Machine Learning Worksheet 5

Tomas Ladek, Michael Kratzer
3602673, 3612903
tom.ladek@tum.de, mkratzer@mytum.de

Problem 1

$$
\begin{aligned}
E_{\mathcal{D}}(w) &= \frac{1}{2}\sum_{n=1}^{N} T_n[\mathbf{W}^T\phi(x_n) - Z_n]^2 \\
&= \frac{1}{2}[T(\Phi W - Z)^T][T(\Phi W - Z)]
\end{aligned}
$$

with

$$
T = \begin{pmatrix} \sqrt{T_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{T_n} \end{pmatrix}
$$

Now, finding the optimal W so that this error function is minimal (using the knowledge about the derivative from the slides and the Matrix Cookbook):

$$
\begin{aligned}
\nabla_W E_{\mathcal{D}}(W) &= \frac{\partial}{\partial W}\frac{1}{2}[T(\Phi W - Z)^T][T(\Phi W - Z)] \\
&= -T^2\Phi^T(\Phi W - Z)
\end{aligned}
$$

Further, setting this to 0 and solving for W:

$$
\begin{aligned}
-T^2\Phi^T(\Phi W - Z) &= 0 \\
-T^2\Phi^T\Phi W + T^2\Phi^T Z &= 0 \\
T^2\Phi^T\Phi W &= T^2\Phi^T Z \\
(T^2\Phi^T\Phi)^{-1}(T^2\Phi^T\Phi W) &= (T^2\Phi^T\Phi)^{-1}T^2\Phi^T Z \\
W &= T^{-2}T^2(\Phi^T\Phi)^{-1}\Phi^T Z \\
W &= (\Phi^T\Phi)^{-1}\Phi^T Z
\end{aligned}
$$

By carefully choosing values of T_n, the error function can be made less sensitive against outliers (that can have a significant impact on the variance of the data noise). Also duplicate data points can be weighted accordingly (e.g. small value of T_i for the data point copies: the error is also made smaller).

Problem 2

The following calculation assumes, that p is equal the the number of rows in the matrix Φ. The normal least squares algorithm is defined as:

$$
E_D(w) = \frac{1}{2}\sum^{N}[W^T\Phi(x_n) - z_n]^2 = \frac{1}{2}(\Phi W - Z)^T(\Phi W - Z) \tag{1}
$$

Augmenting the matrix and the vector by the given values leeds to:

$$\frac{1}{2}(\begin{pmatrix} \phi_0(X_1) & \dots & \phi_{m-1}(X_1) \\ \vdots & \ddots & \vdots \\ \phi_0(X_N) & \dots & \phi_{m-1}(X_N) \\ \sqrt{\lambda} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda} \end{pmatrix} \begin{pmatrix} w_0 \\ \vdots \\ w_{m-1} \end{pmatrix} - \begin{pmatrix} z_0 \\ \vdots \\ z_{m-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix})^T \begin{pmatrix} \phi_0(X_1) & \dots & \phi_{m-1}(X_1) \\ \vdots & \ddots & \vdots \\ \phi_0(X_N) & \dots & \phi_{m-1}(X_N) \\ \sqrt{\lambda} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda} \end{pmatrix} \begin{pmatrix} w_0 \\ \vdots \\ w_{m-1} \end{pmatrix} - \begin{pmatrix} z_0 \\ \vdots \\ z_{m-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}) \tag{2}$$

$$= \frac{1}{2} \begin{pmatrix} w_0\phi_0(X_1) - z_0 + \dots + w_{m-1}\phi_{m-1}(X_1) - z_{m-1} \\ \vdots \\ w_0\phi_0 X_N - z_0 + \dots + w_{m-1}\phi_{m-1}(X_N) - z_{m-1} \\ \sqrt{\lambda} w_0 \\ \vdots \\ \sqrt{\lambda} w_{m-1} \end{pmatrix}^T \begin{pmatrix} w_0\phi_0(X_1) - z_0 + \dots + w_{m-1}\phi_{m-1}(X_1) - z_{m-1} \\ \vdots \\ w_0\phi_0 X_N - z_0 + \dots + w_{m-1}\phi_{m-1}(X_N) - z_{m-1} \\ \sqrt{\lambda} w_0 \\ \vdots \\ \sqrt{\lambda} w_{m-1} \end{pmatrix} \tag{3}$$

$$= \frac{1}{2}(\sum^N [W^T\Phi(x_n)]^2 + \lambda w_0^2 + \dots + \lambda w_{m-1}^2) = \frac{1}{2}(\sum^N [W^T\Phi(x_n)]^2 + \lambda ||w||) = \frac{1}{2}\sum^N [W^T\Phi(x_n)]^2 + \frac{\lambda}{2}||w|| = \widetilde{E}_D(w) \tag{4}$$

Problem 3

$$\begin{aligned} p(W,\beta|Z,X) &\propto p(Z|X,W,\beta)p(W|\beta)p(\beta) && (5) \\ &= p(Z|X,W,\beta)p(W,\beta) && (6) \\ &= \prod_{n=1}^N \mathcal{N}(Z_n|W^T\phi(X_n),\beta^{-1})\mathcal{N}(W|M_0,\beta^{-1}S_0)Gam(\beta|a_0,b_0) && (7) \end{aligned}$$

Problem 4

With $y \sim \mathcal{N}(10,4)$, then $p(y > 10) = 0.5$. 10 is exactly the expected value of the normal distribution. To show this easily we can transform the random variable to a standard gaussian distribution.

$$Z = \frac{y-\mu}{\sigma} = \frac{y-10}{2} \tag{8}$$

To get the probability for $y > 10$ we just have to plug in 10 into the formular for Z and look ap the corresponding value in the cummulative table. This results in $p(y > 10) = 0.5$

Problem 5

Given $y \sim \mathcal{N}(5x+10,4)$ and $x = 1$, then $y \sim \mathcal{N}(15,4)$. Then the expected value of y is just 15.

Problem 6

We just square the norm, which is ok because we are only interested in distances.

$$
\begin{aligned}
||x_1 - x_2||_2^2 &= < x_1 - x_2, x_1 - x_2 > && (9)\\
&= (x_1 - x_2)^T (x_1 - x_2) && (10)\\
&= (x_1^T - x_2^T)(x_1 - x_2) && (11)\\
&= x_1^T x_1 - x_1^T x_2 - x_2^T x_1 + x_2^T x_2 && (12)\\
&= x_1^T x_1 + x_2^T x_2 - 2x_1^T x_2 && (13)
\end{aligned}
$$

Using the kernel: $k(x_1, x_1) + k(x_2, x_2) - 2k(x_1, x_2)$

Problem 7

$$
\begin{aligned}
k(x_1, x_2) &= exp\{-\frac{1}{2}||x_1 - x_2||^2\} && (14)\\
&= exp\{-\frac{1}{2}(x_1^T x_1 + x_2^T x_2 - 2x_1^T x_2)\} && (15)\\
&= exp\{-\frac{1}{2}x_1^T x_1\} exp\{-\frac{1}{2}x_2^T x_2\} exp\{x_1^T x_2\} && (16)\\
&= f(x_1) k(x_1, x_2) f(x_2) && (17)
\end{aligned}
$$

with $f(x) = exp\{-\frac{1}{2}x^T x\}$ and $k(x_1, x_2) = exp\{x_1^T x_2\}$. $f(x)$ is only a scalar and a kernel times a scalar is also a kernel. $x_1^T x_2$ is the linear kernel. Also $exp\{k(x_1, x_2)\}$ is also a kernel.