Exercises
Learning and Intelligent Systems
SS 2017

Series 2, Mar 31, 2017 (Kernels)

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Gunnar Rätch and Prof. Dr. Thomas Hofmann
Web: `http://las.inf.ethz.ch/teaching/lis-s17/`
Email questions to:
Harun Mustafa, `harun.mustafa@inf.ethz.ch`

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 (LaTeXor scan) to `harun.mustafa@inf.ethz.ch` until Tuesday, Apr 11, 2017.

Problem 1 (Kernel Composition):

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

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

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

(c) $k(\boldsymbol{x}, \boldsymbol{y}) := k_i(\boldsymbol{x}, \boldsymbol{y}) k_j(\boldsymbol{x}, \boldsymbol{y})$

Solution 1:

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

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

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

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

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 $k'(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n-1} a_i k_i(\boldsymbol{x}, \boldsymbol{y})$. Then we define the feature map $\Phi = [\Phi'(\boldsymbol{x}), \sqrt{a_n}\Phi(\boldsymbol{x})]$ and follow the same argument as above.

(c) Consider the feature mapping $\Phi : \mathcal{X} \to \mathbb{R}^{d_i \times d_j}$ with $\Phi(\boldsymbol{x})_{kl} = \Phi_i(\boldsymbol{x})_k \Phi_j(\boldsymbol{x})_\ell$ with $\langle \cdot, \cdot \rangle$ defined as the sum of all entries after point-wise multiplication. Then,

$$\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{y}) &= \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle \\
&= \sum_{k=1}^{d_i} \sum_{\ell=1}^{d_j} \Phi(\boldsymbol{x})_{k\ell} \Phi(\boldsymbol{y})_{k\ell} \\
&= \sum_{k=1}^{d_i} \sum_{\ell=1}^{d_j} \Phi_i(\boldsymbol{x})_k \Phi_j(\boldsymbol{x})_\ell \Phi_i(\boldsymbol{y})_k \Phi_j(\boldsymbol{y})_\ell \\
&= \langle \Phi_i(\boldsymbol{x}), \Phi_i(\boldsymbol{y}) \rangle \langle \Phi_j(\boldsymbol{x}), \Phi_j(\boldsymbol{y}) \rangle \\
&= k_i(\boldsymbol{x}, \boldsymbol{y}) k_j(\boldsymbol{x}, \boldsymbol{y})
\end{aligned}$$

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 $\mathbf{B} \in \mathbb{R}^{n \times m}$, and any invertible matrices $\mathbf{A} \in \mathbb{R}^{m \times m}$, and $\mathbf{C} \in \mathbb{R}^{n \times n}$.

$$\left(\mathbf{A}^{-1} + \mathbf{B}^T \mathbf{C}^{-1} \mathbf{B}\right)^{-1} \mathbf{B}^T \mathbf{C}^{-1} = \mathbf{A}\mathbf{B}^T \left(\mathbf{B}\mathbf{A}\mathbf{B}^T + \mathbf{C}\right)^{-1}$$

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

(c) Use the result of part (b) to transform the original ridge regression loss function,

$$R(\boldsymbol{w}) = \|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2,$$

into a new loss function $\hat{R}(\boldsymbol{z})$, such that $\hat{R}(\boldsymbol{z}^*) = R(\boldsymbol{w}^*)$, and $\boldsymbol{z}^* = \operatorname{argmin}_{\boldsymbol{z}} \hat{R}(\boldsymbol{z})$.

(d) Assuming that you are given a kernel $k(\cdot, \cdot)$, express the kernel matrix $\mathbf{K}$ of the data set as a function of the data matrix $\mathbf{X}$, and substitute it in the new loss function $\hat{R}(\boldsymbol{z})$ 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 $\boldsymbol{x}$, assuming that you have already computed $\boldsymbol{z}^*$.

Solution 2:

(a) We multiply both sides by $\left(\mathbf{B}\mathbf{A}\mathbf{B}^T + \mathbf{C}\right)$ from the right. The right side gives $\mathbf{A}\mathbf{B}^T$, and the left hand side gives

$$\begin{aligned}
&\left(\mathbf{A}^{-1} + \mathbf{B}^T \mathbf{C}^{-1} \mathbf{B}\right)^{-1} \mathbf{B}^T \mathbf{C}^{-1} \left(\mathbf{B}\mathbf{A}\mathbf{B}^T + \mathbf{C}\right) \\
=\ &\left(\mathbf{A}^{-1} + \mathbf{B}^T \mathbf{C}^{-1} \mathbf{B}\right)^{-1} \left(\mathbf{B}^T \mathbf{C}^{-1} \mathbf{B}\mathbf{A}\mathbf{B}^T + \mathbf{B}^T\right) \\
=\ &\left(\mathbf{A}^{-1} + \mathbf{B}^T \mathbf{C}^{-1} \mathbf{B}\right)^{-1} \left(\mathbf{B}^T \mathbf{C}^{-1} \mathbf{B}\mathbf{A}\mathbf{B}^T + \mathbf{A}^{-1}\mathbf{A}\mathbf{B}^T\right) \\
=\ &\left(\mathbf{A}^{-1} + \mathbf{B}^T \mathbf{C}^{-1} \mathbf{B}\right)^{-1} \left(\mathbf{B}^T \mathbf{C}^{-1} \mathbf{B} + \mathbf{A}^{-1}\right) \mathbf{A}\mathbf{B}^T \\
=\ &\mathbf{A}\mathbf{B}^T,
\end{aligned}$$

therefore the sides are equal, which proves the identity.

