NYU Paris **Spring 2019**

CSCI-UA 9473 Introduction to Machine Learning

Problem Set 4: Kernels and SVM (Part II)

The exercises vary in difficulty. More advanced exercises are marked with one or two stars (*)

1. Kernels (II)

Exercise 1 (source: Bishop)**.** *Kernels are encoding similarity/dissimilarity between points and thus cannot be defined arbitrarily. A necessary and sufficient condition for a function* $k(\boldsymbol{x}, \boldsymbol{x}')$ *to be a valid kernel is for the Gram matrix* $\boldsymbol{K}$ *whose entries* (i, j) *are defined as* $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$*, to be positive semidefinite for all possible choices of points* $\boldsymbol{x}_i, \boldsymbol{x}_j$[1]*. The easiest way to prove that a Kernel is valid is to start from a simple Kernel, e.g.* $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ *and*

- *Using Fig. 4 prove that the Gaussian kernel,* $\exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$*, is a valid kernel.*
- *Verify properties* (6)*,* (7)*,* (12) *and* (13)*.*

Exercise 2 (Bishop)**.** *The nearest neighbor classifier assigns a new input vector* $\boldsymbol{x}$ *to the same class as that of the nearest input vector* $\boldsymbol{x}_n$ *from the training set. In the simplest case, the distance is taken to be the Euclidean metric* $\|\boldsymbol{x} - \boldsymbol{x}_n\|^2$*. By expressing this rule in terms of scalar products, and then making use of kernels, Give the kernel formulation of the nearest neighbor classifier.*

Exercise 3 (** Bishop)**.** *We consider a non linear feature mapping* $\boldsymbol{\phi}(\boldsymbol{x})$ *and we assume that the loss or error function reads as*

$$L(\boldsymbol{w}) = f(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}), \ldots, \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})) + g(\boldsymbol{w}^T\boldsymbol{w}) \tag{1}$$

where g *is increasing. By writing* $\boldsymbol{w}$ *in the form*

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n \boldsymbol{\phi}(\boldsymbol{x}_n) + \boldsymbol{w}_\perp \tag{2}$$

where $\boldsymbol{w}_\perp \boldsymbol{\phi}(\boldsymbol{x}_n) = 0$ *for all* $\boldsymbol{x}_n$*, show that the value of* $\boldsymbol{w}$ *that minimizes* $L(\boldsymbol{w})$ *takes the form of a linear combination of basis functions* $\boldsymbol{\phi}(\boldsymbol{x}_n)$*,* $n = 1, \ldots, N$*.*

Exercise 4 (source: CMU 10-701, ML, Edmunds, Xing, Gormley)**.** *Consider the one dimensional dataset shown in Fig. 1.*

- *Can you think of a* $1D$ *transformation (i.e* $f(x) : \mathbb{R} \mapsto \mathbb{R}$*) that would make those points linearly separable?*
- *Same question but for with a 2D transformation.*

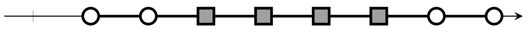

Figure 1. Material for Exercise 4

[1] Recall that a matrix $\boldsymbol{A}$ is positive semidefinite if

- $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x}$
- All the eigenvalues of $\boldsymbol{A}$ are non negative
- $\boldsymbol{A} = \boldsymbol{R}\boldsymbol{R}^T$ for some matrix $\boldsymbol{R}$

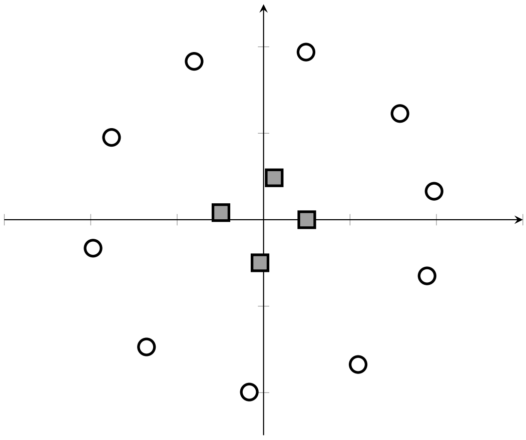

FIGURE 2. Material for Exercise 5

Exercise 5 (source: CMU 10-701, ML, Edmunds, Xing, Gormley). *Consider the dataset illustrated in Fig. 2. Find a* $1D$ *transformation,* $f\ \mathbb{R}^2 \mapsto \mathbb{R}$ *which maps this dataset onto a line on which it becomes linearly separable.*

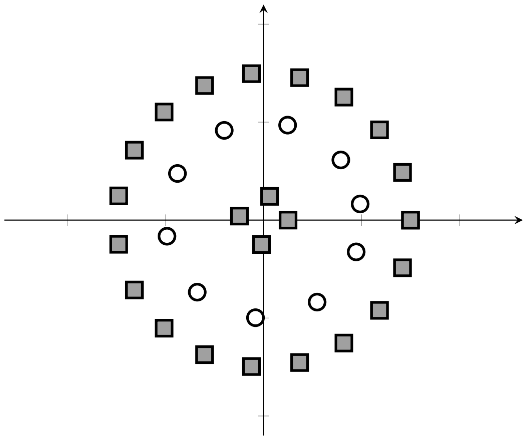

FIGURE 3. Material for Exercise 6

Exercise 6 (source: CMU 10-701, ML, Edmunds, Xing, Gormley). *Consider Fig. 3.*

- *Building upon the ideas of Exercises* (4) *and* (5)*, try to define a transformation that makes this dataset linearly separable.*
- *What is the expression of the kernel corresponding to the transformation you define above?*

