CS 189 Introduction to Machine Learning
Fall 2017 HW8

This homework is due **Friday, October 20 at 10pm.**

1 Getting Started

You may typeset your homework in latex or submit neatly handwritten and scanned solutions. Please make sure to start each question on a new page, as grading (with Gradescope) is much easier that way! Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, "HW[n] Write-Up"
2. Submit all code needed to reproduce your results, "HW[n] Code".
3. Submit your test set evaluation results, "HW[n] Test Set".

After you've submitted your homework, be sure to watch out for the self-grade form.

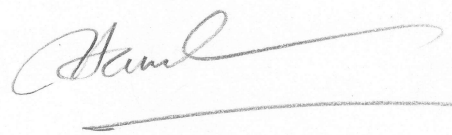
(a) Before you start your homework, write down your team. Who else did you work with on this homework? List names and email addresses. In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

None. I worked alone
Comments : Need the lecture notes for homework

(b) Please copy the following statement and sign next to it:

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

I certify that all solutions are entirely my words & that I have not looked at another student's solutions. I have credited all external sources in this write up.



Problem # 2

a.) $P(x|w_1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}$ $\quad P(x|w_2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}$

Decision boundary:

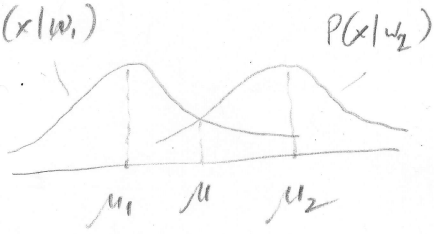
$$P(x|w_1)\,P(w_1)^{1/2} = P(x|w_2)\,P(w_2)^{1/2}$$

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}$$

$$\Rightarrow (x-\mu_1)^2 = (x-\mu_2)^2$$

$$\Rightarrow x-\mu_1 = \mu_2 - x$$

$$\Rightarrow x = \frac{\mu_1+\mu_2}{2}$$



Decision boundary

Decision rule: $x < \frac{\mu_1+\mu_2}{2} \rightarrow x \in$ category 1

$x > \frac{\mu_1+\mu_2}{2} \rightarrow x \in$ category 2.

$x = \underbrace{\frac{\mu_1+\mu_2}{2}}_{\mu} \rightarrow x$ can be of category 1 or 2

b.) $P_e = P((\text{misclassified as } w_1)|w_2)\,P(w_2) + P((\text{misclassified as } w_2)|w_1)\,P(w_1)$

$$= \frac{1}{2}\int_{\mu}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx + \frac{1}{2}\int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}} dx$$

A B

for A: let $z = \frac{x-\mu_1}{\sigma} \Rightarrow dz = \frac{1}{\sigma} dx$

$$x = \mu \Rightarrow z = \frac{\mu-\mu_1}{\sigma} = \frac{\mu_2-\mu_1}{2\sigma} = a$$

$$\Rightarrow A = \frac{1}{2}\int_a^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

for B: let $z = \frac{x-\mu_2}{\sigma} \Rightarrow dz = \frac{1}{\sigma} dx$

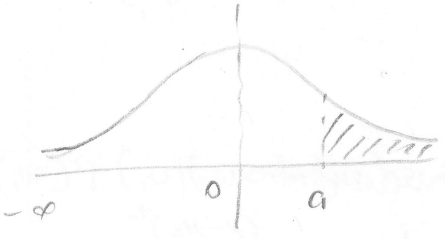
$x = \mu \Rightarrow z = \frac{\mu-\mu_2}{\sigma} = \frac{\mu_1-\mu_2}{2\sigma} = -a$

$$\Rightarrow B = \frac{1}{2}\int_{-\infty}^{-a} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

$$= \frac{1}{2}\int_a^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \quad (\text{since } e^{-\frac{z^2}{2}} \text{ is even function})$$

$$\Rightarrow P_e = A + B = \int_a^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

c.) The integrand of P_e is $\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ of the standard normal distribution form



P_e is the area under the curve (shaded area)

if $a \to \infty \Rightarrow P_e \to 0$

Problem # 3

a.) Gaussian prior w/ smaller variance

b.) TLS allows errors in X & y. OLS only allows errors in y.

