1 Questions from lecture notes

1. The key statistical assumption that we made to make the supervised inductive batch learning problem well-defined is:
TRAINING & DEPLOYMENT DATA ARE SAMPLES FROM SAME LIKELIHOOD.

[0.5 Mark]

2. The Bayes optimal classifier with 0-1 loss is the MODE of the underlying posterior likelihood (written as a function of input).

[0.5 Mark]

3. The component of the generalization error that is independent of the training set size is called: model error.

[0.5 Mark]

4. The formula for loss in logistic regression is given by: $l\,(\text{sign}(a), \text{sign}(b)) \equiv \log(1+e^{-ab})$.

[0.5 Mark]

5. Suppose the underlying likelihood in a linear regression problem satisfies the equations:

$$Y = \sum_{i=1}^{n} \alpha_i X_i + N, \ \mathbb{E}[NX] = 0, \tag{1}$$

where X, Y are the input,output. The necessary and sufficient conditions for parameters of a Bayes optimal linear regressor to be same as α are: $E[XX^T] \succ 0$.

[1 Mark]

6. From bias-variance tradeoff discussion in linear models it is clear that $n \to \infty, m \to \infty$ is sufficient for good generalization. Here, n, m are the number of parameters and training set size respectively. But should n diverge like $\sqrt{m}$ or like m or like m^2 ? $\sqrt{m}$. Fill in this blank with one of the three functions $\sqrt{m}, m, m^2$.

[0.5 Mark]

7. In applications of quantum information theory, the space (manifold) of positive definite (pd) matrices is often encountered and the standard loss function is the squared-Bures-Wasserstein metric: loss between A and B is given by $\left[\text{Tr } A + \text{Tr } B - 2\,\text{Tr}(A^{1/2}BA^{1/2})^{1/2}\right]$. While the standard loss on $\mathbb{R}$ is the squared-loss. Consider a regression problem with input space as that of pd matrices and real outputs, employing standard loss. The Bayes optimal, f^*, for this regression problem is defined by:

$$f^*(x) \equiv \underset{y\in\mathbb{R}}{\arg\min}\ E\left[(y-Y)^2/x\right].$$

Fill this blank with an expression involving the expression for the loss.

[1 Mark]

8. Consider a linear regression problem[1] where the underlying likelihood is such that $p(x, y) = p(x)p(y)$. Assume the mean and variance with $p(x)$ are 3,9 respectively. Assume the mean and variance with $p(y)$ are 4,36 respectively. Then, the simplified expression for the Bayes optimal linear regressor is:

$$f^*_{|\mathcal{L}}(x) = \underline{2/3}\,x.$$

Fill this blank with a numeric constant.

[1.5 Marks]

9. Consider a linear regression problem[2] with training data (input,output pairs): $\mathcal{D} = \{(1, 2), (3, 4)\}$. Then, the simplified expression for the ERM linear regressor is:

$$\hat{f}_{|\mathcal{L}}(x) = \underline{1.4}\,x.$$

Fill this blank with a numeric constant.

[1.5 Marks]

2 Derivations done in lectures

10. Derive a simplified expression for the Bayes optimal restricted to the linear model over inputs with squared loss using only projection theorem as done in lectures.

$w^* = \underset{w \in \mathbb{R}^n}{\operatorname{argmin}}\ E\{(w^T X - Y)^2\}$ $= \underset{w \in \mathbb{R}^n}{\operatorname{argmin}}\ \|w^T X - Y\|_{\mathcal{H}}^2$ $= \underset{w \in \mathbb{R}^n}{\operatorname{argmin}}\ \|w^T E\{XX^T\}^{1/2} \tilde{X} - Y\|_{\mathcal{H}}$ by $\tilde{X} = E\{XX^T\}^{-1/2} X$	$= E\{XX^T\}^{-1/2}$ times $\underset{\bar{w} \in \mathbb{R}^n}{\operatorname{argmin}}\ \|\bar{w}^T \tilde{X} - Y\|$ $= E\{XX^T\}^{-1/2} E\{\tilde{X} Y\}$ $= E\{XX^T\}^{-1} E\{XY\}$ Proj. theorem.

