Neural Networks and Deep Learning Problem Set #2

Si Kai Lee sl3950@columbia.edu

March 7, 2016

Problem A

i

$$
\begin{aligned}
L_{ls} &= \sum_{i=1}^{m} (y^i - Ax^i)^T (y^i - Ax^i) \\
&= \sum_{i=1}^{m} ||y^i - Ax^i||_2^2 \\
&= ||Y - XA||_2^2 \\
\nabla_A L_{ls} &= \nabla_A ||Y - AX||_2^2 \\
&= \nabla_A (Y - AX)^T (Y - AX) \\
&= \nabla_A (Y^T Y - Y^T AX - (AX)^T Y + (AX)^T AX) \\
&\text{Since } Y^T AX \text{ and } (AX)^T Y \text{ are scalars, } Y^T AX = (Y^T AX)^T = (AX)^T Y, \\
&= \nabla_A (Y^T Y - 2(AX)^T Y + X^T A^T AX) \\
&= 2AXX^T - 2XY^{T1} \\
&\text{Equating the above to 0,} \\
&2AXX^T - 2XY^T = 0 \\
&AXX^T = XY^T \\
&A_{ls} = XY^T (XX^T)^{-1}
\end{aligned}
$$

[1] Matrix Cookbook 77

ii

$$
\begin{aligned}
L_r &= \lambda||A||_F^2 + \sum_{i=1}^{m}(y^i - Ax^i)^T(y^i - Ax^i) \\
&= \lambda||A||_F^2 + ||Y - AX||_2^2 \\
\nabla_A L_r &= \nabla_A \lambda||A||_F^2 + ||Y - AX||_2^2 \\
&= 2\lambda A + 2AXX^T - 2XY^T \\
& \text{Equating the above to 0,} \\
& 2\lambda A + 2AXX^T - 2XY^T = 0 \\
& A(XX^T + \lambda I) = XY^T \\
& A_r = XY^T(XX^T + \lambda I)^{-1}
\end{aligned}
$$

iii

Assuming $e^i \sim \mathcal{N}(0, \sigma^2 I)$ and $e^i = y - Ax$, hence $y \sim \mathcal{N}(AX, \sigma^2 I)$

$$
\begin{aligned}
L_n &= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y^i - Ax^i)^T(y^i - Ax^i)) \\
l_n &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(Y - AX)^T(Y - AX) \\
\nabla_A l_n &= -\frac{1}{2\sigma^2}2AXX^T - 2XY^T \\
& \text{Equating the above to 0,} \\
& 2AXX^T - 2XY^T = 0 \\
& A_{MLE} = XY^T(XX^T)^{-1}
\end{aligned}
$$

iv

Assuming $e^i \sim \mathcal{N}(0, \sigma^2 I)$, hence $y \sim \mathcal{N}(XA, \sigma^2 I)$

$$
\begin{aligned}
\Pr(A|X,Y) &= \frac{\Pr(X,Y,A)}{\Pr(X,Y)} = \frac{\Pr(X,Y|A)\Pr(A)}{\Pr(X,Y)} \\
&\propto \Pr(X,Y|A)\Pr(A) \\
&= \exp(-\frac{1}{2\sigma^2}(Y - AX)^T(Y - AX)) * \exp(\frac{1}{2}\operatorname{Tr}[\lambda(A - M)^T(A - M)]) \\
&= \exp(-\frac{1}{2\sigma^2}(Y^TY - 2(AX)^TY + X^TA^TAX) * \exp(\frac{1}{2}\operatorname{Tr}[(A - M)\lambda(A - M)^T]) \\
&= \exp(-\frac{1}{2\sigma^2}(Y^TY - 2(AX)^TY + X^TA^TAX + \frac{1}{2}\operatorname{Tr}(A\lambda A^T - A\lambda M^T - M\lambda A^T - M\lambda M^T) \\
& \text{Removing all terms non-related to } A \text{ as differentiating by } A \text{ later,} \\
&= \exp(-\frac{1}{2\sigma^2}(-2(AX)^TY + X^TA^TAX) + \frac{1}{2}\operatorname{Tr}(A\lambda A^T - 2A\lambda M^T)^2
\end{aligned}
$$

[2] Matrix Cookbook 14

Since $A_{MAP} = \arg\max_A \ \ln \Pr(Y, X|A) + \ln \Pr(A)$

$$\begin{aligned}\nabla_A \ln \Pr(Y, X|A) + \ln \Pr(A) &= -\frac{1}{2\sigma^2}(-2XY^T + 2AXX^T) + \frac{\lambda}{2}(2A - 2M)^3 \\ \text{Equating to 0,}& \\ A(XX^T + \sigma^2 \lambda I) &= XY^T + \sigma^2 \lambda M \\ A_{MAP} &= (XY^T + \sigma^2 \lambda M)(XX^T + \sigma^2 \lambda I)^{-1}\end{aligned}$$

If M is the zero matrix, A_{MAP} would be A_r with a σ^2 shift in the λI regulariser.

v

If the λ in (ii) and variance of the prior in (iv) were 0, then $A_r = A_{MAP} = A_{ls} = A_{MLE}$. (i) should be equal to (iii) as the maximum likelihood estimate of A the same A with minimum squared error as the estimated A would be the one that best fits the data.

[3] Matrix Cookbook 104, 115