Assignment 1 - Report

Pietro Alovisi

11-17-2018

Question 1

The gaussian function is a unimodal distribution, which means that has only one mode and for this particular distribution it coincides with the mean. So in this case we are assuming that value of the deterministic function f for a given $\mathbf{x}$ is the mean value of the distribution of the target. This can be rephrased as assuming a determinsitic model $f(\mathbf{x})$ that generates realizations with a random error ε that distributes as $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Putting everyting togheter we get:

$$\mathbf{t} = f(\mathbf{x}) + \varepsilon$$

A prior oservation about the covariance is that we are assuming homoscedasticity, that is the variance of $\mathbf{t}$ is not dependent on the input vector $\mathbf{x}$.

The spherical covariance matrix means implies two facts:

- All the scalar random variables t_j of the vector $\mathbf{t_i}$ have the same variance σ^2.
- The fact that the covariance matrix is diagonal means that all the output sclar component t_j of the vector $\mathbf{t_i}$ are independent one another.

Question 2

If we do not assume independence of the samples, we must turn to the joint probability distribution

$$p(\mathbf{T}|f, \mathbf{X}) = p(\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_N}|f, \mathbf{X})$$

Question 3

Equation 5 is a linear transformation of a normal distribution which, from its properties, is again a normal distribution equal to:

$$p(\mathbf{t_i}) \sim \mathcal{N}(\mathbf{W}\,\mathbf{x_i}, \sigma^2\mathbf{I})$$

Still assuming conditionally independent samples, from 3 the likelyhood is just:

$$p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{t_i}|\mathbf{W}\,\mathbf{x_i}, \sigma^2\mathbf{I})$$

Which we can also write by vectorising the whole, by noting that since all the $\mathbf{t_i}$ have the same variance, the exponents in the probability density function sum up.

$$\begin{aligned} p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \mathcal{N}(\mathbf{X}\mathbf{W}^T, \mathbf{I}, \sigma^2\mathbf{I}) = \\ = \frac{1}{\sigma^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{1}{2\sigma^2}\sum_i^N (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i})} = \\ = \frac{1}{\sigma^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{1}{2\sigma^2}Tr\left((\mathbf{X}\mathbf{W^T} - \mathbf{T})(\mathbf{X}\mathbf{W^T} - \mathbf{T})^T\right)} \end{aligned}$$

Where we substituted the expression at the exponent $\sum_i^N (\mathbf{W}\mathbf{x_i} - \mathbf{t_i})^T(\mathbf{W}\mathbf{x_i} - \mathbf{t_i})$ with $Tr\left((\mathbf{X}\mathbf{W^T} - \mathbf{T})(\mathbf{X}\mathbf{W^T} - \mathbf{T})^T\right)$ by noting that the summation is just the sum of the diagonal of the matrix $(\mathbf{X}\mathbf{W^T} - \mathbf{T})(\mathbf{X}\mathbf{W^T} - \mathbf{T})^T$.

Question 4

Using L_1 norm will perform some kind of dimensionality reduction by setting some variables to 0.

The two penalization terms are:

$$p(W) = \frac{1}{\sigma^2(2\pi)^{\frac{D}{2}}} \cdot e^{-\frac{tr((W-W_0)(W-W_0)^T)}{2\tau^2}}$$

$w^T w$: for the L_2 norm which is just the Froebenius norm $|w|$

Question 5

We will use the square completion to perform this task. So we assume that the output is normal with the following parametrs:

$$p(W) = \frac{1}{\xi} \cdot e^{-tr((W-W_0)\Sigma^{-1}(W-W_0)^T)} =$$

$$= \frac{1}{\xi} \cdot e^{-tr(W\Sigma^{-1}W^T)} e^{2\cdot tr(W\Sigma^{-1}W_0^T)} e^{-tr(W_0\Sigma^{-1}W_0^T)}$$

Where ξ is just the normlaizing factor to make the integral of the function 1.

Now we will take the product of the prior over W, and the likelyhood $p(\mathbf{t_i}|\mathbf{x_i}, \mathbf{W})$.

$$p(\mathbf{t_i}) = e^{-\frac{1}{2\sigma^2}(\mathbf{t_i}-W\mathbf{x_i})^T(\mathbf{t_i}-W\mathbf{x_i})} \cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)}$$
$$= e^{-\frac{1}{2\sigma^2}tr((\mathbf{t_i}-W\mathbf{x_i})(\mathbf{t_i}-W\mathbf{x_i})^T)} \cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)}$$

