Exercises 6

Curve fitting by linear smoothing

$y_i = f(x_i) + \epsilon_i$

(A) $y_i = \beta x_i + \epsilon_i$

Show that $\hat{y}^* = f(x^*) = \hat{\beta} x^*$

$\hat{f}(x^*) = \sum^{n} w(x_i, x^*) y_i$

From previous exercises,

$\hat{\beta} = (X^T X)^{-1} X^T y$ $\qquad \hat{\beta} \to p \times 1$

$= \dfrac{\sum^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum^{n} (x_i - \bar{x})^2}$

We already substracted means

$\hat{\beta} = \dfrac{\sum x_i y_i}{\sum x_i^2}$ $\qquad f(x^*) = \dfrac{\sum x_i y_i x^*}{\underbrace{\sum x_i^2}_{w}}$

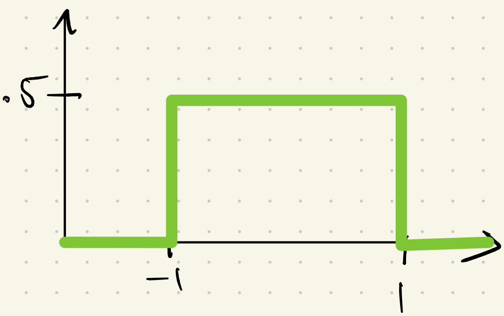
so $\boxed{w(x_i, x^*) = \dfrac{\sum^n x_i x^*}{\sum^n x_i^2}}$

KNN weights the k closest samples equally, whereas here we use all points in the smoother

B) $\int K(x)\,dx = 1 \qquad \int x K(x)\,dx = 0$

$\int x^2 K(x)\,dx > 0$

uniform: $K(x) = \frac{1}{2} I(x) \qquad I(x) \begin{cases} 1, & |x| \leq 1 \\ 0, & \text{otherwise} \end{cases}$



or $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ (Gaussian)

we define $w(x_i, x^*) = \frac{1}{h} K\left(\frac{x_i - x^*}{h}\right)$

h = bandwidth. Weights must be normalized to sum to 1.

write script to:

- simulate noisy data: $y = f(x) + \varepsilon$
- subtract mean from x & y
- fit the kernel smoother
- plot results for a range of bandwidths

$$\hat{y} = \sum_{i}^{n} w(x_i, x^*) y_i$$

$$\hat{y} = Hy \qquad H \equiv \text{matrix of weights}$$

$$\text{LOOCV} = \sum_{1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - H_{ii}} \right)^2 .$$

$$\left[\frac{y - \hat{y}}{1 - \text{diag}(H)} \right]^T \left[\frac{y - \hat{y}}{1 - \text{diag}(H)} \right]$$

Local polynomial regression

$y_i = f(x_i) + \epsilon_i$

$\hat{f}(x) = a = \text{argmin} \sum w_i (y_i - a)^2$

(A) $g_x(u; a) = a_0 + \sum_{k=1}^{D} a_j (u-x)^k$

u: neighbouring points
x: target point

$\hat{a} = \text{argmin} \sum w_i \{ y_i - g_x(x, a) \}^2$

$= \text{argmin} \sum w_i \left(y_i - a_0 - \sum_{1}^{D} a_j (u-x)^k \right)^2$

Define

$$R_x = \underbrace{\begin{bmatrix} (x_1 - x)^0 & (x_1 - x)^1 & \cdots & (x_1 - x)^D \\ \vdots & & & \\ (x_n - x_0)^0 & \cdots & & \end{bmatrix}}_{(n \times p)}$$

