12011923 张旭宇.

problem 1. a). $Y=AX+V \Rightarrow V=Y-AX$:

the least square solution of X is to minize $\|Y-AX\|^2_{Q^{-1}}=J$.

$$J=\|Y-AX\|^2_{Q^{-1}}=(Y-AX)^TQ^{-1}(Y-AX)=(Y^T-X^TA^T)Q^{-1}(Y-AX)$$

$$\frac{\partial J}{\partial X^T}=\frac{\partial(Y^TQ^{-1}Y-X^TA^TQ^{-1}Y-Y^TQ^{-1}AX+X^TA^TQ^{-1}AX)}{\partial X^T}=0-A^TQ^{-1}Y-A^TQ^{-1}Y+2A^TQ^{-1}AX=0$$

$$\Rightarrow 2A^TQ^{-1}AX=2A^TQ^{-1}Y \Rightarrow X^*=(A^TQ^{-1}A)^{-1}A^TQ^{-1}Y.$$

b) $L=(Y-AX)^TQ^{-1}(Y-AX)+\lambda(b^TX-c) \rightarrow$ Lagrangian functions.

$$\frac{\partial L}{\partial X^T}=-A^TQ^{-1}Y-A^TQ^{-1}Y+2A^TQ^{-1}AX+\lambda b=0 \Rightarrow A^TQ^{-1}AX=A^TQ^{-1}Y-\frac{\lambda}{2}b$$

$$\Rightarrow X^*=(A^TQ^{-1}A)^{-1}(A^TQ^{-1}Y-\frac{\lambda}{2}b).\ \text{let } \frac{\lambda}{2}=\lambda_1 \Rightarrow X^*=(A^TQ^{-1}A)^{-1}(A^TQ^{-1}Y-\lambda_1 b)$$

c) Lagrangian functions: $L=(Y-AX)^TQ^{-1}(Y-AX)+\lambda_1(b^TX-c)+\lambda_2(X^TX-d)$

$$\frac{\partial L}{\partial X^T}=-A^TQ^{-1}Y-A^TQ^{-1}Y+2A^TQ^{-1}AX+\lambda_1 b+2\lambda_2X=0$$

$$\Rightarrow 2(A^TQ^{-1}A+\lambda_2)X=2A^TQ^{-1}Y+\lambda_1 b \Rightarrow (A^TQ^{-1}A+\lambda_2)X=A^TQ^{-1}Y+\frac{\lambda_1}{2}b$$

$$\Rightarrow X^*=(A^TQ^{-1}A+\lambda_2)^{-1}(A^TQ^{-1}Y+\frac{\lambda_1}{2}b),\ \text{let } \frac{\lambda_1}{2}=\lambda_1^* \Rightarrow X^*=(A^TQ^{-1}A+\lambda_2)^{-1}(A^TQ^{-1}Y+\lambda_1^*b)$$

problem 2: $X\sim N(x|m_0,\Sigma_0) \Rightarrow AX\sim N(Ax|Am_0,A\Sigma_0A^T)$

~~Given~~ $Y=AX+V$. given X.. the expectation of Y is $AX+E[V]=AX+0=AX$.

a) the variance of Y is $V[X]+V[V]=0+\beta^{-1}I \Rightarrow p(Y|X)\sim N(Y|AX,\beta^{-1}I)$.

b) $P(Y,X)=P(Y|X)\cdot P(X)=N(Y|AX,\beta^{-1}I)\cdot N(X|m_0,\Sigma_0)$ let $z=\begin{pmatrix}x\\y\end{pmatrix}$

consider the log of the joint distribution:

$$\ln p(z)=\ln p(x)+\ln p(y|x)=-\frac{1}{2}(x-m_0)^T\Sigma_0^{-1}(x-m_0)-\frac{1}{2}(Y-AX)^T(\beta^{-1}I)^{-1}(Y-AX)+\text{const}.$$

consider the second order which can be written as:

$$-\frac{1}{2}x^T(\Sigma_0^{-1}+\beta A^TA)x-\frac{1}{2}\beta x^TY+\frac{1}{2}\beta Y^TAx+\frac{1}{2}\beta x^TA^TY=-\frac{1}{2}\begin{pmatrix}x\\y\end{pmatrix}^T\begin{bmatrix}\Sigma_0^{-1}+\beta A^TA & -\beta A^T\\ -\beta A & \beta\end{bmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

$$\text{cov}(z)=R^{-1}=\begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}=\begin{bmatrix}\Sigma_0 & \Sigma_0A^T\\ A\Sigma_0 & A\Sigma_0A^T+\beta^{-1}I\end{bmatrix}.$$

identify the linear term to find the mean $\therefore E(Z) = \begin{pmatrix} m_0 \\ Am_0 \end{pmatrix}$

$$\therefore p(Y,X) \sim N\left[\begin{pmatrix} m_0 \\ Am_0 \end{pmatrix}, \begin{pmatrix} \Sigma_0 & \Sigma_0 A^T \\ A\Sigma_0 & A\Sigma_0 A^T + \beta^{-1} I \end{pmatrix}\right].$$

(c). $p(Y) \sim N(Y \mid Am_0, A\Sigma_0 A^T + \beta^{-1} I)$.

