TRIAL EXAM 2011 SOLUTION

Problem 1 (Ch. 2.4 p. 18-)

~~a) L(X, Y) = loss f~~

a) $L(Y, f(X))$ = loss if we predict by $f(X)$ and Y is correct.

Squared error loss: $L(Y, f(X)) = (Y - f(X))^2$

Optimal?

Choose f by minimizing

$$EPE(f) = E(Y - f(X))^2$$

expected prediction error

$$= E_X E_{Y|X}[(Y - f(X))^2 | X]$$

So given $X = x$ we can let $f(X) = c$ where

$$c = \underset{c}{\operatorname{argmin}} \underbrace{E((Y-c)^2 | X=x)}_{g(c)}$$

Here: $g(c)$ is a $\cup$ parabola

$$g'(c) = E[-2(Y-c) | X=x] = 0$$

$$\Leftrightarrow c = E(Y|X=x). \quad \text{Q.E.D.}$$

b) $f(x) = x^T\beta$

Then can consider

$$EPE(\beta) = E(Y - X^T\beta)^2$$

Now set $\frac{\partial}{\partial\beta} EPE(\beta) = E\left[\frac{\partial}{\partial\beta}(Y - X^T\beta)^2\right] = 0$

$= E\left[\frac{\partial}{\partial\beta}\right]$.

On component form:

$$= E\left[\frac{\partial}{\partial\beta_j}\left(Y - \sum_{i=1}^{p} X_i\beta_i\right)^2\right]$$

$$= E\left[-2X_j\left(Y - \sum_{i=1}^{p} X_i\beta_i\right)\right] \quad ; j = 1, \ldots, p$$

So equation is

$$E[X_j Y] = E\left[X_j \sum_{i=1}^{p} X_i\beta_i\right] \quad ; j = 1, \ldots, p$$

or $E[XY] = E[XX^T\beta] = E[XX^T]\beta$

So $\beta = (E(XX^T))^{-1} E(XY)$ (see (2.16) p. 19)

c) Ch $Y \in \{1, 2, \ldots, k\}$.

$$L(Y, f(X)) = \begin{cases} 0 & \text{if } f(X) = Y \\ 1 & \text{if } f(X) \neq Y \end{cases}$$

$$EPE(f) = P[f(X) \neq Y]$$

$$= \sum_{x} \sum_{g=1}^{k}$$

$$= \sum_{x} \sum_{g=1}^{k} L(g, f(x)) \, P(X=x, Y=g)$$

$$= \sum_{x} \sum_{g=1}^{k} L(g, f(x)) \, P(X=x) \, P(Y=g \mid X=x)$$

$$= \sum_{x} P(X=x) \sum_{g=1}^{k} L(g, f(x)) \, P(Y=g \mid X=x)$$

$$= \sum_{x} P(X=x) \left(1 - P(Y=f(x) \mid X=x)\right)$$

$$= 1 - \sum_{x} P(Y=f(x) \mid X=x) \, P(X=x)$$

Thus - for each x let $f(x) = y$ where

$$y = \operatorname{argmax}_{y} P(Y=y \mid X=x)$$

d) Measures: Let $\mathcal{T}$ = training set

$$Err_{\mathcal{T}} = E[L(Y, \hat{f}(X)) \mid \mathcal{T}]$$

So $\hat{f}(\cdot)$ is a function of training data, while (X, Y) is drawn from joint distr.

If we take expectation over $\mathcal{T}$ also, we get

$$Err = E[Err_{\mathcal{T}}] = E[L(Y, \hat{f}(X))]$$

where randomness is (for fixed N)
① Draw training sample and compute $\hat{f}(\cdot)$
② Draw a new pair (X, Y).

Training error:

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

is too optimistic because same data is used to ~~esti~~ compute $\hat{f}(\cdot)$ as is used to measure fit.

Improvement: Estimate optimism:

Define $Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}[L(Y_i^0, \hat{f}(x_i)) \mid \mathcal{T}]$

where the same x_i as in training set $\mathcal{T}$ are used, but where Y_i^0 are drawn "at each x_i", i.e. from cond. distr. ~~given~~ of Y given $X = x_i$.