c.) Convex functions:

$$f_1(x) = \max\{-x, 0.1x\}$$

$$f_2(x) = x + \frac{x^2}{10}$$

$$f_4(x) = \frac{e^x + e^{-x}}{2} - 1$$

d)

$$\begin{bmatrix} 1 & -0.1 \\ -0.5 & 1 \end{bmatrix}$$

e.) Training error Bias

f.) Variance

g) Validation error

Problem #4

a.) $\det(\Sigma) \geq 0$

$4 - a^2 \geq 0$

$-2 \leq a \leq 2$

b.) 2 $\left.\begin{matrix} 1 \\ 1 \end{matrix}\right\}$ for (x, y, z) for (v_x, v_y, v_z)

Problem #5

$$P(x=x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

(a) $P(x_1, x_2, \dots, x_n \mid \lambda) = \prod_1^n P(x_i \mid \lambda)$ (x_i's are i.i.d)

$$\Rightarrow \ln P(x_1, x_2, \dots, x_n \mid \lambda) = \sum_1^n \ln P(x_i \mid \lambda)$$

$$= \sum_1^n \ln \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \sum_1^n (\ln \lambda^{x_i} + \ln e^{-\lambda} - \ln x_i!)$$

$$= \sum_1^n (x_i \ln \lambda - \lambda - \ln x_i!)$$

$$= \ln \lambda \sum_1^n x_i - n\lambda - \sum_1^n \ln x_i!$$

(b) Consider $f(\lambda) = -\ln P = n\lambda - \ln \lambda \sum_1^n x_i - \sum_1^n \ln x_i!$

$$f'(\lambda) = n - \frac{1}{\lambda} \sum_1^n x_i$$

$$f''(\lambda) = \frac{1}{\lambda^2} \sum_1^n x_i > 0 \quad \forall \lambda \neq 0$$

$\Rightarrow f(\lambda)$ is convex

$\ln P$ max iff $f(\lambda)$ min

$$\Rightarrow f'(\lambda) = 0 \Rightarrow \lambda = \frac{1}{n} \sum_1^n x_i$$

(c)

$$P(\lambda \mid x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n \mid \lambda) P(\lambda)}{P(x_1, \dots, x_n)}$$

$$= \frac{\prod_1^n P(x_i \mid \lambda) P(\lambda)}{\prod_1^n P(x_i)} \leftarrow \text{constant}$$

$$\Rightarrow \ln P(\lambda | x_1, \dots, x_n) = \ln \prod_1^n P(x_i|\lambda) + \ln P(\lambda) - \ln \prod_1^n P(x_i) \quad \text{(const)}$$

$$= \ln \lambda \sum_1^n x_i - n\lambda - \sum_1^n x_i! + \underbrace{\ln \alpha e^{-\alpha\lambda}}_{\ln\alpha - \alpha\lambda}$$

$\ln P(\lambda | x_1, \dots, x_n)$ max iff $g(\lambda) = -\ln P(\lambda | x_1, \dots, x_n)$ min

$$g(\lambda) = (n+\alpha)\lambda - \ln\lambda \sum_1^n x_i + \sum_1^n x_i! - \ln\alpha$$

$$g'(\lambda) = (n+\alpha) - \frac{1}{\lambda}\sum_1^n x_i = 0$$

$$\Rightarrow \lambda = \frac{1}{n+\alpha}\sum_1^n x_i$$

$\lambda_{MAP} < \lambda_{MLE}$ since α appears in the denominator

$$n \to \infty : \quad \alpha = \frac{\sum_1^n x_i / n}{1 + \underbrace{\frac{\alpha}{n}}_{\to 0}} = \mu \quad \text{where } \mu = \frac{1}{n}\sum_1^n x_i$$

