Name: Xun Xue
UNI: xx2241

Problem C

Stochasic Gradient Descent (SGD) mean that we can update the value of parameters each time we calculate a data instead of after we calculate the whole dataset. Here is the peusdo code:

for $i=0,\dots N$ do:

  Initialize d_i randomly and let $t=1$, let $A = X[0]^T X[0]$

  while ($t \leq T$ & stopping condition is not True) do:

    for $j=0,\dots,m$ do:

      $y \leftarrow d - \eta \nabla_{d_{ij}} (-d_{ij}^T A[j]^T A[j] d_{ij})$

      $d_{ij} \leftarrow \frac{y}{\|y\|}$

      $t \leftarrow t+1$

    $\lambda \leftarrow d_{ij}^T X[j]^T X[j] d_{ij}$

    $A \leftarrow X[j]^T X[j] - \sum_{k=0}^{j} \lambda_k d_{ij} d_{ij}^T$

Name: Xun Xue
UNI: xx2241

problem 2.

(i) $\because \int P(x) dx = \int P(y) dy$ and $y = g(x)$

$\therefore \int f(x) dx = \int \frac{1}{255} d g(x)$

$\therefore f(x) dx = \frac{1}{255} d(g(x))$

$\therefore dg(x) = 255 f(x) dx = \frac{255}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$

$\therefore g(x) = \int_0^x \frac{255}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$

(ii)

$P(x) = \int_0^1 \int_0^1 8xyz \, dy dz = 2x$

$P(y) = \int_0^1 \int_0^1 8xyz \, dx dz = 2y$

$P(z) = \int_0^1 \int_0^1 8xyz \, dx dy = 2z$

$E(xyz) = \int_0^1 \int_0^1 \int_0^1 8xyz \cdot x \cdot y \cdot z \, dx dy dz$

$= \int_0^1 \int_0^1 \int_0^1 8x^2y^2z^2 \, dx dy dz$

$= \frac{8}{27}$

$P(xy) = \int_0^1 8xy \, dz = 4xy$

$P(x) \cdot P(y) = 2x \cdot 2y = 4xy = P(xy)$

$\therefore$ x and y conditionally indepent given z.

Problem e

Name: Xun Xue
UNI: xx2241

1. $\mu_{MAP} = \arg\max_{\mu} P(\mu|X)$

$$P(\mu|X) = \frac{P(X|\mu)P(\mu)}{P(X)}$$

$$P(X|\mu) = P(X^{(1)}, X^{(2)}, \dots X^{(m)}|\mu) = \prod_{i=1}^{m} P(X^{(i)}|\mu) = \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}(X^{(i)}-\mu)^T \Sigma^{-1}(X^{(i)}-\mu)\right)$$

$$\log P(X|\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(X^{(i)}-\mu)^T\Sigma^{-1}(X^{(i)}-\mu)$$

$$\frac{\partial}{\partial\mu}\log P(X|\mu) = \frac{1}{2}\sum_{i=1}^{m} \Sigma^{-1}(X^{(i)}-\mu) = \Sigma^{-1}\left(\sum_{i=1}^{m}X^{(i)} - m\mu\right)$$

$$\log P(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0)$$

$$\frac{\partial}{\partial\mu}\log(P(\mu)) = \Sigma^{-1}(\mu_0-\mu)$$

$$\therefore \frac{\partial}{\partial\mu}\log(P(\mu|X)) = \frac{\partial}{\partial\mu}\log\left(\frac{P(X|\mu)P(\mu)}{P(X)}\right) = \frac{\partial}{\partial\mu}\log(P(X|\mu)) + \frac{\partial}{\partial\mu}\log(P(\mu))$$

$$= \Sigma^{-1}\left(\sum_{i=1}^{m}X^{(i)} - m\mu\right) + \Sigma^{-1}(\mu_0-\mu) = 0$$

$$\therefore \sum_{i=1}^{m}X^{(i)} + \mu_0 = (m+1)\mu$$

$$\therefore \mu = \frac{\sum_{i=1}^{m}X^{(i)} + \mu_0}{m+1}$$

$$\therefore \mu_{MAP} = \frac{\sum_{i=1}^{m}X^{(i)} + \mu_0}{m+1}$$

1. $\Sigma_{MAP} = \arg\max_{\Sigma} P(\Sigma|X)$

$$P(\Sigma|X) = \frac{P(X|\Sigma)\,P(\Sigma)}{P(X)}$$

$$\log P(X|\Sigma) = \sum_{i=1}^{m}\left(-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu)\right)$$

$$\frac{d}{d\Sigma}\log P(X|\Sigma) = -\frac{1}{2}m\Sigma^{-1} + \frac{1}{2}\sum_{i=1}^{m}\Sigma^{-1}(x^{(i)}-\mu)(x^{(i)}-\mu)^T\Sigma^{-1}$$

$$\frac{d}{d\Sigma}\log P(\Sigma) = -\frac{1}{2}m\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(\mu-\mu_0)(\mu-\mu_0)^T\Sigma^{-1}$$

$$\frac{d\log P(\Sigma|X)}{d\Sigma} = \frac{d}{d\Sigma}\log\frac{P(X|\Sigma)\,P(\Sigma)}{P(X)} = \frac{d}{d\Sigma}\log P(X|\Sigma) + \frac{d}{d\Sigma}\log P(\Sigma)$$

$$= -m\Sigma^{-1} + \frac{1}{2}\sum_{i=1}^{m}\Sigma^{-1}(x^{(i)}-\mu)(x^{(i)}-\mu)^T\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(\mu-\mu_0)(\mu-\mu_0)^T\Sigma^{-1}$$

$$\therefore \Sigma_{MAP} = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{2}(x^{(i)}-\mu)(x^{(i)}-\mu)^T + \frac{1}{2}(\mu-\mu_0)(\mu-\mu_0)^T\right]$$

2. $\therefore \mu_{MAP} = \dfrac{\sum_{i=1}^{m} x^{(i)} + \mu_0}{m+1}$ $\to \mu_0$

$\therefore \mu_{MAP} - \mu_0 = \dfrac{\sum_{i=1}^{m} x^{(i)} + \mu_0}{m+1} - \mu_0 = \dfrac{\sum_{i=1}^{m} x^{(i)} - m\mu_0}{m+1}$

since $\sum_{i=1}^{m} x^{(i)} = m\mu_0$

$\therefore \mu_{MAP} - \mu_0 = 0$

MAP estimator for this distribution is unbias.

3. $\Sigma_{MAP} = \arg\max_{\Sigma} P(\Sigma|x) = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{1}{2}(x^{(i)} - \mu)(x^{(i)} - \mu)^T + \frac{1}{2}(\mu - \mu_0)(\mu - \mu_0)^T\right]$

$\Sigma_{MLE} = \arg\max_{\Sigma} P(x|\Sigma) = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu)(x^{(i)} - \mu)^T$.

Compare Σ_{MAP} and Σ_{MLE}, we can learn that MAP combiles the possibility information of data sets and the parameter itself, while MLE only contains the possibility of data sets.

With MAP, we can avoid overfitting. However, we must have a prior distribution so that we can perform MAP.