Let $W = \text{diag}(w_1 \dots w_n)$

$$\hat{a} = \text{argmin}\left[(y-Ra)^T W (y-Ra)\right]$$

$$\frac{\partial}{\partial a}(y-Ra)^T W (y-Ra) = 0$$

$$\frac{\partial}{\partial a}\Big(\cancel{y^T W y} - (Ra)^T W y - y^T W Ra$$

$$+ (Ra)^T W Ra\Big) = 0$$

$$\frac{\partial}{\partial a}\Big(\underbrace{-a^T R^T W y}_{\text{scalar}} - y^T W Ra + a^T R^T W Ra\Big) = 0$$

$$\frac{\partial}{\partial a}\Big(-2a^T R^T W y + a^T R^T W Ra\Big) = 0$$

$R \Rightarrow (n \times D)$

$$-2R^T W y + \left(R^T W R + (R^T W R)^T\right) a = 0$$

$$-2R^T W y + \left(R^T W R + R^T (R^T W)^T\right) a = 0$$

$$-2R^T W y + 2R^T W R a = 0$$

$$\boxed{a = \underbrace{(R^T W R)^{-1} R^T W}_{H \quad D \times n} y}$$

B) $\hat{f}(x) = \dfrac{\sum w_i(x) y_i}{\sum w_i(x)}$ For $D = 1$

Show:

where $w_i(x) = K\left(\dfrac{x - x_i}{h}\right)\{S_2(x) - (x_i - x)S_1(x)\}$

$S_j(x) = \sum K\left(\dfrac{x - x_i}{h}\right)(x_i - x)^j$

At the target point, $g_x(u = x; a) = \hat{a}_0$

Y tho?

$$\hat{f}(x) = \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{e_0^T} \left[\begin{bmatrix} 1 & \cdots & 1 \\ (x_1 - x) & \cdots & (x_n - x) \end{bmatrix} \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix} \begin{bmatrix} (x_1 - x) & \vdots \\ \vdots & \vdots \\ (x_n - x) & \vdots \end{bmatrix} \right]^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ (x_1 - x) & \cdots & (x_n \end{bmatrix} \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \end{bmatrix} \left\{ \begin{bmatrix} w_1 & w_2 & \cdots & w_n \\ w_1(x_1 - x) & \cdots & & w_n(x_1 - x \end{bmatrix} \begin{bmatrix} 1 & (x_1 - x) \\ 1 & \vdots \\ 1 & (x_n - x) \end{bmatrix} \right\}^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ (x_1 - x) & \cdots & (x_n \end{bmatrix} \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= [1 \; 0] \begin{bmatrix} \sum w_i & \sum w_i (x_i - x) \\ \sum w_i (x_i - x) & \sum w_i (x_i - x)^2 \end{bmatrix}^{-1}$$

$$\begin{bmatrix} w_1 \cdots w_n \\ w_1 (x_1 - x) \cdots \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \end{bmatrix}$$

$$= [1, 0] \times \frac{1}{\sum w_i \sum w_i (x_i - x)^2 - \left[\sum w_i (x_i - x)\right]^2}$$

$$\times \begin{bmatrix} \sum w_i (x_i - x)^2 & - \sum w_i (x_i - x) \\ \sum w_i (x_i - x) & \sum w_i \end{bmatrix}$$

$$\times \begin{bmatrix} \sum y_i w_i \\ \sum y_i w_i (x_i - x \end{bmatrix}$$

$$= \frac{\sum w_i (x_i - x)^2 \sum y_i w_i - \sum w_i (x_i - x) \sum y_i w_i (x_i - x)}{\sum w_i \sum w_i (x_i - x)^2 - \left(\sum w_i (x_i - x)\right)^2}$$

desired form

$$\hat{f}(x) = \frac{\sum w_i y_i (S_2 - (x_i - x) S_1)}{\sum w_i (S_2 - (x_i - x) S_1)}$$

where $S_1 = \sum w_i (x_i - x)$

$S_2 = \sum w_i (x_i - x)^2$

$$= \frac{\sum w_i (x_i - x)^2 \sum y_i w_i - \sum w_i (x_i - x) \sum y_i w_i (x_i - x)}{\sum w_i \sum w_i (x_i - x)^2 - \left(\sum w_i (x_i - x)\right)^2}$$

$$= \frac{\sum w_i y_i \left(\sum w_i (x_i - x)^2 - (x_i - x) \sum w_i (x_i - x)\right)}{\sum w_i \left(\sum w_i (x_i - x)^2 - (x_i - x) \sum w_i (x_i - x)\right)}$$

(c) $E(\hat{f}(x)) = E\left(\frac{\sum w_i^* y_i}{\sum w_i^*}\right)$ $y_i = f(x_i) + \epsilon_i$