(d). let $X = HY + U$, then $p(X|Y) \sim N(X|HY, L)$. $p(X|Y) \propto p(Y|X)p(X)$.

$$-\frac{1}{2}(X-HY)^T L^{-1}(X-HY) \propto -\frac{1}{2}(Y-AX)^T(\beta^{-1}I)^{-1}(Y-AX) - \frac{1}{2}(X-m_0)^T\Sigma_0^{-1}(X-m_0).$$

$$\Rightarrow \begin{cases} L^{-1} = \beta A^T A + \Sigma_0^{-1} \\ HY = L\,(A^T(\beta^{-1}I)^{-1}Y + \Sigma_0^{-1}m_0) = L(\beta A^T Y + \Sigma_0^{-1} m_0). \end{cases}$$

$$\Rightarrow p(X|Y=y, \beta, m_0, \Sigma_0) \sim N\left[(\beta A^T A + \Sigma_0^{-1})^{-1}(\beta A^T Y + \Sigma_0^{-1} m_0), (\beta A^T A + \Sigma_0^{-1})^{-1}\right].$$

(e). $p(\hat{Y}|Y=y, \beta, m_0, \Sigma_0) = p(\hat{Y}|X)\cdot p(X|Y=y, \beta, m_0, \Sigma_0) = N(\hat{Y}|AX, \beta^{-1}I)\cdot N(m_{MAP}, \Sigma_{MAP})$.

$\Rightarrow p(\hat{Y}|Y=y, \beta, m_0, \Sigma_0) = N(\hat{Y}|Am_{MAP}, A\Sigma_{MAP}A^T + \beta^{-1}I)$.

$\cancel{m_{MAP} =}$ $\Sigma_{MAP} = (\beta A^T A + \Sigma_0^{-1})^{-1}$. $m_{MAP} = \Sigma_{MAP}(\beta A^T Y + \Sigma_0^{-1} m_0)$.

(f). $\because Y = AX + V$. $\Rightarrow$ $\cancel{p(\hat{Y},Y) \sim N(0,}$

problem III. posterior distribution = likelihood function × the prior.

Assume $p(w|D, \beta, m_0, \alpha) \sim N(m_{MAP}, \Sigma_{MAP})$.

$$P_{MAP} = \prod_{n=1}^{N} p(y_n|w, \phi_n, \beta)\cdot p(w)_{prior}.$$

$$\Rightarrow \log P_{MAP} = \sum_{n=1}^{N}(y_n - w^T\phi_n)^T\beta^{-1}(y_n - w^T\phi_n) + (w-m_0)^T(\alpha^{-1}I)^{-1}(w-m_0) = (w-m_{MAP})^T\Sigma_{MAP}^{-1}(w-m_{MAP})$$

$$\Rightarrow \Sigma_{MAP}^{-1} = (\alpha^{-1}I)^{-1} + \sum_{n=1}^{N}\phi_n^T\beta^{-1}\phi_n. \quad m_{MAP} = \Sigma_{MAP}\left[(\alpha^{-1}I)^{-1}m_0 + \beta\sum_{n=1}^{N}\phi_n y_n\right]$$

$$\Rightarrow \Sigma_{MAP} = \left[(\alpha^{-1}I)^{-1} + \sum_{n=1}^{N}\phi_n^T\beta^{-1}\phi_n\right]^{-1} \quad m_{MAP} = \Sigma_{MAP}\left[(\alpha^{-1}I)^{-1}m_0 + \beta\sum_{n=1}^{N}\phi_n y_n\right]$$

$$p(\hat{y}|\hat{x}, D, \beta, m_0, \alpha) \sim N(\phi_n m_{MAP}, \phi_n\Sigma_{MAP}\phi_n^T + \beta^{-1}).$$

problem IV. posterior distribution = likelihood function × the prior

$$p(w|D,m_0,\alpha)=p(t|w)p(w)=\prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}\cdot N(w|m_0,\alpha^{-1}I).$$

$$\ln p(w|t)=-\frac{1}{2}(w-m_0)^T(\alpha^{-1}I)^{-1}(w-m_0)+\text{const}.$$

W_{MAP} can be obtained by maximizing the posterior distribution, which means minimizing error.

