Problem c. Stochastic Gradient Descent

Haoyang Chen
hc2812

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

initialize d_i randomly and $t = 1$

While ($t \leq T$ & Stopping condition is not True) do:

np.random.shuffle(X)

for $k = 0, \cdots,$ X.shape[0] - 1 do:

$D_k = X[k]^T X[k] - \sum_{j=0}^{i-1} \lambda_j d_j d_j^T$

$y = d_i - \gamma \nabla_{d_i}(-d_i^T D_k d_i)$

$d_i = \frac{y}{\|y\|}$

$t = t + 1$

$\lambda_i = d_i^T X^T X d_i$

Neural Networks and Deep Learning

Haoyang Chen hc2812 HW1.

Theory Part:

Problem 4:

(i) Y is uniformly distributed when $x \in [0, 255]$

$$\therefore y = cdf(x) = \int_0^x p(u)du = \int_0^x \frac{f(u)}{\int_0^{255} f(t)dt} du \simeq \int_0^x f(u)du = \int_0^x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(x-127.5)$$

Thus the mapping function $g(x)$ is the cumulative distribution function $\Phi(x-127.5)$

(ii) $p(X=x) = \int_0^\infty \int_0^\infty p(X=x, Y=y, Z=z)\,dydz = \int_0^1 \int_0^1 8xyz\,dydz = 2x \cdot [y^2]_0^1 \cdot [z^2]_0^1 = 2x, \quad x \in [0,1]$

$p(Y=y) = \int_0^\infty \int_0^\infty p(X=x, Y=y, Z=z)\,dxdz = \int_0^1 \int_0^1 8xyz\,dydz = 2y \cdot [x^2]_0^1 \cdot [z^2]_0^1 = 2y, \quad y \in [0,1]$

$p(Z=z) = \int_0^\infty \int_0^\infty p(X=x, Y=y, Z=z)\,dxdy = \int_0^1 \int_0^1 8xyz\,dxdy = 2z \cdot [x^2]_0^1 \cdot [y^2]_0^1 = 2z, \quad z \in [0,1]$

$$\therefore p(x) = \begin{cases} 2x, & x \in [0,1] \\ 0, & \text{otherwise} \end{cases} \quad p(y) = \begin{cases} 2y, & y \in [0,1] \\ 0, & \text{otherwise} \end{cases}, \quad p(z) = \begin{cases} 2z, & z \in [0,1] \\ 0, & \text{otherwise} \end{cases}$$

$$E(xyz) = \int_0^\infty \int_0^\infty \int_0^\infty xyz\, p(x,y,z)\,dxdydz = \int_0^1 \int_0^1 \int_0^1 xyz \cdot 8xyz\,dxdydz = 8\left[\frac{x^3}{3}\right]_0^1 \cdot \left[\frac{y^3}{3}\right]_0^1 \cdot \left[\frac{z^3}{3}\right]_0^1 = \frac{8}{27}$$

x and y are conditionally independent given z.

Proof: $P(x,y|z) = \frac{P(x,y,z)}{P(z)} = \frac{8xyz}{2z} = 4xy, \quad \text{for } x,y,z \in [0,1]$

$P(x|z) = \frac{P(x,z)}{P(z)} = \frac{\int_0^\infty p(x,y,z)dy}{p(z)} = \frac{\int_0^1 8xyz\,dy}{2z} = \frac{4xz}{2z} = 2x, \quad \text{for } x,z \in [0,1]$

$p(y|z) = \frac{p(y,z)}{p(z)} = \frac{\int_0^\infty p(x,y,z)dx}{p(z)} = \frac{\int_0^1 8xyz\,dx}{2z} = \frac{4yz}{2z} = 2y, \quad \text{for } y,z \in [0,1]$

$$\therefore p(x,y|z) = \begin{cases} 4xy, & \text{for } x,y,z \in [0,1] \\ 0, & \text{otherwise} \end{cases} \quad p(x|z) = \begin{cases} 2x, & \text{for } x,z \in [0,1] \\ 0 & \text{otherwise} \end{cases}, \quad p(y|z) = \begin{cases} 2y, & \text{for } y,z \in [0,1] \\ 0, & \text{otherwise} \end{cases}$$

$\therefore p(x,y|z) = p(x|z)\,p(y|z) \quad \therefore$ x and y are conditionally independent given z #

Problem e:

1. $P(\mu|X,\Sigma) \propto P(X|\mu,\Sigma)P(\mu|\mu_0,\Sigma_0) \propto \prod_{i=1}^{m} \exp(-\frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu)) \cdot \exp(-\frac{1}{2}(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0))$

$= \exp(-\frac{1}{2}\left((\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0) + \sum_{i=1}^{m}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu)\right)$

$\propto \exp(-\frac{1}{2}(\mu-\mu_n)^T\Sigma_n^{-1}(\mu-\mu_n)) = N(\mu|\mu_m,\Sigma_m)$

where $\mu_m = (\Sigma_0^{-1}+m\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + m\Sigma^{-1}\bar{x})$, $\bar{x} = \sum_{i=1}^{m} x^{(i)}/m$

$\Sigma_m^{-1} = \Sigma_0^{-1} + m\Sigma^{-1}$

Thus $\mu_{MAP} = \mu_m$ which maximize $P(\mu|X,\Sigma)$

I guess professor means $\Sigma_{MAP} = \Sigma_m$, but in bayesian statistics, it's wired. Σ_m is related to the posterior distribution of μ, not the posterior distribution of Σ. In bayesian statistics, if Σ is unknown, $\Sigma_m^{-1} = \Sigma_0^{-1} + m\Sigma_{MAP}^{-1}$, and $\Sigma_{MAP}^{-1} = \arg\max_{\Sigma} P(\Sigma|X)$.

2. $E(\mu_{MAP}) = E((\Sigma_0^{-1}+m\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + m\Sigma^{-1}\bar{x})) \neq E(\bar{x}) = \mu$

$E(\Sigma_{MAP}) = E((\Sigma_0^{-1}+m\Sigma^{-1})^{-1}) \neq \Sigma$

Thus it's biased estimators

3. $L(X,\mu,\Sigma) = \prod_{i=1}^{m} \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu))$

$\propto |\Sigma|^{-\frac{m}{2}} \exp(-\frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu))$

$l(X,\mu,\Sigma) = \log L(X,\mu,\Sigma) \propto -\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu))$

$\frac{\partial l(X,\mu,\Sigma)}{\partial \mu} = 0 \Rightarrow \mu_{MLE} = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$

$\frac{\partial l(X,\mu,\Sigma)}{\partial \Sigma} = 0 \Rightarrow \Sigma_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{MLE})(x^{(i)}-\mu_{MLE})^T = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\bar{x})(x^{(i)}-\bar{x})^T$

Comparing the MLE & MAP, Σ_{MLE} is an unbiased estimator while Σ_{MAP} is biased