Define

$$op \equiv Err_{in} - \overline{err}$$

and $\omega \equiv E_y(op)$

where E_y is expected value over the training set with fixed $x_1, x_2, \dots, x_N$

For many loss functions we have

$$E_y(Err_{in}) = E_y(\overline{err}) + \underbrace{\frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)}_{\omega}$$

$$\Rightarrow E_y(\overline{err}) + 2 \cdot \frac{d}{N}\sigma_\varepsilon^2$$

often when model is $Y = f(X) + \varepsilon$

so we can estimate $\widehat{Err_{in}} = \overline{err} + 2 \cdot \frac{d}{N}\sigma_\varepsilon^2$

e) K-fold Cross-Validation

Divide data in K approx. equal sets ($K = 5$ or 10 usual)

Let $\kappa : \{1, \dots, N\} \to \{1, \dots, K\}$ be the indexing function

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

is the CV-estimate of prediction error.

Bootstrapping:

From traing data $((x_1,y_1),\dots,(x_N,y_N))$
draw with replacement N times:

$$((x_1^*,y_1^*),\dots,(x_N^*,y_N^*))$$

Do this B times!
Use these to compute $\hat{f}^{*b}(\cdot)$
Then measure of prediction error is e.g.

$$\hat{Err}^{(.632)} = 0.368\,\overline{err} + 0.632\cdot \hat{Err}^{(1)}$$

where

$$\hat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{C^{-i}}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

These are appropriate when there are too few data to have separate validation and test sets: Use all data as training data!

Problem 2

a) Cubic spline is given by:

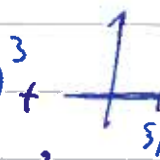
- K knots $\xi_1, \xi_2, \dots, \xi_K$
- Between knots given by 3rd order polynomials
- Continuous at knots
  Continuous first and second derivatives at knots.

[Natural Cubic spline: linear outside boundary knots.

Dimension of cubic splines:

K+1 regions with 4 coefficients	$4K+4$
$-$ K knots $\cdot$ 3 restrictions	$-3K$

$\Rightarrow$ $K+4$ basis functions needed.
These can be given as in exercise
(Must check that they are
- linearly independent
- requirements at knots:

Consider $g(x) = (x-\xi_k)_+^3$  So $g(\xi_k+) = g(\xi_k-) = 0$

$g'(x) = 3(x-\xi_k)_+^2$ $\quad g'(\xi_k+) = g'(\xi_k-) = 0$

$g''(x) = 6(x-\xi_k)_+$ $\quad g''(\xi_k+) = g''(\xi_k-) = 0.$

b) Natural requires

for $x < \xi_1$ is $f(x)$ linear

now the given $f(x)$ is for $x < \xi_1$

$$f(x) = \sum_{j=0}^{3} \beta_j \cdot x^j$$

so linear $\Rightarrow \beta_2 = \beta_3 = 0$.

For $x > \xi_k$ is

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \theta_k (x - \xi_k)^3 \qquad \text{(without + 's)}$$

So

$$f'(x) = \beta_1 + \sum_{k=1}^{K} \theta_k 3(x - \xi_k)^2$$

$$f''(x) = \sum_{k=1}^{K} 6\theta_k (x - \xi_k)$$

$$f'''(x) = \sum_{k=1}^{K} 6\theta_k$$

Hence: $f'''(x) = 0 \Rightarrow \sum_{k=1}^{K} \theta_k = 0$

$$f''(x) = 0 \Rightarrow \underbrace{\sum_{k=1}^{K} \theta_k x}_{=0} = \sum_{k=1}^{K} \theta_k \xi_k$$

$$\Rightarrow \sum_{k=1}^{K} \theta_k \xi_k = 0$$

c) For natural splines we need

$\underbrace{N+4}_{\text{cubic spline}} - 4 = N$ basis functions

(4 ↑ coeff. for x^2 and x^3 term are 0 outside both ξ_1 and ξ_K.)

Now note that

$$\sum_{k=1}^{K} \theta_k = 0 \Rightarrow \boxed{\theta_K = -\sum_{k=1}^{K-1} \theta_k} \quad (*)$$

$$\sum_{k=1}^{K} \xi_k \theta_k = 0 \Rightarrow \sum_{k=1}^{K-1} \xi_k \theta_k + \xi_K \theta_K = 0$$