$\alpha \to$ the mean of samples

Thus $\lambda_{MAP} \to \lambda_{MLE}$ when $n \to \infty$

Problem # 6

(a) OLS w/ X' & y' : $\frac{1}{2}\|X'w' - y'\|_2^2 \leftarrow \min$

the close form solution : $w' = (X'^T X')^{-1} X'^T y'$

Transform OLS w/ X' & y' into Ridge Regression w/ X & y :

$$\frac{1}{2}\|X'w' - y'\|_2^2 = \frac{1}{2}\|(X_e + C)w' - y'\|^2 \qquad X' = \begin{bmatrix} X \\ ce_1^T \\ \vdots \\ ce_d^T \end{bmatrix} = \underbrace{\begin{bmatrix} X \\ 0^T \\ \vdots \\ 0^T \end{bmatrix}}_{X_e} + \underbrace{\begin{bmatrix} 0 \\ ce_1^T \\ \vdots \\ ce_d^T \end{bmatrix}}_{C}$$

$$= \underbrace{\frac{1}{2}\|X_e w' - y'\|_2^2}_{= \frac{1}{2}\|Xw' - y\|_2^2} - \underbrace{\frac{1}{2} \times 2 \times (X_e w' - y')^T C w'}_{= [(Xw'-y)^T 0 \dots 0] \begin{bmatrix} 0 \\ \vdots \\ 0 \\ ce_1^T w' \\ \vdots \\ ce_d^T w' \end{bmatrix} \to 0} + \underbrace{\frac{1}{2}\|Cw'\|_2^2}_{\frac{c^2}{2}\|w'\|_2^2}$$

since the last d rows are 0

$$= \frac{1}{2}\|Xw' - y\|_2^2 + \frac{c^2}{2}\|w'\|_2^2 \leftarrow \text{ridge regression}$$

Now transform w' into w :

$$w' = \underbrace{\left([X^T ce_1 \dots ce_d] \begin{bmatrix} X \\ ce_1^T \\ \vdots \\ ce_d^T \end{bmatrix} \right)^{-1}}_{(X_e^T + C^T)(X_e + C)} \underbrace{[X\ ce_1^T \dots ce_d^T] \begin{bmatrix} y \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{Xy}$$

$$= X_e^T X_e + X_e^T C + C^T X_e + C^T C$$

we know that :

$$X_e = \begin{bmatrix} X \\ 0^T \\ 0^T \\ \vdots \\ 0^T \end{bmatrix} \qquad C = \begin{bmatrix} 0 \\ ce_1^T \\ \vdots \\ ce_d^T \end{bmatrix} \qquad (X_e + C = X')$$

$$\Rightarrow X_e^T X_e = X^T X \qquad X_e^T C = C^T X_e = 0 \qquad C^T C = c^2 I_d$$

$\Rightarrow \quad w' = (X^T X + c^2 I_d)^{-1} X y$

compared to $\hat{w} = (X^T X + \lambda I_d)^{-1} X y$

$\Rightarrow \quad c^2 = \lambda \quad$ or $\quad c = \sqrt{\lambda}$

Thus, we can convert Ridge regression into OLS by extending X and y as in the question because the ridge regression problem is equivalent to the extended OLS and the rigde regression solution is equivalent to the OLS solution with $c = \sqrt{\lambda}$ as we prove above.

(b) Ridge regression $\rightarrow$ OLS

$$\min_{w'} f(w') = \min_{w'} \frac{1}{2} \| X'w' - y' \|_2^2$$

the step size of gradient descent for the above OLS will be

$$\gamma' = \frac{2}{\lambda_{min}(X'^T X') + \lambda_{max}(X'^T X')}$$

we know from (a) that : $X'^T X' = X^T X + c^2 I$

$$= X^T X + \lambda I$$

If λ_i is the eigen value of $X^T X$, then $\lambda_i + \lambda$ is the eigen value of $X^T X + \lambda I = X'^T X'$ since

$$X^T X \vec{v} = \lambda_i \vec{v} \quad \forall \vec{v}$$

and $\quad \lambda I \vec{v} = \lambda \vec{v} \quad \forall \vec{v}$

$$\Rightarrow (X^T X + \lambda I) \vec{v} = (\lambda_i + \lambda) \vec{v} \quad \forall i$$

Thus $\lambda_{max}(X'^T X') = \lambda_{max}(X^T X) + \lambda$

$= M + \lambda$

$\lambda_{min}(X'^T X') = \lambda_{min}(X^T X) + \lambda$

$= m + \lambda$

$\Rightarrow$ the stepsize for the extended OLS (ridge regression)

is $\gamma' = \frac{1}{\lambda_{max}(X'^T X') + \lambda_{min}(X'^T X')} = \frac{1}{M + m + 2\lambda}$