$$E=\frac{1}{2}(w-m_0)^T(\alpha^{-1}I)^{-1}(w-m_0)-\sum_{n=1}^{N} t_n\ln y_n+(1-t_n)\ln(1-y_n).$$

$$S_N^{-1}=\cancel{-\nabla\nabla\ln p}\ -\nabla\nabla E=(\alpha^{-1}I)^{-1}+\sum_{n=1}^{N} y_n(1-y_n)\phi_n^T\phi_n,\quad W_{MAP}=S_N\Big[(\alpha^{-1}I)^{-1}m_0+\sum_{n=1}^{N}\phi_n y_n^2(1-y_n)\Big].$$

$$\Rightarrow p(w|D,m_0,\alpha)\sim N(w|W_{MAP},S_N).$$

for posterior predictive distribution. $p(t_1|x,D,m_0,\alpha)=\int p(t_1|x,w)\cdot p(w|D,m_0,\alpha)\,dw\ \cancel{=\int\sigma(w^T\phi(x))p}$

$$\approx\int\sigma(w^T\phi(x))\,p(w|D,m_0,\alpha)\,dw.$$

let $a=w^T\phi(x)$. $\sigma(w^T\phi)=\int\delta(a-w^T\phi(x))\sigma(a)\,da$, δ is delta function.

$$\int\sigma(w^T\phi(x))p(w|D,m_0,\alpha)\,dw=\int\sigma(a)\,p(a)\,da,\quad p(a)=\int\delta(a-w^T\phi(x))p(w|D,m_0,\alpha)\,dw.$$

Then calculating each moment and switch the integration of a and w.

$$\mu_a=E[a]=\int p(a)\,a\,da=\int p(w|D,m_0,\alpha)\,w^T\phi(x)\,dw=W_{MAP}^T\phi(x).$$

$$\sigma^2(a)=Var[a]=\int p(a)\{a^2-E[a]^2\}\,da=\int p(w|D,m_0,\alpha)\{[w^T\phi(x)]^2-[W_{MAP}^T\phi(x)]^2\}\,dw=\phi(x)S_N\phi(x)$$

$$p(w^T\phi(x)|t)=\int\sigma(w^T\phi(x))\,\cancel{p(a)}\,p(w^T\phi(x))\,d[w^T\phi(x)]=\int\sigma(w^T\phi(x))N(w^T\phi(x)|\mu_a,\sigma^2(a))\,da$$

$$p(t_2|x,D,m_0,\alpha)=1-p(t_1|x,D,m_0,\alpha).$$

problem V. (1) $\cancel{y=\sigma(a_2)}\ \cancel{\frac{\partial y}{\partial a_2}=}$ $y=\sigma(a_2)=\frac{1}{1+e^{-a_2}}$, $\frac{\partial y}{\partial a_2}=\frac{e^{-a_2}}{(1+e^{-a_2})^2}=\frac{1}{1+e^{-a_2}}\cdot\frac{e^{-a_2}}{1+e^{-a_2}}=y(1-y)$

$$\frac{\partial y}{\partial w^{(2)}}=\frac{\partial y}{\partial a_2}\frac{\partial a_2}{\partial w^{(2)}}=y(1-y)z,\quad \frac{\partial y}{\partial a_1}=\frac{\partial y}{\partial a_2}\frac{\partial a_2}{\partial z}\frac{\partial z}{\partial a_1}=y(1-y)w^{(2)}h'(a_1)$$

$$\frac{\partial y}{w^{(1)}}=\frac{\partial y}{\partial a_2}\frac{\partial a_2}{\partial z}\frac{\partial z}{\partial a_1}\frac{\partial a_1}{\partial w^{(1)}}=y(1-y)w^{(2)}h'(a_1)x$$

$$\frac{\partial y}{\partial x}=\frac{\partial y}{\partial a_1}\frac{\partial a_1}{\partial x}=y(1-y)w^{(2)}h'(a_1)w^{(1)}$$

(2) $L=-\log(y^t(1-y)^{1-t})\ \cancel{=\log t}\ =-[t\log y+(1-t)\log(1-y)]$. $\frac{\partial L}{\partial y}=-\left[\frac{t}{y}+\frac{(1-t)}{1-y}\right]=\frac{y-t}{y(1-y)}$

for $w^{(2)}$: $\frac{\partial L}{\partial w^{(2)}}=\frac{\partial L}{\partial y}\frac{\partial y}{\partial w^{(2)}}=\frac{y-t}{y(1-y)}\cdot y(1-y)z=(y-t)z$. $w^{(2)new}=w^{(2)old}-\alpha\cdot z(y-t)$.

