Machine Learning 1 EX 06

01.12.2015

1. **KL divergence of two Gaussians:**

a) According to a Linear Algebra Calculation Rule: $\mathbf{tr}(\mathbf{AB}) = \mathbf{tr}(\mathbf{BA})$

Notice that $(\mathbf{x} - \mu_\mathbf{i})^\mathbf{T}\mathbf{\Sigma^{-1}}(\mathbf{x} - \mu_\mathbf{i})$ is a scalar, apply the rule, we get:

$$\begin{aligned}(\mathbf{x} - \mu_\mathbf{i})^\mathbf{T}\mathbf{\Sigma^{-1}}(\mathbf{x} - \mu_\mathbf{i}) &= \mathbf{tr}\left((\mathbf{x} - \mu_\mathbf{i})^\mathbf{T}\mathbf{\Sigma^{-1}}(\mathbf{x} - \mu_\mathbf{i})\right) \\ &= \mathbf{tr}\left((\mathbf{x} - \mu_\mathbf{i})^\mathbf{T}\left(\mathbf{\Sigma^{-1}}(\mathbf{x} - \mu_\mathbf{i})\right)\right) \\ &= \mathbf{tr}\left(\left(\mathbf{\Sigma^{-1}}(\mathbf{x} - \mu_\mathbf{i})\right)(\mathbf{x} - \mu_\mathbf{i})^\mathbf{T}\right) \\ &= \mathbf{tr}\left(\mathbf{\Sigma^{-1}}(\mathbf{x} - \mu_\mathbf{i})(\mathbf{x} - \mu_\mathbf{i})^\mathbf{T}\right)\end{aligned}$$

b) Notice that:

$$\mathbb{E}_{\mathbf{X_1}}[\mathbf{x}] = \mu_\mathbf{1}, \qquad \mathbb{E}_{\mathbf{X_1}}[\mathbf{xx^T}] = \mathbf{\Sigma_1} + \mu_\mathbf{1}\mu_\mathbf{1}^\mathbf{T}$$

As Expectations are additive, the trace operator can exchange with Expectation operator, we get:

$$\begin{aligned}\mathbb{E}_{\mathbf{X_1}}\left[\mathbf{tr}\left(\mathbf{\Sigma_2^{-1}}(\mathbf{xx^T} - \mathbf{2x}\mu_\mathbf{2}^\mathbf{T} + \mu_\mathbf{2}\mu_\mathbf{2}^\mathbf{T})\right)\right] &= \mathbb{E}_{\mathbf{X_1}}\left[\mathbf{tr}\left(\mathbf{\Sigma_2^{-1}xx^T}\right)\right] - 2\mathbb{E}_{\mathbf{X_1}}\left[\mathbf{tr}\left(\mathbf{\Sigma_2^{-1}x}\mu_\mathbf{2}^\mathbf{T}\right)\right] + \mathbb{E}_{\mathbf{X_1}}\left[\mathbf{tr}\left(\mathbf{\Sigma_2^{-1}}\mu_\mathbf{2}\mu_\mathbf{2}^\mathbf{T}\right)\right] \\ &= \mathbf{tr}\left(\mathbb{E}_{\mathbf{X_1}}\left[\mathbf{\Sigma_2^{-1}xx^T}\right]\right) - 2\mathbf{tr}\left(\mathbb{E}_{\mathbf{X_1}}\left[\mathbf{\Sigma_2^{-1}x}\mu_\mathbf{2}^\mathbf{T}\right]\right) + \mathbf{tr}\left(\mathbf{\Sigma_2^{-1}}\mu_\mathbf{2}\mu_\mathbf{2}^\mathbf{T}\right) \\ &= \mathbf{tr}\left(\mathbf{\Sigma_2^{-1}}\mathbb{E}_{\mathbf{X_1}}\left[\mathbf{xx^T}\right]\right) - 2\mathbf{tr}\left(\mathbf{\Sigma_2^{-1}}\mathbb{E}_{\mathbf{X_1}}\left[x\right]\mu_\mathbf{2}^\mathbf{T}\right) + \mathbf{tr}\left(\mathbf{\Sigma_2^{-1}}\mu_\mathbf{2}\mu_\mathbf{2}^\mathbf{T}\right) \\ &= \mathbf{tr}\left(\mathbf{\Sigma_2^{-1}\Sigma_1} + \mathbf{\Sigma_2^{-1}}\mu_\mathbf{1}\mu_\mathbf{1}^\mathbf{T}\right) - 2\mathbf{tr}\left(\mathbf{\Sigma_2^{-1}}\mu_\mathbf{1}\mu_\mathbf{2}^\mathbf{T}\right) + \mathbf{tr}\left(\mathbf{\Sigma_2^{-1}}\mu_\mathbf{2}\mu_\mathbf{2}^\mathbf{T}\right) \\ &= \mathbf{tr}\left(\mathbf{\Sigma_2^{-1}}\left(\mathbf{\Sigma_1} + \mu_\mathbf{1}\mu_\mathbf{1}^\mathbf{T} - \mathbf{2}\mu_\mathbf{1}\mu_\mathbf{2}^\mathbf{T} + \mu_\mathbf{2}\mu_\mathbf{2}^\mathbf{T}\right)\right)\end{aligned}$$

c)

$$\begin{aligned}\mathbf{tr}\left(\mu_\mathbf{1}^\mathbf{T}\mathbf{\Sigma_2^{-1}}\mu_\mathbf{1} - 2\mu_\mathbf{1}\mathbf{\Sigma_2^{-1}}\mu_\mathbf{2} + \mu_\mathbf{2}^\mathbf{T}\mathbf{\Sigma_2^{-1}}\mu_\mathbf{2}\right) &= \mathbf{tr}\left(\mu_\mathbf{1}^\mathbf{T}\mathbf{\Sigma_2^{-1}}\left(\mu_\mathbf{1} - \mu_\mathbf{2}\right) - \left(\mu_\mathbf{1} - \mu_\mathbf{2}\right)^T\mathbf{\Sigma_2^{-1}}\mu_\mathbf{2}\right) \\ &= \mathbf{tr}\left(\mu_\mathbf{1}^\mathbf{T}\mathbf{\Sigma_2^{-1}}\left(\mu_\mathbf{1} - \mu_\mathbf{2}\right)\right) - \mathbf{tr}\left(\left(\mu_\mathbf{1} - \mu_\mathbf{2}\right)^T\mathbf{\Sigma_2^{-1}}\mu_\mathbf{2}\right)\end{aligned}$$