$= E\left(\frac{w_1^* y_1 + w_2^* y_2 \dots}{w_1^* + w_2^* + w_3^*}\right)$ $E(y_i) = f(x_i)$

$= E\left(\frac{w_1^* y_1}{\sum w_i^*}\right) + E\left(\frac{w_2^* y_2}{\sum w_i^*} \dots\right)$

$= \frac{w_1^*}{\sum w_i^*} f(x_1) + \frac{w_2^*}{\sum w_i^*} f(x_2) \dots$ $\hat{f}(x) \to$ estimator

$= \boxed{\frac{\sum w_i^* f(x_i)}{\sum w_i^*}}$

$\hat{a}$ but isolate y

or? $E(\hat{f}(x)) = E(e_1 H y)$

$= e_1 H E(y) = \boxed{e_1 H f(x)}$

$H \to (1000 \times 100)$

$\text{var}(\hat{f}(x)) = \text{var}(e_1 H y)$

$= e_1 H \,\text{var}(y) (e_1 H)^T$ $H^T e_1^T$

$= \boxed{e_1 H \sigma^2 (e_1 H)^T}$

$(1 \times D)(D \times n)(n \times D)(D \times 1)$

(D)

$$E(y) = \mu$$
$$Var(y) = \Sigma = \sigma^2 I$$

$$r = y - \hat{y} = y - Hy$$

$$E(\hat{\sigma}^2) = E\left(\frac{\|r\|_2^2}{n - 2tr(H) + tr(H^T H)}\right)$$

$$= E\left(\frac{(y - Hy)^T (y - Hy)}{n - 2tr(H) + tr(H^T H)}\right)$$

$$= E\left(\frac{y^T y - 2y^T H y + y^T H^T H y}{n - 2tr(H) + tr(H^T H)}\right)$$

$$= \frac{E(y^T y) - 2E(y^T H y) + E(y^T H^T H y)}{n - 2tr(H) + tr(H^T H)}$$

$$= \frac{tr(\Sigma) + \mu^T \mu - 2tr(H\Sigma) - 2\mu^T H \mu + tr(H^T H \Sigma) + 2\mu^T H^T H \mu}{n - 2tr(H) + tr(H^T H)}$$

$$= \frac{tr(\sigma^2 I) + \mu^T \mu - 2tr(H\sigma^2 I) - 2\mu^T H \mu + tr(H^T H \sigma^2 I) + 2\mu^T H^T H \mu}{n - 2tr(H) + tr(H^T H)}$$

$$y = f(x) + \varepsilon \qquad E(y) = \mu = f(x)$$

$$= \frac{\operatorname{tr}(\sigma^2 I) + f(x)^T f(x) - 2\operatorname{tr}(H\sigma^2 I) - 2 f(x)^T H f(x) + \operatorname{tr}(H^T H \sigma^2 I) + 2 f(x)^T H^T H f(x)}{n - 2\operatorname{tr}(H) + \operatorname{tr}(H^T H)}$$

$$= \frac{\sigma^2 [n - 2\operatorname{tr}(H) + \operatorname{tr}(H^T H)] + f(x)^T [I - 2H + H^T H] f(x)]}{n - 2\operatorname{tr}(H) + \operatorname{tr}(H^T H)} \qquad \operatorname{tr}(I) = n$$

$$= \boxed{\sigma^2 + \frac{f(x)^T [(I-H)^T (I-H)] f(x)}{n - 2\operatorname{tr}(H) + \operatorname{tr}(H^T H)}}$$

$$\rightarrow [\underbrace{(I-H) f(x)}_{\text{bias vector}}]^T [(I-H) f(x)]$$

$\boxed{\text{Gaussian processes}}$

$[f(x_1) \cdots f(x_n)]$ has a multivariate normal

$\mathcal{X} \rightarrow$ set of indices

$f \sim GP(m, c)$

$E\{f(x_1)\} = m(x_1)$

$\text{cov}(f(x_1), f(x_2)) = c(x_1, x_2)$

(A) $c(x_1, x_2) = \tau_1^2 \exp\left(-\frac{1}{2} \left\{ \frac{d(x_1, x_2)}{b} \right\}^2 \right)$

$+ \tau_2^2 \delta(x_1, x_2)$

$\mathcal{X} \in [0, 1]$

B) $f \sim GP(m, c)$

I call it f because it's smooth. But it's like y data wo noise.