for $w^{(1)}$: $\frac{\partial L}{\partial w^{(1)}}=\frac{\partial L}{\partial y}\frac{\partial y}{\partial w^{(1)}}=\frac{y-t}{y(1-y)}\cdot y(1-y)w^{(2)}h'(a_1)x=(y-t)w^{(2)}h'(a_1)x$.

$$w^{(1)new}=w^{(1)old}-\alpha(y-t)w^{(2)}h'(a_1)x.$$

problem VI: a) Assume $p(t|x)$ has the property of Gaussian distribution, which can be affected by noise.

$p(t|x,w,\beta) = N(t|y(x,w), \beta^{-1})$, $p(w|D,\beta,m_0,\alpha) \propto p(w|m_0,\alpha) \cdot p(D|w,\beta)$.

Using Laplace approximation: $\ln p(w|D,\beta,m_0,\alpha) = -\frac{1}{2}(w-m_0)^T(\alpha^{-1}I)^{-1}(w-m_0) - \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,w)-t_n\}^2(\beta^{-1})^{-1}$ + cons

$= -\frac{1}{2}(w-m_0)^T\alpha(w-m_0) - \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,w)-t_n\}^2$ + cons.

$\Rightarrow \Sigma^{-1} = (\alpha^{-1}I)^{-1} + \beta H = \alpha I + \beta H$. $W_{map} = \Sigma[(\alpha^{-1}I)^{-1}m_0 + \sum_{n=1}^{N} H y_n(1-y_n)]$

$\rightarrow p(w|D,\beta,m_0,\alpha) \sim N(w|W_{map}, \Sigma)$.

for posterior predictive distribution, $p(t|x,D,\beta,m_0,\alpha) = \int p(t|x,w)\, p(w|D,\beta,m_0,\alpha)\, dw$

$y(x,w) \simeq y(x,W_{map}) + g^T(w-W_{map})$, $g = \nabla_w y(x,w)\big|_{w=W_{map}}$.

$p(t|x,w,\beta) \simeq N(t|y(x,W_{map}) + g^T(W-W_{map}), \beta^{-1})$; $p(t|D,\beta,m_0,\alpha) = N(t|y(x,W_{map}), \sigma^2(x))$

$\sigma^2(x) = \beta^{-1} + g^T A^{-1} g = \beta^{-1} + g^T(\alpha I + \beta H)^{-1} g$

b). $p(w|D,\alpha) = p(D|w) \cdot p(w|\alpha)$. $p(D|w) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}$, $\ln p(D|w) = \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$

W_{map} can be obtained by maximizing the posterior distribution, which ~~can be~~ means minizing error.

$E(w) = -\ln p(D|w) + \frac{1}{2}w^T(\alpha^{-1}I)^{-1}w = \frac{1}{2}w^T\alpha w - \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$

$A^{-1} = -\nabla\nabla[E(w)] = -\nabla(-\nabla E(w)) = -\nabla(-\alpha w + \beta H w) = \alpha I + \beta H$.

$W_{map} = A[(\alpha^{-1}I)^{-1}m_0 + \sum_{n=1}^{N} H y_n(1-y_n)] \Rightarrow p(w|D,\alpha) \sim N(w|W_{map}, A)$.

for posterior predictive distribution, $p(t|x,D,\alpha) \simeq p(t|x,W_{map})$.

$a(x,w) \simeq a_{map}(x) + b^T(w-W_{map})$. $b^T = \nabla a(x,W_{map})$, $\sigma_a^2(x) = b^T(x)A^{-1}b(x)$.

$p(t|D,\alpha) = N(\sigma|a(W_{map},x), b^T A^{-1} b)$.

problem VII a).i).

	SVM	logistic regression.
loss function.	$\frac{1}{2}\|W\|^2$	$\sum_{n=1}^{N} t_n\log y_n + (1-t_n)\log(1-y_n) + \frac{\lambda}{2}\|W\|^2$
restrictions	$t_n[W^T\phi(x_n)+b] \geq 1$. $n=1,\cdots,N$.	no.
analysis.	ensure correct classification through hard restrictions.	put the correctness of results in the loss function.
prediction.	$W^T\phi(x_n)+b \begin{cases} \geq 1 & \text{positive} \\ \leq -1 & \text{negative} \end{cases}$	$\sigma[y(W,x)]$, $\rightarrow$ the probability.
linear or not.	without kernel function, ~~is~~ linear classification.	can do non-linear classification.

