Parametric regression (cont'd) and nonparametric methods

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 6

multivariate regression

$$y = f(x) + \varepsilon \quad , \quad x \in \mathbb{R}^D \ , \ f : \mathbb{R}^D \to \mathbb{R}$$

Estimate the "true model" $f(x)$ with $\underbrace{g(x|w)}$ assume to be linear:

$$\mathcal{X} = \{ \mathbb{X}_n \}_{n=1}^{N} = \{ (x_n, y_n) \}_{n=1}^{N}$$

$$g(x|w) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

Following the assumption that $\varepsilon \sim N(0, \sigma^2)$, similarly to the 1D case, we will have to minimize an error function

$$E(w_0, w_1, \cdots, w_D \mid \mathcal{X})$$

$$= \frac{1}{2} \sum_{n=1}^{N} \Big(y_n - \underbrace{(w_0 + w_1 x_{n1} + w_2 x_{n2} + \cdots + w_D x_{nD})}_{n\text{-th entry of } Xw} \Big)^2$$

Write $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1D} \\ 1 & x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & & & & \\ 1 & x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix} \in \mathbb{R}^{N \times (D+1)}$ data matrix

$$w = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix} \in \mathbb{R}^{D+1} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N$$

$$E(w|\mathcal{X}) = \frac{1}{2} \| y - Xw \|^2$$

Setting $\frac{\partial E}{\partial w} = \frac{1}{2} \cdot 2 \cdot X^T(Xw - y) = \underbrace{X^T(Xw-y) = 0}_{X^TXw = X^Ty}$

yields $w = (X^TX)^{-1} X^T y =: \hat{w}_{ls}$

Ridge regression

$$\hat{w}_{MAP} = \underset{w}{\operatorname{argmax}} \; \log p(w|x)$$

$$= \underset{w}{\operatorname{argmax}} \; \log p(x|w) + \log p(w)$$

Assume $p(w) \sim \mathcal{N}(0, \tilde{\sigma}^2 I)$

Then $p(w) = \prod_{j=0}^{D} \mathcal{N}(w_j | 0, \tilde{\sigma}^2)$

$$\propto \prod_{j=0}^{D} \exp\left(-\frac{w_j^2}{2\tilde{\sigma}^2}\right)$$

That implies $\log p(w) = -\sum_{j=0}^{D} \frac{w_j^2}{2\tilde{\sigma}^2} + \text{const.}$

$$\hat{w}_{MAP} = \underset{w}{\operatorname{argmin}} \; \underbrace{\frac{1}{2}\|y - Xw\|^2 + \frac{\lambda}{2}\|w\|^2}_{E_{ridge}}$$

"ridge regression"

Setting $\dfrac{\partial E_{ridge}}{\partial w} = X^T(Xw - y) + \lambda w = 0$

yields $X^T X w + \lambda I w = X^T y$

That is $(X^T X + \lambda I)\, w = X^T y$

which gives $\underline{w = (X^T X + \lambda I)^{-1} X^T y =: \hat{w}_{ridge}}$

Next, let's compare $\hat{w}_{ridge}$ with $\hat{w}_{ls}$

Apply SVD to X, say $X = U \Delta V^T$,

where $\Delta = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_D)$ and $U^T U = V^T V = I$

Then $X^T = V \Delta U^T$ and

$$X^T X = V \Delta \underbrace{U^T U}_{I} \Delta V^T = V \Delta^2 V^T$$

That implies

$$X \hat{w}_{ridge} = U \Delta V^T (V \Delta^2 V^T + \lambda I)^{-1} V \Delta U^T y$$

$$= U \Delta V^T \left(V (\Delta^2 + \lambda I) V^T \right)^{-1} V \Delta U^T y$$

$$= U \Delta \underbrace{V^T V}_{I} (\Delta^2 + \lambda I)^{-1} \underbrace{V^T V}_{I} \Delta U^T y$$

$$= U \underbrace{\Delta (\Delta^2 + \lambda I)^{-1} \Delta}_{\text{diagonal}} U^T y \quad = U \Delta^2 (\Delta^2 + \lambda I)^{-1} U^T y$$

$$= \boxed{\sum_{j=0}^{D} u_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j^T y} \quad X \hat{w}_{ridge}$$

On the other hand,

$$X \hat{w}_{LS} = X (X^T X)^{-1} X^T y$$

$$= U \Delta \underbrace{V^T V}_{I} \Delta^{-2} \underbrace{V^T V}_{I} \Delta U^T y$$

$$= U \underbrace{\Delta \Delta^{-2} \Delta}_{I} U^T y$$

$$= U U^T y = \boxed{\sum_{j=0}^{D} u_j u_j^T y} \quad X \hat{w}_{LS}$$

" Ridge regression gives a shrunken version of estimation compared with regular linear regression. "