$P(f^* | f) = ?$

$f \sim \mathcal{N}(m, C)$

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m \\ m^* \end{bmatrix}, \begin{bmatrix} C(x, x) & C(x^*, x) \\ C(x^*, x)^T & C^*(x^*, x^*) \end{bmatrix} \right)$$

From exercises 1,

$$x_1 | x_2 \sim \mathcal{N}(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$$\boxed{\text{so } f^* | f \sim \mathcal{N}(m^* + C(x^*, x)^T C(x, x)^{-1} (f - m), C^*(x^*, x^*) - C(x^*, x)^T C(x, x)^{-1} C(x^*, x))}$$

(c) $y|\theta \sim \mathcal{N}(R\theta, \Sigma)$

$\theta \sim \mathcal{N}(m, V)$

$y = R\theta + \Sigma$

$\theta = m + V = \theta + 0$

$$\begin{bmatrix} y \\ \theta \end{bmatrix} = \begin{bmatrix} R \\ I \end{bmatrix} \theta + \begin{bmatrix} I \\ 0 \end{bmatrix} \Sigma$$

affine transformation of a MVN is also an MVN

$$\begin{bmatrix} f(x_1) \\ f(x_2) \end{bmatrix} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, C \right)$$

$\boxed{\text{In nonparametric regression \& spatial smoothing}}$

(A) $y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$

Prior $f \sim GP(0, C)$

derive $P(f \mid y)$

try to use previous proof

$$P(f \mid y) \propto P(y \mid f) P(f)$$

$$\propto \exp\left(-\frac{1}{2}(y-f)^T(\sigma^2 I)^{-1}(y-f)\right) \qquad m=0$$

$$\times \exp\left(-\frac{1}{2}(f-\cancel{m})^T (C)^{-1}(f-\cancel{m})\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}\left(\cancel{y^T y} - 2f^T y + f^T f\right) + f^T C^{-1} f\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[f^T\left[(\sigma^2)^{-1} I + C^{-1}\right] f - \frac{2f^T y}{\sigma^2}\right)\right)$$

we have $\boxed{\Sigma = \left((I\sigma^2)^{-1} + C^{-1}\right)^{-1}}$

$Km = y$

$\boxed{m = \frac{1}{\sigma^2} \Sigma y}$

$Km = \frac{y}{\sigma^2}$

(B) $f \sim GP(0, C)$

derive $E(f(x^*) | y)$

$m = \frac{\Sigma y}{\sigma^2}$

$y_i = f(x_i) + e_i$

$f(x) \sim N(m, C)$

$y = f(x) + \sigma^2 I = \overset{=0}{m} + C + \sigma^2 I$

$$\begin{bmatrix} y \\ f^* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}\right).$$

$$\sim N\left(0, \begin{bmatrix} C + \sigma^2 I & C(x, x^*) \\ C(x^*, x) & C(x^*, x^*) \end{bmatrix}\right)$$

As before,

$$\boxed{f^*|y = \mathcal{N}\left(C(x^*,x)\left[C(x,x)+\sigma^2 I\right]^{-1} y,\ C(x^*,x^*) - C(x^*,x)^T \left[C(x,x)+\sigma^2 I\right]^{-1} C(x^*,x) \right)}$$

The only difference is that for noisy data $\sigma^2 \neq 0$

we see that

$E(f^*|y) = Hy$ where

$$\boxed{H = C(x^*,x)\left[C(x,x)+\sigma^2 I\right]^{-1}}$$

D) $y_i = f(x_i) + \varepsilon_i \Rightarrow y \sim \mathcal{N}(f, \sigma^2 I)$

$f \sim GP(0, C)$

$$P(y) = \int P(y, f)\,df$$

$$= \int P(y|f)\,P(f)\,df$$

$$= \int \exp\left(-\frac{1}{2}\left[(y-f)^T(\sigma^2 I)^{-1}(y-f)\right.\right.$$