Problem #7

$$Y = M^* + N$$

$N_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0,1) \Rightarrow \Sigma_N = I$

$$\Rightarrow P(N) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)^{1/2}}} \exp\left(-\frac{1}{2}(N-0)^T \Sigma^{-1} (N-0)\right)$$

($\Sigma = I$; $\det(\Sigma)^{1/2} \to 1$)

$$\Rightarrow P(Y|M) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}(Y-M)^T(Y-M)\right)$$

$$P(Y|M) = \prod_{i,j}^{d} P(y_{ij}|M) \quad \text{(since } N_{ij} \text{ iid so } y_{ij} \text{ iid)}$$

$$\Rightarrow \log P(Y|M) = \sum_{i,j}^{d} \ln \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}(y_{ij}-m_{ij})^2\right)$$

$$= \sum_{i,j}^{d} \left(-\frac{1}{2}(y_{ij}-m_{ij})^2 - \ln \sqrt{(2\pi)^d}\right)$$

($\ln \sqrt{(2\pi)^d}$ ← const)

$$= -\frac{1}{2}\|Y-M\|_F^2 + \text{const.}$$

$\Rightarrow$ $\max \ln P(Y|M)$ is equivalent to $\min \|Y-M\|_F^2$

i.e $$\hat{M} = \underset{M \in \mathcal{C}}{\operatorname{argmin}} \|Y-M\|_F^2$$

(b) Applying the Eckart-Young theorem directly

$$\| Y - M_k \|_F \leq \| Y - M \|_F$$

thus

$$\hat{M} = \underset{M : \text{rank}(M) = d-1}{\arg\min} \| Y - M \|_F = M_{d-1}$$

where $M_{d-1} = \sum_{i=1}^{d-1} \sigma_i \vec{u}_i \vec{v}_i^T$ σ_i is the $(d-1)$ largest eigenvalues of YY^T

$$Y = U \Sigma V^T$$

i.e

$$\Sigma_{d-1} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_{d-1} \\ & & & & 0 \end{bmatrix}$$

that is we decompose Y into $U \Sigma V^T$. Then take $(d-1)$ largest singular value of Y and corresponding $\vec{u}_i, \vec{v}_i$ to built $M_k = \sum_{i=1}^{d-1} \sigma_i \vec{u}_i \vec{v}_i^T$

c.) $Y = U \Sigma V^{-1}$

$M_k = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T$ where σ_i is the k largest singular values of Y and $\vec{v}_i$, $\vec{u}_i$ are the corresponding singular vectors.

Problem #8

(a)

$$C = \Sigma_{xx}^{-1/2} \Sigma_{x\mathcal{Q}} \Sigma_{\mathcal{Q}\mathcal{Q}}^{-1/2}$$

$$\hat{x} = U^T \Sigma_{xx}^{-1/2} x \qquad \hat{\mathcal{Q}} = V^T \Sigma_{\mathcal{Q}\mathcal{Q}}^{-1/2} \mathcal{Q}$$

$$E[\hat{x}\hat{\mathcal{Q}}^T] = E[U^T \underbrace{\Sigma_{xx}^{-1/2} x \mathcal{Q}^T \Sigma_{\mathcal{Q}\mathcal{Q}}^{-1/2}}_{C = U\Lambda V^T} V]$$

$$= E[U^T U \Lambda V^T V] = E[\Lambda] = \Lambda$$

$$E[\hat{x}\hat{x}^T] = E[U^T \Sigma_{xx}^{-1/2} x x^T \Sigma_{xx}^{-1/2} U]$$

$$= U^T \Sigma_{xx}^{-1/2} \underbrace{E[xx^T]}_{I} \Sigma_{xx}^{-1/2} U$$

$$= U^T \underbrace{\Sigma_{xx}^{-1}}_{I} U = U^T I U = I$$

$$E[\hat{\mathcal{Q}}\hat{\mathcal{Q}}^T] = E[V^T \Sigma_{\mathcal{Q}\mathcal{Q}}^{-1/2} \mathcal{Q} \mathcal{Q}^T \Sigma_{\mathcal{Q}\mathcal{Q}}^{-1/2} V]$$

$$= V^T \Sigma_{\mathcal{Q}\mathcal{Q}}^{-1/2} \underbrace{E[\mathcal{Q}\mathcal{Q}^T]}_{I} \Sigma_{\mathcal{Q}\mathcal{Q}}^{-1/2} V$$

$$= V^T \Sigma^{-1/2} \Sigma^{-1/2} V = V^T I V = I$$

b.)

