It is not mandatory to submit solutions and sample solutions will be published in two weeks. If you choose to submit your solution, please send an e-mail from your ethz.ch address with subject Exercise2 containing a PDF (LPXor scan) to harun.mustafa@inf.ethz.ch until Tuesday, Apr 11, 2017.

Problem 1 (Kernel Composition):
Assume that $k_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $i = 1, \ldots, n$, are kernels with corresponding features mappings $\Phi_i : \mathcal{X} \to \mathbb{R}^{d_i}$. For each definition of $k$ below, prove that $k$ is also a kernel by finding the corresponding mapping $\Phi : \mathcal{X} \to \mathbb{R}^d$.

(a) $k(x, y) := x^T M y$, for $x, y \in \mathbb{R}^d$, and some symmetric positive semidefinite matrix $M \in \mathbb{R}^{d \times d}$.

(b) $k(x, y) := \sum_{i=1}^n a_i k_i(x, y)$, for $a_1, \ldots, a_n > 0$. Hint: start by proving the fact for $n = 2$, then use mathematical induction.

(c) $k(x, y) := k_i(x, y) k_j(x, y)$

Solution 1:

(a) Since $M$ is symmetric positive semi-definite, it has an eigendecomposition of the form $M = V \Sigma V^T$, where $V \in \mathbb{R}^{d \times d}$ is orthogonal, and $\Sigma \in \mathbb{R}^{d \times d}$ is diagonal containing the (non-negative) eigenvalues of $M$. Consider the feature mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ with $\Phi(x) = \Sigma^{1/2} V^T x$. Then,

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle = \left\langle \Sigma^{1/2} V^T x, \Sigma^{1/2} V^T y \right\rangle = \left(\Sigma^{1/2} V^T x\right)^T \Sigma^{1/2} V^T y = x^T V \Sigma^{1/2} \Sigma^{1/2} V^T y = x^T V \Sigma V^T y = x^T M y$$

(b) Consider the feature mapping $\Phi : \mathcal{X} \to \mathbb{R}^{d_i + d_j}$ with $\Phi(x) = [\sqrt{a_i} \Phi_i(x), \sqrt{a_j} \Phi_j(x)]$. Then,

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle = \left\langle [\sqrt{a_i} \Phi_i(x), \sqrt{a_j} \Phi_j(x)], [\sqrt{a_i} \Phi_i(y), \sqrt{a_j} \Phi_j(y)] \right\rangle = \langle \sqrt{a_i} \Phi_i(x), \sqrt{a_i} \Phi_i(y) \rangle + \langle \sqrt{a_j} \Phi_j(x), \sqrt{a_j} \Phi_j(y) \rangle = a_i k_i(x, y) + a_j k_j(x, y)$$

For the induction step, suppose that a feature map $\Phi' : \mathcal{X} \to \mathbb{R}^{\sum_{i=1}^{n-1} d_i}$ exists, inducing the kernel $k'(x, y) = \sum_{i=1}^{n-1} a_i k_i(x, y)$. Then we define the feature map $\Phi = [\Phi'(x), \sqrt{a_n} \Phi_n(x)]$ and follow the same argument as above.
(c) Consider the feature mapping $\Phi : \mathcal{X} \rightarrow \mathbb{R}^{d_1 \times d_2}$ with \( \Phi(x)_{kl} = \Phi_i(x)_{k} \Phi_j(x)_{\ell} \) with \( \langle \cdot , \cdot \rangle \) defined as the sum of all entries after point-wise multiplication. Then,
\[
k(x, y) = \langle \Phi(x), \Phi(y) \rangle = \sum_{k=1}^{d_1} \sum_{\ell=1}^{d_2} \Phi(x)_{k\ell} \Phi(y)_{k\ell} = \sum_{k=1}^{d_1} \sum_{\ell=1}^{d_2} \Phi_i(x)_{k} \Phi_j(x)_{\ell} \Phi_i(y)_{k} \Phi_j(y)_{\ell} = \langle \Phi_i(x), \Phi_j(y) \rangle \langle \Phi_j(x), \Phi_j(y) \rangle = k_i(x, y) k_j(x, y)
\]

Problem 2 (Kernelized Linear Regression):

In this exercise you will derive the kernelized version of linear regression.

(a) Prove that the following identity holds for any matrix \( B \in \mathbb{R}^{n \times m} \), and any invertible matrices \( A \in \mathbb{R}^{m \times m} \), and \( C \in \mathbb{R}^{n \times n} \).
\[
(A^{-1} + B^T C^{-1} B)^{-1} B^T C^{-1} = AB^T (BAB^T + C)^{-1}
\]

(b) Remember the solution of ridge regression, \( w^* = (X^T X + \lambda I)^{-1} X^T y \). Use the matrix identity of part (a) to prove that \( w^* \) lies in the row space of \( X \), that is, it can be written as \( w^* = X^T z^* \) for some \( z^* \in \mathbb{R}^n \).