Notice that for any square matrix $\mathbf{A}$, $\mathbf{tr}(\mathbf{A}) = \mathbf{tr}(\mathbf{A^T})$

Also otice that $\mathbf{\Sigma_2^{-1}}$ is symmetric, we can conclude that: $\mathbf{\Sigma_2^{-1}} = (\mathbf{\Sigma_2^{-1}})^T$

$$
\begin{aligned}
\mathbf{tr}\left(\mu_1^{\mathbf{T}}\boldsymbol{\Sigma}_2^{-1}(\mu_1-\mu_2)\right)-\mathbf{tr}\left((\mu_1-\mu_2)^T\boldsymbol{\Sigma}_2^{-1}\mu_2\right) &= \mathbf{tr}\left(\mu_1^{\mathbf{T}}\boldsymbol{\Sigma}_2^{-1}(\mu_1-\mu_2)\right)-\mathbf{tr}\left(\mu_2^{\mathbf{T}}\boldsymbol{\Sigma}_2^{-1}(\mu_1-\mu_2)\right)\\
&= \mathbf{tr}\left(\mu_1^{\mathbf{T}}\boldsymbol{\Sigma}_2^{-1}(\mu_1-\mu_2)-\mu_2^{\mathbf{T}}\boldsymbol{\Sigma}_2^{-1}(\mu_1-\mu_2)\right)\\
&= \mathbf{tr}\left((\mu_1-\mu_2)^{\mathbf{T}}\boldsymbol{\Sigma}_2^{-1}(\mu_1-\mu_2)\right) \qquad \textit{This is a scalar}\\
&= (\mu_1-\mu_2)^{\mathbf{T}}\boldsymbol{\Sigma}_2^{-1}(\mu_1-\mu_2)
\end{aligned}
$$

2. **Cost function of SSA:**

a)

$$
\begin{aligned}
\sum_{\mathbf{i=1}}^{\mathbf{N}}\mathbf{D_{KL}}[\mathcal{N}(\mu_{\mathbf{i}}^{\mathbf{s}},\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}})\parallel\mathcal{N}(\mathbf{0},I)] &= \sum_{\mathbf{i=1}}^{\mathbf{N}}\frac{\mathbf{1}}{\mathbf{2}}\left(\log\frac{\mathbf{det}(I)}{\mathbf{det}\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}}}+\mathbf{tr}(I^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}})+(\mathbf{0}-\mu_{\mathbf{i}}^{\mathbf{s}})^{\mathbf{T}}I^{-\mathbf{1}}(\mathbf{0}-\mu_{\mathbf{i}}^{\mathbf{s}})-\mathbf{N}\right)\\
&= \frac{\mathbf{1}}{\mathbf{2}}\sum_{\mathbf{i=1}}^{\mathbf{N}}\left(-\log\mathbf{det}\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}}+\mathbf{tr}(\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}})+\mu_{\mathbf{i}}^{\mathbf{s}^{\mathbf{T}}}\mu_{\mathbf{i}}^{\mathbf{s}}-\mathbf{N}\right)\\
&= \frac{\mathbf{1}}{\mathbf{2}}\left(\sum_{\mathbf{i=1}}^{\mathbf{N}}(-\log\mathbf{det}\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}}+\mu_{\mathbf{i}}^{\mathbf{s}^{\mathbf{T}}}\mu_{\mathbf{i}}^{\mathbf{s}})+\sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathbf{tr}(\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}})-\mathbf{N})\right)\\
\sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathbf{tr}(\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}})) &= \mathbf{tr}(\sum_{\mathbf{i=1}}^{\mathbf{N}}\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}})\\
&= \mathbf{tr}(\sum_{\mathbf{i=1}}^{\mathbf{N}}\mathbf{B}\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}}\mathbf{B}^{\mathbf{T}})\\
&= \mathbf{tr}(\mathbf{B}\sum_{\mathbf{i=1}}^{\mathbf{N}}\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}}\mathbf{B}^{\mathbf{T}})\\
&= \mathbf{tr}(\mathbf{BN}I\mathbf{B}^{\mathbf{T}})\\
&= \mathbf{Ntr}(\mathbf{BB}^{\mathbf{T}})\\
&= \mathbf{Ntr}(I)\\
&= \mathbf{N}^{\mathbf{2}}
\end{aligned}
$$

Therefore

$$
\sum_{\mathbf{i=1}}^{\mathbf{N}}\mathbf{D_{KL}}[\mathcal{N}(\mu_{\mathbf{i}}^{\mathbf{s}},\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}})\parallel\mathcal{N}(\mathbf{0},I)] = \frac{\mathbf{1}}{\mathbf{2}}\sum_{\mathbf{i=1}}^{\mathbf{N}}(-\log\mathbf{det}\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{s}}+\mu_{\mathbf{i}}^{\mathbf{s}^{\mathbf{T}}}\mu_{\mathbf{i}}^{\mathbf{s}})
$$