(ii). ν-SVM | least square regression

	ν-SVM	least square regression
loss function	$\frac{1}{2}\|\|w\|\|^2 + C\sum_{n=1}^{N} E_\varepsilon(y_n - t_n)$	$\frac{1}{2}[y-(w^Tx+b)]^2 + \frac{1}{2}\lambda w^T w$.
restriction	$t_n \le y(x_n) + \varepsilon + \xi_n$ $t_n \ge y(x_n) - \varepsilon - \xi_n$ $\xi_n \ge 0$	no.
prediction	$w^T x + b$	$w^T x + b$.
Analysis	E_ε function of ν-SVM can be various, which can prevent outliner's effect.	it is hard to handle the outliner.

b). $f(x) = \frac{1}{1+e^{-x}}$. ① differentiable ② $f(x) \in (0,1)$, which is suitable for two-class classification.
③ can handle the outliner. ④ convert natural parameters into Bernouli parameters,
⑤ $f'(x) = f(x)[1-f(x)]$.

(c) (i). sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$, which is suitable for probability output.
sigmoid function and tanh function can encounter the problem of gradient dispersion or disappearance because $f(x)[1-f(x)] \le \frac{1}{4}$, when $f(x) \le 1$.. relu function is easy to compute, but dead neuron will appear when it's less than 0.

(ii). sigmoid function is commonly used for two-class classification.
tanh function and sigmoid function are used in RNN model.
ReLU function is widely used in the hidden layers of MLP, CNN. and transformer model.

(d). Jacobian: $J_{ki} = \frac{\partial y_k}{\partial x_i}$. Jacobian matrix is used to measure the sensitivity of the current model to inputs. The higher the value, the less ability to recognize the current input. Because the output will change dramatically in a certain direction ~~of~~ if the input changes inslightly. The value of Jacobian matrix of a good model is small.

Hessian matrics: $\frac{\partial^2 E}{\partial w_{ji} \partial w_{lk}}$, which is used for SGD, $x^* = x^0 - H^{-1} J^T$.

(e). Exponential family distribution is memory-free. $P(T>t+s|T>t) = P(T>s)$. This is a Poisson distribution process that allows problem to be computed in engineering practice. It can be considered that the observed variables are independently and identitily distribution and using maximum likelihood method to compute. T-distribution and f-distribution aren't exponential family distribution.

(f) The essence of ML is bring Pdata and Pmodel as close as possible

KL divergence can ~~measure~~ measure the similarity of two distributions. And it has a huge effect on dimensionality reduction.. What's more, KL divergence is used in loss function.

$D_{KL}(p\|q) = \sum_{i=1}^{N} p(x_i)\log\left[\frac{p(x_i)}{q(x_i)}\right]$, ~~$p(x_i)q(x_i)$~~ $p(x_i)(\log p(x_i))$ is fixed. ~~$H(p,q)$~~

$H(p,q) = -\sum_{i=1}^{n} p(x_i)\log q(x_i)$ means cross entropy.

(g) The purpose of regularization skills for NNs is to make neural network has good generalization while keeping easy. This means that the small $\frac{\partial y}{\partial x}$, the stronger the ability to recognize inputs. Data augmentation techniques can prevent models from learning too fine-grained.

(h) Considering, $Y = AX + V$, X, V is Gaussian distribution, then $p(Y|X)$, $p(X,Y)$, $p(Y)$ are also Gaussian distribution, which is very easy to analyze. Also, Gaussian distribution belongs to exponential family distribution, which means it has the same properties of exponential family distribution.

(i). For MAP model, there exists prior distribution, $p(w|D) \propto p(D|w)\,p(w)$. The existence of prior distribution makes the w of model fit a distribution, which can prevent value of w from being abnormal.

Problem VIII. discriminative approach is to calculate $p(C|X)$.
generative approach is to calculate $\frac{p(X|C)p(C)}{p(X)}$ or $p(X,C)$.

discriminative approach is to calculate the differences between different classes, while generative approach is to learn peculiarity of each class.

Generative approach: advantages: ~~only a sm~~ requiring only a small amount of data to drive model.

disadvantages: the accuracy isn't very high. And prior distribution $p(c)$ is required. This does not apply to cases where the prior distribution of some samples differs greatly from the actual prior distribution.

Discriminative approach: advantages: high accuracy, can define peculiarity.

disadvantages: the features of data can't be missing too much. limited range of application.

Example: determine whether it is a spam email, count the frequency of each word, including $x_1, x_2, \cdots x_N$

Discriminative approach: ~~that~~ input is the frequency of each word X, two MLP Layers + sigmoid output probability.

Generative approach: train two models, determine whether it is a spam email.