DeepLearning HW2

Sirui Tan, st2957

March 7, 2016

1 Problem a

1.1 (i)

Let $Y = [y_1, \cdots, y_m]$ and $X = [x_1, \cdots, x_m]$, we have

$$\mathcal{L}_{ls} = Tr[(Y - AX)^T(Y - AX)]$$

Take the derivative

$$\begin{aligned}\Delta_A \mathcal{L}_{ls} &= \left[\Delta_A (Y - AX)^T(Y - AX)\right]^T \\ &= \Delta_A (Y^T Y - X^T A^T Y - Y^T AX + X^T A^T AX) \\ &= 0\end{aligned}$$

Simplify the equation yields

$$2AXX^T = YX^T$$

Therefore

$$A_{ls} = YX^T(XX^T)^{-1}$$

1.2 (ii)

According to (i)

$$\mathcal{L}_r = \lambda Tr(A^T A) + Tr\left[(Y - AX)^T(Y - AX)\right]$$

Take the derivative

$$\Delta_A \mathcal{L}_r = \lambda A^T + 2AXX^T - 2YX^T = 0$$

Which yields

$$A_r = YX^T(XX^T + \lambda I)^{-1}$$

1.3 (iii)

According to the problem

$$y \sim \mathcal{N}(Ax, \sigma^2 I)$$

Therefore the likelihood can be expressed as

$$\prod_{i=1}^{m} p(y_i|x_i, A) = \frac{1}{(2\pi)^(n/2)\sigma} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - Ax_i)^T(y_i - Ax_i)\right]$$

In order to maximize likelihood, $\sum_{i=1}^{m}(y_i - Ax_i)^T(y_i - Ax_i)$ is to be minimized. According to (i).

$$A_{ml} = YX^T(XX^T)^{-1}$$

1.4 (iv)

$$\begin{aligned} p(A|y_1, \cdots, y_m, x_1, \cdots, x_m) &\propto p(y_1, \cdots, y_m|A, x_1, \cdots, x_m)p(A) \\ &\propto \left[\prod_{i=1}^{m} p(y_i|A, x_i)\right] p(A) \\ &\propto \exp\left[\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - Ax_i)^T(y_i - Ax_i)\right] \exp\left[-\frac{1}{2}Tr\left[\lambda, (A-M)^T(A-M)\right]\right] \end{aligned}$$

In order to maximize a posterori,

$$\begin{aligned} \Delta_A \left[\sum_{i=1}^{m}(y_i - Ax_i)^T(y_i - Ax_i) + Tr\left[\lambda(A-M)^T(A-M)\right]\right] &= 0 \\ (2XX^T + \lambda I)A - 2YX^T - 2\lambda M &= 0 \end{aligned}$$

Therefore

$$A_{MAP} = (2YX^T + 2\lambda M)(XX^T + \lambda I)^{-1}$$

If $M = 0$,

$$A_{MAP} = YX^T(XX^T + \lambda I)^{-1}$$

1.5 (iv)

For (i) and (iii), the relationship is two different ways of looking at the given problem, (i) looks at the problem from a deterministic point of view and find a good estimate by minimize the square error. (ii), on the other hand, looks at the problem from a stochastic point of view, assuming Y is the deterministic version plus white noise, and find a good fit by maximize the observation likelihood.

For (ii) and (iv), the basic idea is identical. The only difference is that (ii) uses a regularization term to stabilize estimate result and (iv) uses Bayes rule to maximize the posterior probability instead of the likelihood itself.