$\Rightarrow$ using (*)

$$\sum_{k=1}^{K-1} (\xi_k - \xi_K)\theta_k = 0 \quad (**)$$

or $\sum_{k=1}^{K-2} (\xi_k - \xi_K)\theta_k = -(\xi_{K-1} - \xi_K)\theta_K$

Now under constraints in b) is

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \theta_k (x-\xi_k)_+^3 = \beta_0 + \beta_1 x + \sum_{k=1}^{K-1} \theta_k (x-\xi_k)_+^3 + \theta_K (x-\xi_K)_+^3$$

$$\overset{(*)}{=} \beta_0 + \beta_1 x + \sum_{k=1}^{K-1} \theta_k (x-\xi_k)_+^3 - \sum_{k=1}^{K-1} \theta_k (x-\xi_K)_+^3$$

$$= \beta_0 + \beta_1 x + \sum_{k=1}^{K-1} \theta_k \left[(x-\xi_k)_+^3 - (x-\xi_K)_+^3 \right]$$

$$= \beta_0 + \beta_1 x + \sum_{k=1}^{K-1} \theta_k (\xi_K - \xi_k) d_k(x)$$

~~But using (**) then~~

$$= \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \theta_k (\xi_K - \xi_k) d_k(x)$$

$$+ \theta_{K-1} (\xi_K - \xi_{K-1}) d_{K-1}(x)$$

~~$= \sum_{k=1}^{K-2} (\xi_k - \xi_K) \theta_{K-1} d_K$ by (**)~~

by (**) is this $= \sum_{k=1}^{K-2} (\xi_k - \xi_K) \theta_k d_{K-1}(x)$

$$= \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \theta_k (\xi_K - \xi_k) (d_k(x) - d_{K-1}(x))$$

$$= \beta_0 N_1(x) + \beta_1 N_2(x) + \sum_{k=1}^{K-2} \theta_k (\xi_K - \xi_k) N_{k+2}(x)$$

which gives a linear combination of the $N_1(\cdot), \ldots, N_K(\cdot)$ } Q.E.D.

Alternatively:

$$\beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \alpha_k N_{k+2}(x)$$

$$= \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \alpha_k [d_k(x) - d_{K-1}(x)]$$

$$= \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \alpha_k d_k(x) - d_{K-1}(x) \sum_{k=1}^{K-2} \alpha_k$$

$$= \beta_0 + \beta_1 x + \sum_{k=1}^{K-2} \frac{\alpha_k}{\xi_K - \xi_k} (x - \xi_k)_+^3 - \sum_{k=1}^{K-2} \frac{\alpha_k}{\xi_K - \xi_k} (x - \xi_K)_+^3$$

$$- \frac{(x - \xi_{K-1})_+^3}{\xi_K - \xi_{K-1}} \sum_{k=1}^{K-2} \alpha_k + \frac{(x - \xi_K)_+^3}{\xi_K - \xi_{K-1}} \sum_{k=1}^{K-2} \alpha_k$$

Coeff to θ_k

For $k \le K-2$: $\theta_k = \frac{\alpha_k}{\xi_K - \xi_k}$

$$\theta_{K-1} = -\left(\sum_{k=1}^{K-2} \alpha_k\right) \cdot \frac{1}{\xi_K - \xi_{K-1}}$$

$$\theta_K = -\sum_{k=1}^{K-2} \frac{\alpha_k}{\xi_K - \xi_k} + \frac{\sum_{k=1}^{K-2} \alpha_k}{\xi_K - \xi_{K-1}}$$

Now

$$\sum_{k=1}^{K}\theta_k = \sum_{k=1}^{K-2}\frac{\alpha_k}{\xi_K-\xi_k} - \frac{\sum_{k=1}^{K-2}\alpha_k}{\xi_K-\xi_{K-1}} - \text{def same} = 0$$

$$\sum_{k=1}^{K}\xi_k\theta_k = \sum_{k=1}^{K-2}\frac{\xi_k\alpha_k}{\xi_K-\xi_k} - \frac{\left(\sum_{k=1}^{K-2}\alpha_k\right)\xi_{K-1}}{\xi_K-\xi_{K-1}} - \left(\sum_{k=1}^{K-2}\frac{\alpha_k}{\xi_K-\xi_k}\right)\xi_K + \frac{\left(\sum_{k=1}^{K-2}\alpha_k\right)\xi_K}{\xi_K-\xi_{K-1}}$$

$$= \sum_{k=1}^{K-2}\frac{\alpha_k}{\xi_K-\xi_k}(\xi_k-\xi_K)$$

$$= -\sum_{k=1}^{K-2}\alpha_k$$

$$= 0$$