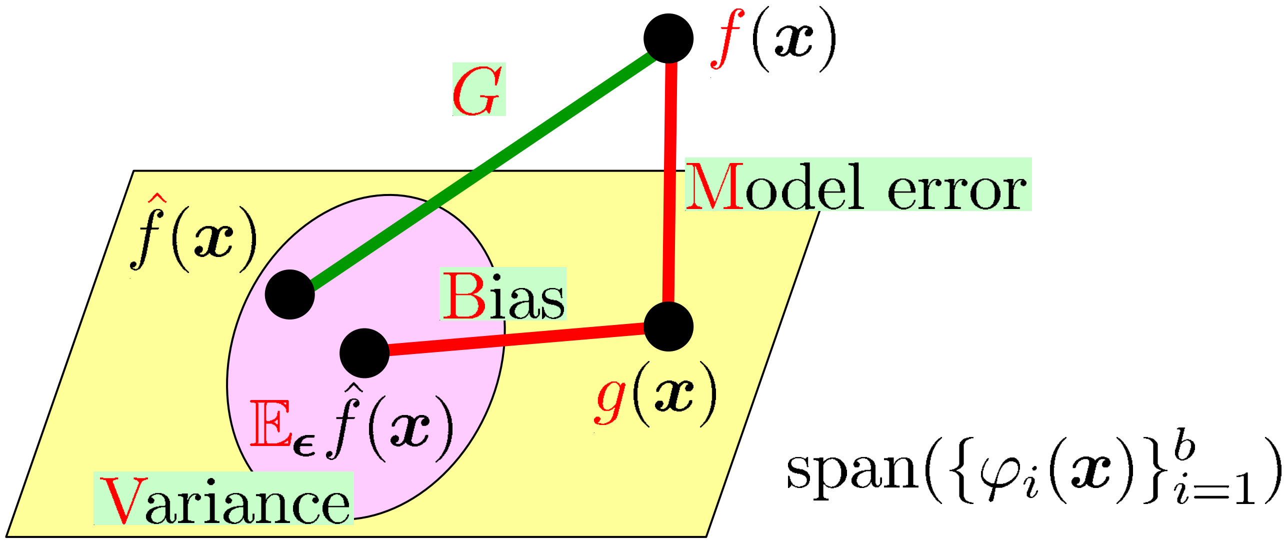
Bias/Variance Decomposition

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

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

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

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



Assumption

- We assume that model is correct
  - $r(\boldsymbol{x}) = 0$: model error vanishes
  - LS is unbiased: bias vanishes
- Only variance remains!

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

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

$$= \sigma^2 \mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

$$\propto \mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

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

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.
- Optimize training input density based on

$$\mathrm{tr}(\boldsymbol{U}(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

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

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

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

When Model Is Not Correct

- When model is not correct, least-squares is no longer unbiased (even asymptotically) due to $p(\boldsymbol{x}) \neq q(\boldsymbol{x})$.
- Instead, the following importance-weighted LS is asymptotically unbiased.

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

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

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

Solution of IWLS

- IWLS learning result is given by

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

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

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

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

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

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

Asymptotic Unbiasedness of IWLS

- We show

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\alpha}}] \to \boldsymbol{\alpha}^* \text{ as } n \to \infty$$

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

$$\boldsymbol{X}\boldsymbol{\alpha}^* = (g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_n))^\top$$
$$\boldsymbol{z}_r = (r(\boldsymbol{x}_1), \ldots, r(\boldsymbol{x}_n))^\top$$
$$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$$

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$
$$f(\boldsymbol{x}) = g(\boldsymbol{x}) + r(\boldsymbol{x})$$
$$g(\boldsymbol{x}) = \sum_{i=1}^{b} \alpha_i^* \varphi_i(\boldsymbol{x})$$

- $\mathbb{E}_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\alpha}}] = \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{L}_W \boldsymbol{y}]$

$$= (\boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D} \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{y}]$$
$$= (\boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{D} (\boldsymbol{X}\boldsymbol{\alpha}^* + \boldsymbol{z}_r)$$
$$= \boldsymbol{\alpha}^* + (\tfrac{1}{n}\boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X})^{-1} \tfrac{1}{n}\boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{z}_r$$

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\epsilon_i] = 0$$

Asymptotic Unbiasedness of IWLS

- $$[\tfrac{1}{n}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{z}_r]_k = \frac{1}{n}\sum_{i=1}^{n}\varphi_k(\boldsymbol{x}_i)r(\boldsymbol{x}_i)\frac{q(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}$$
$$\to \int_{\mathcal{D}} \varphi_k(\boldsymbol{x})r(\boldsymbol{x})\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}p(\boldsymbol{x})d\boldsymbol{x} \qquad \text{as } n \to \infty$$
(Law of large numbers)
$$= \int_{\mathcal{D}} \varphi_k(\boldsymbol{x})r(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} = 0$$

- $$[\tfrac{1}{n}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X}]_{i,j} = \mathcal{O}(1) \qquad \text{as } n \to \infty$$

- Thus, $\mathbb{E}_{\boldsymbol{\epsilon}}[\hat{\boldsymbol{\alpha}}] \to \boldsymbol{\alpha}^* \text{ as } n \to \infty$.

(Q.E.D.)

Active Learning with IWLS

- Variance of IWLS is

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

- Optimize training input distribution based on

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

$$\boldsymbol{U}_{i,j} = \int \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x} \qquad \boldsymbol{X}_{i,j} = \varphi_j(\boldsymbol{x}_i) \qquad \{\boldsymbol{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} p(\boldsymbol{x})$$

$$\boldsymbol{L}_W = (\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{D} \qquad \boldsymbol{D} = \mathrm{diag}\left(\frac{q(\boldsymbol{x}_1)}{p(\boldsymbol{x}_1)}, \frac{q(\boldsymbol{x}_2)}{p(\boldsymbol{x}_2)}, \ldots, \frac{q(\boldsymbol{x}_n)}{p(\boldsymbol{x}_n)}\right)$$

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 8th (Fri.)
- E-mail submission is also accepted!

sugi@cs.titech.ac.jp

Schedule

- July 8th : preparation for workshop (no lecture)
- July 15th : preparation for workshop (no lecture)
- July 22nd : mini-workshop (starting from 10:40 !)