Pattern Recognition and Machine Learning Exercises

Jesper Stemann Andersen

February 16, 2007

Solutions to exercises found in Pattern Recognition and Machine Learning by Christopher M. Bishop.

Exercise 2.35

i.) Given (2.59), $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$, we wish to prove (2.62):

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

Let's consider the covariance matrix, $\boldsymbol{\Sigma} = \operatorname{cov}[\boldsymbol{x}]$, as given by equation (2.63):

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{x}\mathbb{E}[\boldsymbol{x}]^{\mathrm{T}} - \mathbb{E}[\boldsymbol{x}]\boldsymbol{x}^{\mathrm{T}} + \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \boldsymbol{x}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{x}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \\
&= \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}
\end{aligned}$$

Which was what we were hoping to prove. From line 2 to the 3, we use (2.59), from line 3 to 4 we use in addition that the expectation function is linear (and thus can be distributed across a sum) and that $\mathbb{E}[\boldsymbol{\mu}] = \boldsymbol{\mu}$.

ii.a.) Now, using the just proven (2.62), we wish to show:

$$\mathbb{E}[\boldsymbol{x_n}\boldsymbol{x_m}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + I_{nm}\boldsymbol{\Sigma}$$

where I_{nm} is the (n, m) entry of the identity matrix and $\boldsymbol{x_n}$ denotes a multidimensional data point sampled from a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We assume data points to be i.i.d.

Since the data points are assumed to be independent $\operatorname{cov}[\boldsymbol{x_n}, \boldsymbol{x_m}] = 0$ for $n \neq m$. What we wish to show is then equivalent to the expression for the covariance between two random variables, as given by (1.42):

$$\begin{aligned}
\operatorname{cov}[\boldsymbol{x_n}, \boldsymbol{x_m}] &= \mathbb{E}[\boldsymbol{x_n}\boldsymbol{x_m}^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{x_n}]\mathbb{E}[\boldsymbol{x_m}^{\mathrm{T}}] \\
\mathbb{E}[\boldsymbol{x_n}\boldsymbol{x_m}^{\mathrm{T}}] &= \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}
\end{aligned}$$

ii.b.) Given $\boldsymbol{\mu_{ML}}$, we wish to prove $\mathbb{E}[\boldsymbol{\Sigma_{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}$, using (2.122):

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\Sigma_{ML}}] &= \mathbb{E}\left[\frac{1}{N}\Sigma_{i=1}^{N}(\boldsymbol{x_i} - \boldsymbol{\mu_{ML}})(\boldsymbol{x_i} - \boldsymbol{\mu_{ML}})^{\mathrm{T}}\right] \\
&= \frac{1}{N}\mathbb{E}\left[\Sigma_{i=1}^{N}(\boldsymbol{x_i}\boldsymbol{x_i}^{\mathrm{T}} - \boldsymbol{x_i}\boldsymbol{\mu_{ML}}^{\mathrm{T}} - \boldsymbol{\mu_{ML}}\boldsymbol{x_i}^{\mathrm{T}} + \boldsymbol{\mu_{ML}}\boldsymbol{\mu_{ML}}^{\mathrm{T}})\right] \\
&= \frac{1}{N}\Sigma_{i=1}^{N}\left(\mathbb{E}[\boldsymbol{x_i}\boldsymbol{x_i}^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{x_i}\boldsymbol{\mu_{ML}}^{\mathrm{T}}] - \mathbb{E}[\boldsymbol{\mu_{ML}}\boldsymbol{x_i}^{\mathrm{T}}] + \mathbb{E}[\boldsymbol{\mu_{ML}}\boldsymbol{\mu_{ML}}^{\mathrm{T}}]\right) \\
&= \frac{1}{N}\Sigma_{i=1}^{N}\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} - \frac{1}{N}\mathbb{E}[\boldsymbol{x_i}\Sigma_{j=1}^{N}\boldsymbol{x_j}^{\mathrm{T}}] - \frac{1}{N}\mathbb{E}[\Sigma_{j=1}^{N}\boldsymbol{x_j}\boldsymbol{x_i}^{\mathrm{T}}] + \frac{1}{N^2}\mathbb{E}[\Sigma_{j=1}^{N}\boldsymbol{x_j}\Sigma_{k=1}^{N}\boldsymbol{x_k}^{\mathrm{T}}]\right) \\
&= \frac{1}{N}\Sigma_{i=1}^{N}\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} - \frac{1}{N}\Sigma_{j=1}^{N}\mathbb{E}[\boldsymbol{x_i}\boldsymbol{x_j}^{\mathrm{T}}] - \frac{1}{N}\Sigma_{j=1}^{N}\mathbb{E}[\boldsymbol{x_j}\boldsymbol{x_i}^{\mathrm{T}}] + \frac{1}{N^2}\Sigma_{j=1}^{N}\Sigma_{k=1}^{N}\mathbb{E}[\boldsymbol{x_j}\boldsymbol{x_k}^{\mathrm{T}}]\right)
\end{aligned}$$

We can now apply the sum, $\Sigma_{j=1}^{N}\mathbb{E}[\boldsymbol{x_i}\boldsymbol{x_j}^{\mathrm{T}}] = N\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\Sigma_{ML}}] &= \frac{1}{N}\Sigma_{i=1}^{N}\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} - \frac{1}{N}(N\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}) - \frac{1}{N}(N\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}) + \frac{1}{N^2}\Sigma_{j=1}^{N}(N\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma})\right) \\
&= \frac{1}{N}\Sigma_{i=1}^{N}\left(\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \frac{1}{N}\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \frac{1}{N}\boldsymbol{\Sigma} + \frac{1}{N}(N\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma})\right) \\
&= \frac{1}{N}\Sigma_{i=1}^{N}\left(\boldsymbol{\Sigma} - \frac{1}{N}\boldsymbol{\Sigma}\right) \\
&= \boldsymbol{\Sigma} - \frac{1}{N}\boldsymbol{\Sigma} = \frac{N-1}{N}\boldsymbol{\Sigma}
\end{aligned}$$

Which was what we set out to prove.