(c) Use the result of part (b) to transform the original ridge regression loss function,
\[
R(w) = \|Xw - y\|^2 + \lambda \|w\|^2,
\]
into a new loss function \( \hat{R}(z) \), such that \( \hat{R}(z^*) = R(w^*) \), and \( z^* = \arg \min_z \hat{R}(z) \).

(d) Assuming that you are given a kernel \( k(\cdot, \cdot) \), express the kernel matrix \( K \) of the data set as a function of the data matrix \( X \), and substitute it in the new loss function \( \hat{R}(z) \) to obtain the kernelized version of the ridge regression loss function.

(e) To complete the kernelized version of ridge regression, show how you would predict the value \( y \) of a new point \( x \), assuming that you have already computed \( z^* \).

Solution 2:

(a) We multiply both sides by \( (BAB^T + C) \) from the right. The right side gives \( AB^T \), and the left hand side gives
\[
(A^{-1} + B^T C^{-1} B)^{-1} B^T C^{-1} (BAB^T + C)
\]
\[
= (A^{-1} + B^T C^{-1} B)^{-1} (B^T C^{-1} BAB^T + B^T)
\]
\[
= (A^{-1} + B^T C^{-1} B)^{-1} (B^T C^{-1} BAB^T + A^{-1} AB^T)
\]
\[
= (A^{-1} + B^T C^{-1} B)^{-1} (B^T C^{-1} B + A^{-1}) AB^T
\]
\[
= AB^T,
\]
therefore the sides are equal, which proves the identity.
(b) Using the above matrix identity with $A = \frac{1}{\lambda}I$, $B = X$, and $C = I$, we get

$$
\begin{align*}
(\lambda I + X^TX)^{-1} X^T &= \frac{1}{\lambda}X^T \left( \frac{1}{\lambda}XX^T + I \right)^{-1} \\
&= X^T (XX^T + \lambda I)^{-1}.
\end{align*}
$$

Therefore, $w^* = (X^TX + \lambda I)^{-1} X^T y = X^T (XX^T + \lambda I)^{-1} y$, and $w^*$ is in the row space of $X$, since it can be written as $w^* = X^T z^*$, if we define $z^* = (XX^T + \lambda I)^{-1} y$.

(c) For any $z \in \mathbb{R}^n$, substituting $w = X^T z$ in $R(w)$, we get

$$
\hat{R}(z) = R(X^T z) = \|XX^T z - y\|_2^2 + \lambda \|X^T z\|_2^2 = \|XX^T z - y\|_2^2 + \lambda z^T XX^T z.
$$

By definition, it holds that $R(w^*) = R(X^T z^*) = \hat{R}(z^*)$. It also holds that $z^* = \arg\min_z \hat{R}(z)$. Assume to the contrary that $\exists \bar{z}$, such that $\hat{R}(\bar{z}) < \hat{R}(z^*)$. Then, if we define $\bar{w} = X^T \bar{z}$, we get

$$
R(\bar{w}) = \hat{R}(\bar{z}) < \hat{R}(z^*) = R(w^*),
$$

which contradicts the definition of $w^*$.

(d) The kernel matrix can be written as $K = XX^T$, which we can substitute into $\hat{R}$ to get

$$
\hat{R}(z) = \|Kz - y\|_2^2 + \lambda z^T Kz.
$$

(e) We would predict the value of point $x$ as

$$
y = w^T x = (X^T z)^T x = z^T X x = \sum_{i=1}^{n} z_i x_i^T x = \sum_{i=1}^{n} z_i k(x_i, x),
$$

from which we see that we can also predict using only the kernel, without the need for any operations in the feature space.

Problem 3 (Classifiers):
The following figure shows three classifiers trained on the same data set. One of them is a \( k \)-nearest neighbor classifier, and the other two are support vector machines (SVMs) using a quadratic and a Gaussian kernel respectively. Based on the shape of the decision boundary, can you guess which plot corresponds to which classifier?

Solution 3:
Plot (b) corresponds to the quadratic kernel SVM. Because of the quadratic kernel, the decision boundary is a second-order curve, in this case, an ellipse. Plot (c) corresponds to the \( k \)-NN classifier. The decision boundary is notably non-smooth, because of the nearest neighbor classification rule. (Increasing \( k \) would make it smoother.) Finally, plot (a) corresponds to the Gaussian kernel SVM.