LASSO regression

$$\hat{w}_{LASSO} = \underset{w}{\operatorname{argmin}} \ \frac{1}{2} \| y - Xw \|^2 + \lambda \|w\|_1$$

where $\|w\|_1 = \sum_{j=0}^{D} |w_j|$

MAP estimator with a different prior

$$p(w) \propto \exp\left(-\alpha \sum_{j=0}^{D} |w_j| \right)$$

Laplacian prior

"No closed-form solution."

bias-variance decomposition

decomposition of expected loss

true output — our model

- In general, suppose we have a loss function $L(t, y(\mathbf{x}))$. The overall average expected loss is

$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

- In regression, the loss is given by

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

decomposition of expected loss

our model true model true output

- It does not hurt if we write

$$
\begin{aligned}
\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\
&= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2
\end{aligned}
$$

where the expectation is taken over t.

decomposition of expected loss

$$
\begin{aligned}
\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\
&= \underbrace{\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2}_{①} + \underbrace{2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\}}_{②} + \underbrace{\{\mathbb{E}[t|\mathbf{x}] - t\}^2}_{③}
\end{aligned}
$$

- Let's calculate

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

$$= \iint ① \, p(x,t)\, dx\, dt + \iint ② \, p(x,t)\, dx\, dt + \iint ③ \, p(x,t)\, dx\, dt$$

decomposition of expected loss

$$\{y(\mathbf{x}) - t\}^2 = \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2$$
$$= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2$$

1

- Let's calculate

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

First, $\iint (y(x) - \mathbb{E}[t|x])^2 p(x,t)\, dx\, dt$

$= \int \underbrace{\left(\int p(x,t)\, dt\right)}_{p(x)} (y(x) - \mathbb{E}[t|x])^2\, dx$

$= \int (y(x) - \mathbb{E}[t(x])^2 p(x)\, dx$

decomposition of expected loss

$$\{y(\mathbf{x})-t\}^2 = \{y(\mathbf{x})-\mathbb{E}[t|\mathbf{x}]+\mathbb{E}[t|\mathbf{x}]-t\}^2$$
$$= \{y(\mathbf{x})-\mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x})-\mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}]-t\} + \{\mathbb{E}[t|\mathbf{x}]-t\}^2$$

2

- Let's calculate

$$\mathbb{E}[L] = \iint \{y(\mathbf{x})-t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

Second,

$$2\iint \big(y(x)-\mathbb{E}[t|x]\big)\big(\mathbb{E}[t|x]-t\big)\,p(x,t)\,dx\,dt$$

$$= 2\iint \big(y(x)-\mathbb{E}[t|x]\big)\big(\mathbb{E}[t|x]-t\big)\,p(t|x)\,p(x)\,dx\,dt$$

$$= 2\int \big(y(x)-\mathbb{E}[t|x]\big)\Big(\int \big(\mathbb{E}[t|x]-t\big)\,p(t|x)\,dt\Big)\,p(x)\,dx$$

$$\int \big(\mathbb{E}[t|x]-t\big)\,p(t|x)\,dt = \mathbb{E}[t|x]-\int t\,p(t|x)\,dt = \mathbb{E}[t|x]-\mathbb{E}[t|x]$$

$$= 0$$

decomposition of expected loss

$$\{y(\mathbf{x}) - t\}^2 = \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2$$
$$= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \underbrace{\{\mathbb{E}[t|\mathbf{x}] - t\}^2}_{3}$$

- Let's calculate

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

third,

$$\iint (\underbrace{\mathbb{E}[t|x]}_{\text{true model}} - \underbrace{t}_{\text{true output}})^2\, p(x,t)\, dx\, dt = \text{"noise"}$$

decomposition of expected loss

- We have

$$\mathbb{E}[L] = \underbrace{\int \{\underset{\text{our model}}{y(\mathbf{x})} - \underset{\text{true model}}{h(\mathbf{x})}\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}}_{?} + \underbrace{\int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t}_{\text{noise}}$$

"unavoidable"

where

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\,\mathrm{d}t$$

- NOTE, our $y(\mathbf{x})$ depends on the dataset (sample) $\mathcal{D}$!!! Let's denote it as $y(\mathbf{x}) = y(\mathbf{x}; \mathcal{D})$ to explicitly emphasize this fact.