[2 Marks]

3 Problems from other course page resources

11. Consider a problem where the underlying likelihood is defined by $p(x|y) \sim \mathcal{N}(\mu_y, \Sigma_y), y = 1, 0$. Let $p(y) = \begin{cases} 0.4 & y = 1 \\ 0.6 & y = 0 \end{cases}$. Assume the loss is the 0-1 loss. Derive a simplified expression for the Bayes optimal.

[1]feature map $\phi(x) = x$.
[2]feature map $\phi(x) = x$.

$p(1/x) \propto e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) + \ln 0.4 - \frac{1}{2}\ln\lvert\Sigma_1\rvert}$ $p(0/x) \propto e^{-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0) + \ln 0.6 - \frac{1}{2}\ln\lvert\Sigma_0\rvert}$	$f^*(x) = 1$ if $\text{sign}(g(x)) \leq 0$, 0 else $g(x) = \frac{1}{2}x^T(\Sigma_1^{-1} - \Sigma_0^{-1}) - x^T(\Sigma_1^{-1}\mu_1 - \Sigma_0\mu_0)$ $\ln 0.6/0.4 + \frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1 - \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0 - \frac{1}{2}\ln\frac{\lvert\Sigma_0\rvert}{\lvert\Sigma_1\rvert}$

[1 Mark]

If $\Sigma_1 = \Sigma_0$, then show that the model error with linear model over the feature map, $\phi(x) = \begin{bmatrix} x \\ 1 \end{bmatrix}$, is exactly zero.

$g(x)$ becomes $-x^T(\Sigma_1^{-1}\mu_1 - \Sigma_0\mu_0) + \ln 0.6/0.4 + \frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1 - \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0 - \frac{1}{2}\ln\frac{\lvert\Sigma_0\rvert}{\lvert\Sigma_1\rvert}$ so linear

[0.5 Mark]

Prove that this condition is not necessary for the model error being zero[3]. In other words, provide a different condition on $\mu_1, \mu_0, \Sigma_1, \Sigma_0$, such that the model error is exactly zero.

model error is zero iff quadratic reduces to an affine or linear or constant in terms of its sign.   For alternate, $\Sigma_1 \neq \Sigma_0$, i.e. sign of quadratic be a constant $\iff$ discriminant ≤ 0, $x^TAx + 2b^Tx + c \geq 0 \; \forall x \iff c - b^TA^{-1}b \geq 0$. So Alternate is (A is invertible $A \succ 0$) $\Sigma_1^{-1} - \Sigma_0^{-1} \succ 0$, $\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0 + 2\ln\frac{0.6}{0.4} - \ln\lvert\Sigma_0\rvert/\lvert\Sigma_1\rvert$ $-(\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_0)^T(\Sigma_1^{-1} - \Sigma_0^{-1})^{-1}(\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1) \geq 0$

[4 Marks]

[3] is the solution provided for the practice problems partially wrong? Go home and think about the necessary and sufficient conditions for 0 model error. If you think you got them, meet me to discuss.

INDIAN INSTITUTE OF TECHNOLOGY HYDERABAD

ADDITIONAL SHEET

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Roll No. :

1 First line on Pg 12 of summary notes

2 First line, second para, Pg 16 of notes

3 First line, second para, Pg 20 of notes

4 Second last line, Second last para, Pg 32 of notes

5 last para Pg 23 - first para Pg 24 of notes

or

6 Second last para in Pg 23 of notes

6 ~~Necessity in notes~~

$x_1, \ldots x_n$ are lin independent $\Leftrightarrow$ $E[xx^T] \succ 0$

Marks to either answer.