Exercise 7 (source: Jaakkola). *Consider a dataset with 2 points in 1D:* $(x_1 = 0, y_1 = -1)$ *and* $(x_2 = \sqrt{2}, y_2 = 1)$*. Consider mapping each point to 3D space using the feature vector* $\phi(x) = [1, \sqrt{2}x, x^2]$*. The max margin classifier has the form*

$$\min \quad \|\boldsymbol{w}\|^2 \tag{3}$$

$$\text{subject to} \quad y_1(\boldsymbol{w}^T\phi(\boldsymbol{x}_1) + w_0) \geq 1 \tag{4}$$

$$y_2(\boldsymbol{w}^T\phi(\boldsymbol{x}_2) + w_0) \geq 1 \tag{5}$$

- *Write down a vector that is parallel to the optimal vector* $\boldsymbol{w}$*. (hint: recall that* $\boldsymbol{w}$ *is perpendicular to the decision boundary between the two points in the 3D scene)*
- *What is the value of the margin that is achieved by this* $\boldsymbol{w}$*? (recall that the margin is the distance from each support vector to the decision boundary. Then think about the geometry of* 2 *points in space with a line separating one from the other)*

- *Solve for $\boldsymbol{w}$ using the fact that the margin is equal to $1/\|\boldsymbol{w}\|$.*
- *Solve for w_0 using your value for $\boldsymbol{w}$ as well as the optimization formulation above. (Hint: the points will be on the decision boundaries so the inequalities will be tight, i.e equality will hold.)*
- *Write down the form of the discriminant function $f(x) = w_0 + \boldsymbol{w}^T\boldsymbol{\phi}(x)$ as an explicit function of x.*

For any valid kernels $k_1(\boldsymbol{x},\boldsymbol{x}')$ and $k_2(\boldsymbol{x},\boldsymbol{x}')$, the following new kernels remain valid

$$k(\boldsymbol{x},\boldsymbol{x}') = ck_1(\boldsymbol{x},\boldsymbol{x}') \tag{6}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = f(\boldsymbol{x})k_1(\boldsymbol{x},\boldsymbol{x}')f(\boldsymbol{x}') \tag{7}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = q(k_1(\boldsymbol{x},\boldsymbol{x}')) \tag{8}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = \exp(k_1(\boldsymbol{x},\boldsymbol{x}')) \tag{9}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = k_1(\boldsymbol{x},\boldsymbol{x}') + k_2(\boldsymbol{x},\boldsymbol{x}') \tag{10}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = k_1(\boldsymbol{x},\boldsymbol{x}')k_2(\boldsymbol{x},\boldsymbol{x}') \tag{11}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = k_3(\phi(\boldsymbol{x}),\phi(\boldsymbol{x}')) \tag{12}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}' \tag{13}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = k_a(\boldsymbol{x}_a,\boldsymbol{x}_{a'}) + k_b(\boldsymbol{x}_b,\boldsymbol{x}_{b'}) \tag{14}$$
$$k(\boldsymbol{x},\boldsymbol{x}') = k_a(\boldsymbol{x}_a,\boldsymbol{x}_{a'})k_b(\boldsymbol{x}_b,\boldsymbol{x}_{b'}) \tag{15}$$

In the equalities above, $c > 0$ is a positive constant, f is any function, q is a polynomial with non negative coefficients, $\phi(\boldsymbol{x})$ is a function from $\boldsymbol{x}$ to $\mathbb{R}^M$, $k_3(\cdot,\cdot)$ is a valid kernel in $\mathbb{R}^M$, $\boldsymbol{A}$ is a symmetric positive semidefinite matrix

Figure 4. Kernel Properties. Source: Bishop, Pattern Recognition and ML.

2. Support Vector machines (SVM)

Support vector machines (a.k.a sparse vector/kernel machines) extend the traditional separation hyperplane by looking for the plane that maximizes the "margin" which is defined as the perpendicular distance of the closest point to the separating hyperplane. Recall that the distance of any prototype $\boldsymbol{x}_n$ (or feature $\boldsymbol{\phi}(\boldsymbol{x}_n)$) with target $t_n = +1/-1$, to the hyperplane $y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ can be written as

$$\frac{t_n y(\boldsymbol{x}_n)}{\|\boldsymbol{w}\|} = \frac{t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b)}{\|\boldsymbol{w}\|} \tag{16}$$

Finding the maximum margin hyperplane can then be done by minimizing the distance over the prototypes $\boldsymbol{x}_n$ and then looking for the plane that maximizes this smallest distance, i.e.