Remark: $\mathcal{D}$ was denoted as χ before

decomposition of expected loss

- We expand $\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2$ as

$$
\begin{aligned}
&\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad = \quad \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\qquad\quad +2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}.
\end{aligned}
$$

decomposition of expected loss

- We expand $\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2$ as

$$
\begin{aligned}
&\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad = \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\
&\qquad +2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}.
\end{aligned}
$$

- Taking the expectation over the ensemble of datasets, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] \\
&\quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}
\end{aligned}
$$

decomposition of expected loss

- Therefore, we have

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

$$
\begin{aligned}
(\text{bias})^2 &= \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\text{variance} &= \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\text{noise} &= \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t
\end{aligned}
$$

decomposition of expected loss

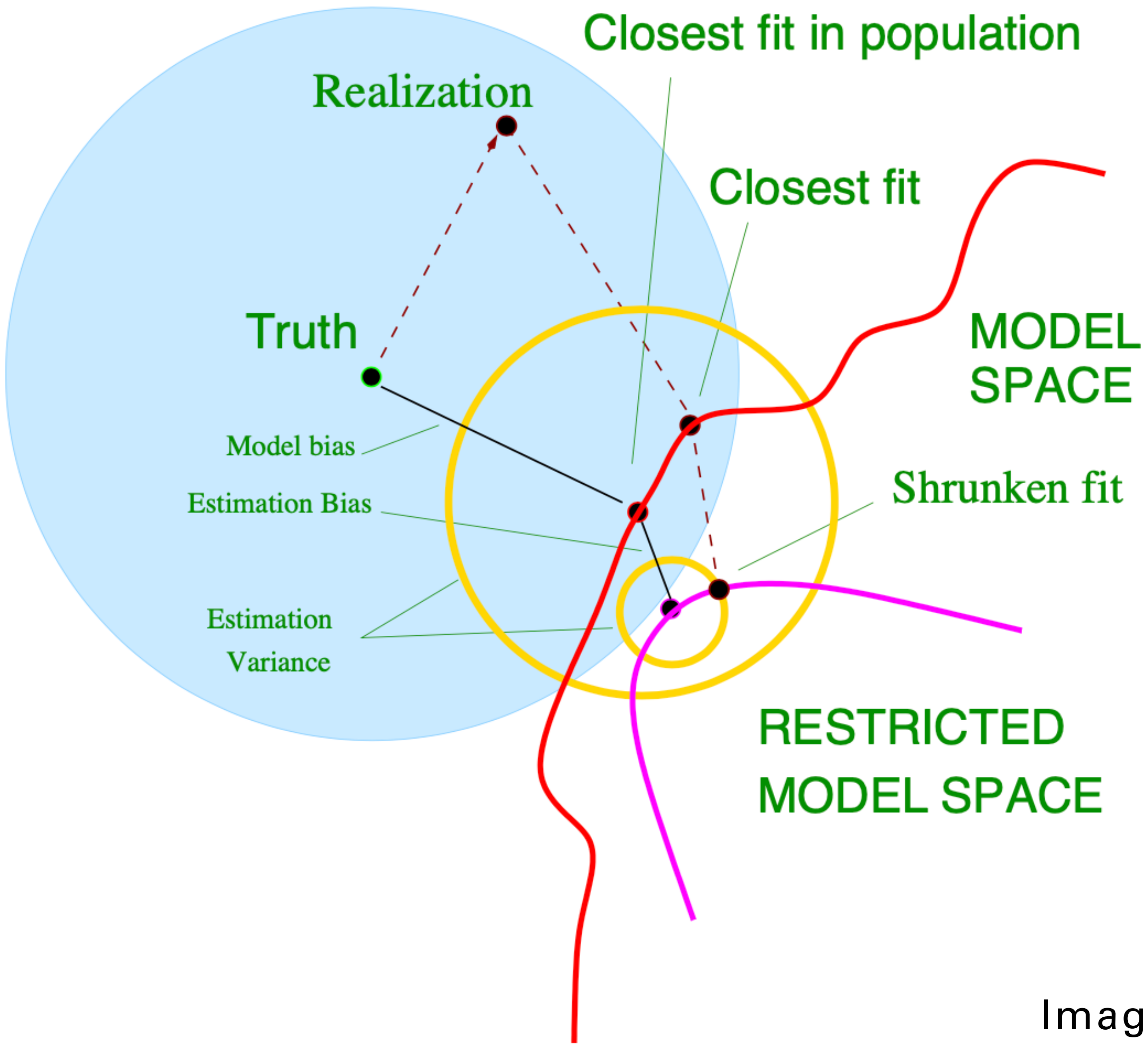

Image taken from [HaTF] Page 225

Questions?

Reference

- *Multivariate regression:*
    - *[Al] Ch.5.8*
    - *[HaTF] Ch.3.2*
    - [Bi] Ch.3.2 (bias-var)
- *Ridge and LASSO:*
    - *[HaTF] Ch.3.4*
- *Overview nonparametric methods:*
    - *[Al] Ch.8.1-8.3*
    - *[Bi] Ch.2.5.1*
- *Nonparametric classification and regression:*
    - *[Al] Ch.8.4-8.6, 8.8*
    - *[HaTF] Ch.6.6, 6.1.1-6.1.2*
    - *[Bi] Ch.2.5.2, 6.3.1*