$$\left.+ f^T C^{-1} f\right]$$

$$= \int \exp\left(-\frac{1}{2}\left(y^T(\sigma^2 I)^{-1} y - 2f^T(\sigma^2 I)^{-1} y\right.\right.$$

$$\left. + f^T(\sigma^2 I)^{-1} f + f^T C^{-1} f\right)$$

$$= \int \exp\left(-\frac{1}{2}\left[y^T(\sigma^2 I)^{-1} y + f^T\left[(\sigma^2 I)^{-1}\right.\right.\right.$$

$$\left.+ C^{-1}\right] f - 2f^T(\sigma^2 I)^{-1} y\Big)$$

complete the square

$$(f - m^*)^T K (f - m^*)$$

$$= f^T K f - 2 f^T K m^* + m^{*T} K m^*$$

$$\boxed{\Sigma = \left[(\sigma^2 I)^{-1} + C^{-1}\right]^{-1}}$$

$$\Sigma^{-1} m^* = (\sigma^2 I)^{-1} y$$

$$\boxed{m^* = \Sigma (\sigma^2 I)^{-1} y}$$

MVN integrates to $\frac{1}{cst}$

so $$P(y) = \int \overbrace{\exp \frac{-1}{2} \left[(f - m^*)^T \Sigma^{-1} (f - m^*) \right.}^{\text{MVN integrates to } \frac{1}{cst}}$$

$$\left. \underbrace{- m^{*T} \Sigma^{-1} m^* + y^T (\sigma^2 I)^{-1} y}_{\text{does not depend on } f} \right] df$$

$$\propto \exp\left(\frac{-1}{2}\left(-m^{*T} \Sigma^{-1} m^* + y^T (\sigma^2 I)^{-1} y\right)\right)$$

$$\times \frac{1}{cst}$$ cst $\neq f(y)$

$$\propto \exp\left[-\frac{1}{2}\left(y^T (\sigma^2 I)^{-1} y - y^T (\sigma^2 I)^{-1} \Sigma (\sigma^2 I)^{-1} y \right)\right]$$

$$\propto \exp\left(-\frac{1}{2}\left[y^T (\sigma^2 I)^{-1} y - y^T \left[(\sigma^2 I)^{-1} \left((\sigma^2 I)^{-1} + C\right)^{-1} (\sigma^2 I)^{-1} y\right]\right.\right.$$

$$\propto \exp\left(-\frac{1}{2}\left(y^T \left[(\sigma^2 I)^{-1} - (\sigma^2 I)^{-1} \left((\sigma^2 I)^{-1} + C^{-1}\right)^{-1} (\sigma^2 I)^{-1} \right] y \right)\right.$$

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U (C^{-1} + V A^{-1} U)^{-1} V A^{-1}$$

$$U = V = 1$$

$$(A + C)^{-1} = A^{-1} - A^{-1} (C^{-1} + A^{-1})^{-1} A^{-1}$$

$$\Leftrightarrow \quad A^{-1} (C^{-1} + A^{-1}) A^{-1} = A^{-1} - (A + C)^{-1}$$

$$A = \sigma^2 I$$
$$C = C$$

$$\propto \exp\left(-\frac{1}{2} \left(y^T \left[\cancel{(\sigma^2 I)^{-1}} - \cancel{(\sigma^2 I)^{-1}} + (\sigma^2 I + C)^{-1}\right] y \right)\right.$$

$$\propto \exp\left(-\frac{1}{2} \left(y^T (\sigma^2 I + C)^{-1} y \right)\right.$$

$$\boxed{P(y) \sim \mathcal{N}(0, \sigma^2 I + C)}$$

(E) evaluate $P(y|\sigma^2, b)$

empirical Bayes to estimate σ^2

$f^*|y = Hy$

$$
\begin{array}{c|ccc}
 & x_1 & x_2 & x_3 \\
\hline
y_1 & & & \\
y_2 & & & \\
y_3 & & &
\end{array}
$$

EERC

not linear models

Cory

Deep L

Stat ML

fair & trans

MC

internet courses

Phil manuel

Deep learning

CS240

Machine