$$E[(Y - w^T\hat{X})^2] = E[(Y - w^T\hat{X})^T (Y - w^T\hat{X})]$$

$$= E[Y^TY - Y^Tw^T\hat{X} - \hat{X}^TwY + \hat{X}^Tww^T\hat{X}]$$

$$= E[Y^TY] - E[Y^Tw^T\hat{X}] - E[\hat{X}^TwY] + E[\hat{X}^Tww^T\hat{X}]$$

$E[Y^2]$ $\quad w^T E[Y\hat{X}^T]$ $\quad w E[\hat{X}^TY]$ $\quad \|w\|_2^2 \underbrace{E[\hat{X}^TX]}_{I}$

$$E[Y^TX] = E[\hat{X}^TY] = E[Y\hat{X}]$$

Since $Y \in \mathbb{R}$

$$\Rightarrow E[(Y - w^T\hat{X})^2] = E[Y^2] - 2w^T E[Y\hat{X}] + \|w\|_2^2.$$

c.) $\hat{w} = \arg\min E[(Y - w^T\hat{X})^2] + \|w\|_{LCA}^2$

$$= \arg\min E[Y^2] - 2w^T E[Y\hat{X}] + \|w\|_2^2 + \sum_1^d \frac{1-\lambda_i}{\lambda_i}(w_i)^2$$

$\|w\|_2^2 + \sum_1^d \frac{1}{\lambda_i}(w_i)^2 - \|w\|_2^2$

$f(\vec{w})$

$$f'_{w_i}(w) = 0 - 2E[Y(X)_i] + 2w_i \frac{1}{\lambda_i} = 0$$

$$\Rightarrow w_i = \lambda_i E[Y(X)_i]$$

(d)

$$E[\|\tilde{w} - \hat{w}\|_2^2] = E[\|\tilde{w} - E(\tilde{w})\|_2^2]$$

$$= \sigma(\tilde{w}) =$$

$$\tilde{w}_i = \lambda_i \frac{1}{n} \sum_{j=1}^{n} y^j (x^j)_i$$

$$\sigma(\tilde{w}) = E[\tilde{w}^2] + \cancel{E[\tilde{w}]^2 \to 0}$$

$$= \sum_1^d \lambda_i^2 \frac{1}{n^2} \left(\sum_1^n y^j (x^j)_i\right)^2$$

$$= \sum_1^d \lambda_i^2 \frac{1}{n^2} n Y^2$$

$$= \sum_1^d \lambda_i^2 \frac{1}{n} Y^2 \qquad Y^2 \le 1$$

$$\le \sum_1^d \frac{\lambda_i^2}{n}$$

$$\Rightarrow E[\|\tilde{w} - \hat{w}\|_2^2] \le \sum_1^d \frac{\lambda_i^2}{n}$$

Problem # 9

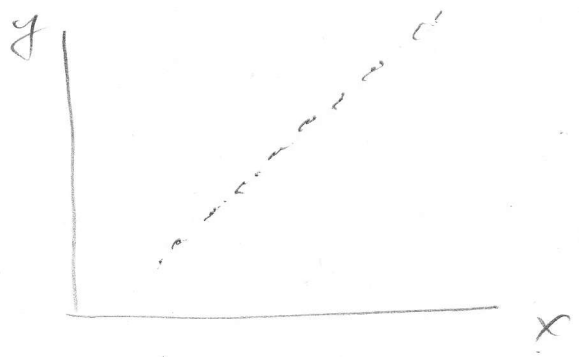
Multiple choice

a.) Select function convex over $\mathbb{R}$ (3 pts)

(i) $f(x) = x^3$ (iii) $f(x) = \cosh(x)$

(ii) $f(x) = \tanh x$ (iv) $f(x) = \sinh(x)$

b) Give the best covariance matrix of the diagram (3 pts)



(c) True or false? (4 pts)

(i) PCA can be use in case of noise with large variance?

(ii) CCA is effected by noise variance?

Solution

a.) $f(x) = \sinh(x)$

b) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

c.) (i) False (ii) False