6 Based on Eqn. (10.2) and (11.1) in notes

$\frac{n}{m} \to 0$ $\quad \frac{\sqrt{m}}{m} \to 0$ $\quad \frac{m}{m} \to ?$ $\quad \frac{m^2}{m} \to \infty$

7 Eqn (6.1) in notes . Substitute square loss.

8 Eqn (8.7) $w^* = E[xx^T]^{-1} E[xy]$ from notes

$E[xy] = E[x]E[y]$ $\quad (\because x \perp\!\!\!\perp y)$

$= 3 \times 4 = 12$

$$E\{xx^T\} = E\{x^2\} = var(x) + (E\{x\})^2$$
$$= 9 + 3^2 = 18$$

$\therefore w^* = \frac{12}{18} = \frac{2}{3}$ $\quad \therefore f^*_L(x) = w^{*T} x$
$$= \frac{2}{3} x$$

(9) From notes (8.8) $\hat{w} = \left(\frac{1}{m}\sum_i x_i x_i^T\right)^{-1}\left(\frac{1}{m}\sum_i x_i y_i\right)$

$$= \left(\frac{1}{2}(1\times1 + 3\times3)\right)^{-1}\left(\frac{1}{2}(1\times2 + 3\times4)\right)$$

$$= \frac{14}{10} = 1.4$$

$\therefore f_L(x) = \hat{w}^T x = 1.4x.$

(10) Theorem 3 on Pg 3 in hand-written derivations.

(11) $p(y/x) \propto p(x/y)p(y)$

$$p(1/x) \propto e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)} \; 0.4 / |\Sigma_1|^{1/2}$$

$$= e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) + \ln 0.4 - \frac{1}{2}\ln|\Sigma_1|}$$

$$\text{IIIly } p(0/x) \propto e^{-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0) + \ln 0.6 - \frac{1}{2}\ln|\Sigma_0|}$$

$$p(1/x) \geq p(0/x) \iff -\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) + \ln 0.4 - \frac{1}{2}\ln|\Sigma_1|$$

$$\geq -\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0) + \ln 0.6 - \frac{1}{2}\ln|\Sigma_0|$$

$$\therefore f^*(x) = \begin{cases} 1 & \text{if } \frac{1}{2} x^T(\Sigma_1^{-1} - \Sigma_0^{-1})x - x^T(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0) + \frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1 - \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0 + \ln 0.6/0.4 - \frac{1}{2}\ln |\Sigma_0|/|\Sigma_1| \leq 0 \\ 0 & \text{else} \end{cases}$$

If $\Sigma_1 = \Sigma_0$, then

$$f^*(x) = \begin{cases} 1 & \text{if } \underbrace{-x^T(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0)}_{w^*} + \underbrace{\frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1 - \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0 + \ln\frac{0.6}{0.4} - \frac{1}{2}\ln\frac{|\Sigma_0|}{|\Sigma_1|}}_{b_0^*} \leq 0 \\ 0 & \text{else} \end{cases}$$

i.e. $f^*(x) = -\text{sign}(-w^{*T}x + b_0^*)$ $(-1 \to 0)$

$\therefore$ Model error is zero.

Can model error be zero → quadratic → linear/affine
→ quadratic → const.

Can sign of quadratic behave like constant without $\Sigma_1 = \Sigma_0$?

Yes! if its a non-negative quadratic or non-positive quadratic } Discriminant ≤ 0

② marks if you write this.

INDIAN INSTITUTE OF TECHNOLOGY HYDERABAD

ADDITIONAL SHEET

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Roll No. :

$x^T A x + 2b^T x + c \geq 0 \quad \forall x$

$\Leftrightarrow \| A^{1/2} x + A^{-1/2} b \|^2 + c - b^T A^{-1} b \geq 0 \quad \forall x$ (assuming $A \succ 0$)

If you are interested in knowing what happens if $A \succeq 0$ then please meet me.

or if A is symmetric only.

$\Leftrightarrow c - b^T A^{-1} b \geq 0$

$\therefore$ Alternate condition is

$$\Sigma_1^{-1} - \Sigma_0^{-1} \succ 0, \quad \mu_1^T \Sigma_1^{-1} \mu_1 - \mu_0^T \Sigma_0^{-1} \mu_0 + 2 \ln \frac{0.6}{0.4} - \ln \frac{|\Sigma_0|}{|\Sigma_1|} - (\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1)^T (\Sigma_1^{-1} - \Sigma_0^{-1})^{-1} (\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1) \geq 0$$

or ....

See Assignment Week 3 for comprehensive answers.