Since they have the same dimensions, we can do merge the two traces:

$$p(\mathbf{W}|\mathbf{t_i}, \mathbf{x_i}) = e^{-\frac{1}{2\sigma^2}(\mathbf{t_i}-W\mathbf{x_i})^T(\mathbf{t_i}-W\mathbf{x_i})} \cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)} =$$
$$= e^{-\frac{1}{2\sigma^2}tr((\mathbf{t_i}-W\mathbf{x_i})(\mathbf{t_i}-W\mathbf{x_i})^T)} \cdot e^{-\frac{1}{2\tau^2}tr((W-W_0)(W-W_0)^T)} =$$
$$= e^{-\frac{1}{2\sigma^2}tr(\mathbf{t_i}\mathbf{t_i}^T)} e^{\frac{1}{\sigma^2}tr(W\mathbf{x_i}\mathbf{t_i}^T)} e^{-\frac{1}{2\sigma^2}tr(W\mathbf{x_i}\mathbf{x_i}^TW^T)} e^{-\frac{1}{2\tau^2}tr(WW^T)} e^{\frac{1}{\tau^2}tr(WW_0^T)} e^{-\frac{1}{2\tau^2}tr(W_0W_0^T)} =$$
$$= e^{-tr(W(\frac{1}{2\tau^2}\mathbf{I}+\frac{1}{2\sigma^2}\mathbf{x_i}\mathbf{x_i}^T)W^T)} e^{tr(W(\frac{1}{\sigma^2}\mathbf{x_i}\mathbf{t_i}^T+\frac{1}{\tau^2}W_0^T))} e^{-tr(\frac{1}{2\sigma^2}\mathbf{t_i}\mathbf{t_i}^T+\frac{1}{2\tau^2}W_0W_0^T)}$$

Then assuming the independence of the t_i we can get to the full posterior.

$$p(\mathbf{W}|\mathbf{T}, \mathbf{X}) = e^{-tr(W(\frac{1}{2\tau^2}\mathbf{I}+\frac{1}{2\sigma^2}\sum_i \mathbf{x_i}\mathbf{x_i}^T)W^T)} e^{tr(W(\frac{1}{\sigma^2}\mathbf{x_i}\mathbf{t_i}^T+\frac{1}{\tau^2}W_0^T))} e^{-tr(\frac{1}{2\sigma^2}\mathbf{t_i}\mathbf{t_i}^T+\frac{1}{2\tau^2}W_0W_0^T)}$$

$$p(\mathbf{W}|\mathbf{T}, \mathbf{X}) = e^{-tr(W(\frac{1}{2\tau^2}\mathbf{I}+\frac{1}{2\sigma^2}\mathbf{X}^T\mathbf{X})W^T)} e^{tr(W(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{T}+\frac{1}{\tau^2}W_0^T))} e^{-tr(\frac{1}{2\sigma^2}\mathbf{T}^T\mathbf{T}+\frac{1}{2\tau^2}W_0W_0^T)}$$

Where we substituted $\sum_i \mathbf{x_i}\mathbf{t_i}^T$ with $\mathbf{X}^T\mathbf{T}$. This can be demostrated to be true, and so I do in appendix A.

Question 6

Question 7

Question 8

Question 9

Question 10

Question 11

Question 12

Question 13

Question 14

Question 15

Question 16

Question 17

Question 18

Question 19

Question 20

Question 21

Question 22

Question 23

Question 24

Appenddix A

Demonstration of $\sum_i \mathbf{x_i}\mathbf{x_i}^T = \mathbf{X}^T\mathbf{X}$

Suppose we have a matrix X:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{x_N}^T \end{bmatrix}$$

We can decompose this matrix as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \dots + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{x_n}^T \end{bmatrix}$$

From the previous it immediatly follows that its transpose can be expressed as:

$$\mathbf{X}^T = \begin{bmatrix} \mathbf{x_1} & \mathbf{0} & \dots & \mathbf{0} & \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{x_2} & \dots & \mathbf{0} & \end{bmatrix} + \dots + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{x_n} & \end{bmatrix}$$

Now if we multiply the decomposed version of the matrix $\mathbf{X}^T$ together with $\mathbf{X}$ and apply the distributive property we get:

$$\mathbf{X}^T\mathbf{X} = \left(\begin{bmatrix} \mathbf{x_1} & \mathbf{0} & \dots & \mathbf{0} & \end{bmatrix} + \dots + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{x_n} & \end{bmatrix} \right) \cdot \left(\begin{bmatrix} \mathbf{x_1}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{x_2}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} + \dots + \begin{bmatrix} \mathbf{0}^T \\ \mathbf{0}^T \\ \dots \\ \mathbf{x_n}^T \end{bmatrix} \right)$$

From which we can see that the multiplication of any corresponding matrices containing the same vector $\mathbf{x_i}$ we get:

$$\begin{bmatrix} \mathbf{0} & \dots & \mathbf{x_i} & \dots & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{0}^T \\ \dots \\ \mathbf{x_i}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} = \mathbf{x_i} \cdot \mathbf{x_i}^T$$

While by multipling two matrices containing different vectors we get:

$$\begin{bmatrix} \mathbf{0} & \dots & \mathbf{x_j} & \dots & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{0}^T \\ \dots \\ \mathbf{x_i}^T \\ \dots \\ \mathbf{0}^T \end{bmatrix} = \mathbf{0}$$

We can therfore conclude that:

$$\mathbf{X}^T\mathbf{X} = \sum_i \mathbf{x_i}\mathbf{x_i}^T$$