(b) Using the above matrix identity with $\mathbf{A} = \frac{1}{\lambda}\mathbf{I}$, $\mathbf{B} = \mathbf{X}$, and $\mathbf{C} = \mathbf{I}$, we get

$$\begin{aligned}
&\left(\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\
&= \frac{1}{\lambda}\mathbf{X}^T\left(\frac{1}{\lambda}\mathbf{X}\mathbf{X}^T + \mathbf{I}\right)^{-1} \\
&= \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}\right)^{-1}.
\end{aligned}$$

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

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

$$\begin{aligned}
\hat{R}(\boldsymbol{z}) &= R(\mathbf{X}^T\boldsymbol{z}) \\
&= \|\mathbf{X}\mathbf{X}^T\boldsymbol{z} - \boldsymbol{y}\|_2^2 + \lambda\|\mathbf{X}^T\boldsymbol{z}\|_2^2 \\
&= \|\mathbf{X}\mathbf{X}^T\boldsymbol{z} - \boldsymbol{y}\|_2^2 + \lambda\boldsymbol{z}^T\mathbf{X}\mathbf{X}^T\boldsymbol{z}.
\end{aligned}$$

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

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

which contradicts the definition of $\boldsymbol{w}^*$.

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

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

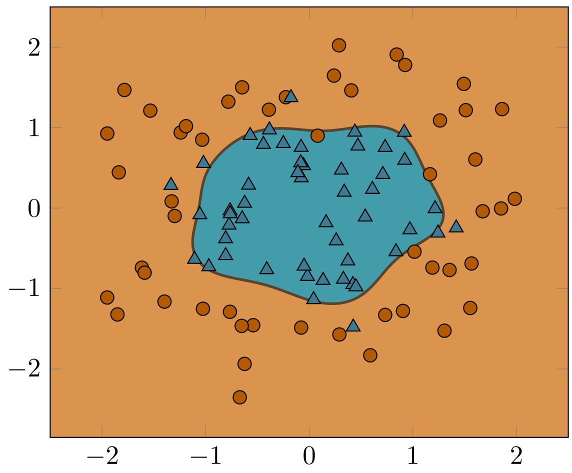
(e) We would predict the value of point $\boldsymbol{x}$ as

$$\begin{aligned}
\boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{x} &= \left(\mathbf{X}^T\boldsymbol{z}\right)^T\boldsymbol{x} \\
&= \boldsymbol{z}^T\mathbf{X}\boldsymbol{x} \\
&= \sum_{i=1}^n z_i\boldsymbol{x}_i^T\boldsymbol{x} \\
&= \sum_{i=1}^n z_i k(\boldsymbol{x}_i, \boldsymbol{x}),
\end{aligned}$$

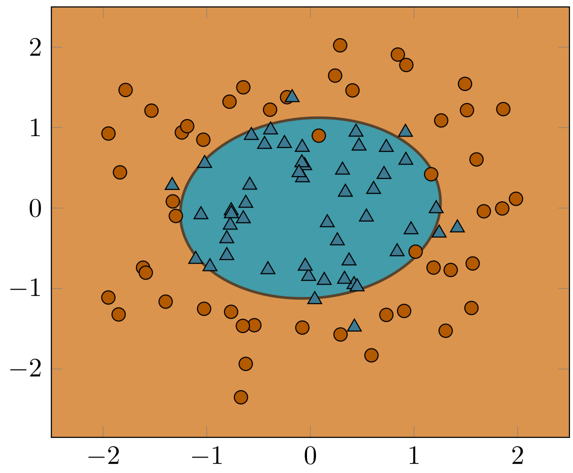
from which we see that we can also predict using only the kernel, without the need for any operations in the feature space.

Problem 3 (Classifiers):

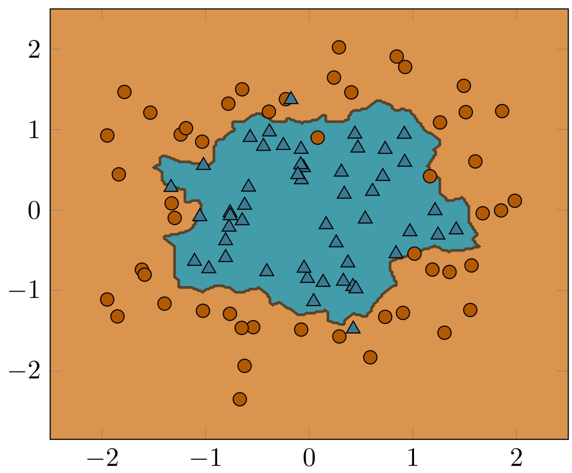
The following figure shows three classifiers trained on the same data set. One of them is a k-nearest neighbor classifier, and the other two are support vector machines (SVMs) using a quadratic and a Gaussian kernel respectively. Based on the shape of the decision boundary, can you guess which plot corresponds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



(a)



(b)



(c)

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.