$$\underset{\boldsymbol{w},b}{\text{argmax}} \left\{\frac{1}{\|\boldsymbol{w}\|}\min_n \left[t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b)\right]\right\} \tag{17}$$

One can show that the classifier which solves this optimization problem has the form

$$y(\boldsymbol{x}) = \sum_{n\in\mathcal{S}} a_n t_n k(\boldsymbol{x},\boldsymbol{x}_n) + b \tag{18}$$

Where

$$b = \frac{1}{N_{\mathcal{S}}}\sum_{n\in\mathcal{S}}\left(t_n - \sum_{m\in\mathcal{S}} a_m t_m k(\boldsymbol{x}_n,\boldsymbol{x}_m)\right) \tag{19}$$

Here $\mathcal{S}$ denotes the set of support vectors (that is the vectors that are lying on the boundary) and $k(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x}')$. This is the second important properties of those classifiers and the reason why they are called "spare" kernel machines.

n	$x_{n,1}$	$x_{n,2}$	t_n	a_n
1	1	2	-1	0
2	2	2	-1	*0*
3	3	1	-1	0
4	3	3	-1	1/4
5	2	9	$+1$	0
6	4	6	$+1$	1/8
7	6	4	$+1$	1/8
8	6	5	$+1$	0

TABLE 1. Data for exercise 11.

Exercise 8 (Warm-up: distance between a point and a plane)**.** *Find the distance between* $(3, 7, -5)$ *and the following*

- *The* x*-*y *plane*
- *The* x*-*z *plane*
- *The* y*-*z *plane*

Exercise 9. *Quadratic optimization programs are programs of the form*

$$\min \quad \frac{1}{2}\boldsymbol{x}^T\boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x} \tag{20}$$

$$\textit{subject to} \quad \boldsymbol{G}\boldsymbol{x} \geq \boldsymbol{h}, \quad \textit{and} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{21}$$

Show that the calculation of the maximum margin hyperplane can be written under that form.

Exercise 10. *We consider a general hyperplane* $(\boldsymbol{w}, b)$ *where* $\boldsymbol{w} = (w_1, \ldots, w_N)$ *and* $b \in \mathbb{R}$*. Show that for any two points* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$ *that are lying on the plane, the difference* $\boldsymbol{x}_1 - \boldsymbol{x}_2$ *is orthogonal to the plane.*

Exercise 11 (source: INFOMPR, Pattern recognition, Utrecht)**.** *we consider the data given in table 1. In this table,* x_1 *denote the first coordinate of each training point,* x_2 *the second coordinate,* t_n *is the target label and the* a_n *are the coefficients of the SVM classifier which reads as*

$$y(\boldsymbol{x}) = b + \sum_{n=1}^{N} a_n t_n \boldsymbol{x}^T \boldsymbol{x}_n \tag{22}$$

The bias b *can be computed as*

$$b = t_m - \sum_{n=1}^{N} a_n t_n \boldsymbol{x}_m^T \boldsymbol{x}_n \tag{23}$$

For any support vector $\boldsymbol{x}_m$

- *What are the support vectors in this exercise?*
- *Compute the bias* b
- *What is the SVM prediction for* $x_1 = 0$*,* $x_2 = 7$
- *Give the equation* $y = \boldsymbol{w}^T\boldsymbol{x} + b$ *of the maximum margin hyperplane and plot it together with the training points.*

Exercise 12 (adapted from ISL, James, Witten, Hastie, Tibshirani)**.** *We now consider the observations summarized in table 2.*

- *Sketch the observations.*
- *Sketch the optimal separating hyperplane and provide the equation of this hyperplane (of the form* $y(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$*)*
- *On your sketch indicate the margin of the classifier*
- *Indicate the support vectors*

- *Give the expression of the maximal margin hyperplane if we replace the seventh observation by* $(4, 2)$.
- *We associate* $+1$ *targets* t_n *to the red points and* -1 *target to the blue point. Show that the plane in item 2 can be written equivalently as*

$$y(\boldsymbol{x}) = b + \sum_{n=1}^{4} a_n k(\boldsymbol{x}, \boldsymbol{x}_n) \tag{24}$$

for some a_n *and* b *of the form*

$$b = t_m - \sum_{n=1}^{4} a_n t_n k(\boldsymbol{x}_n, \boldsymbol{x}_m), \tag{25}$$

where the $\boldsymbol{x}_n$ *are chosen among the training points and* $k(\boldsymbol{x}, \boldsymbol{x}_n) = \langle \boldsymbol{x}, \boldsymbol{x}_n \rangle$.

Exercise 13 (*Bishop)**.** *The final expression of the Maximum margin classifier is*

$$y(\boldsymbol{x}) = \sum_{n=1}^{N} a_n t_n k(\boldsymbol{x}, \boldsymbol{x}_n) + b \tag{26}$$

Show that the value ρ *of the margin for the maximum margin hyperplane is given by*

$$\frac{1}{\rho^2} = \sum_{n=1}^{N} a_n \tag{27}$$

Exercise 14 (**source: Bishop)**.** *Show that, irrespective of the dimensionality of the data space, a dataset consisting of just two data points, one from each class, is sufficient to determine the location of the maximum-margin hyperplane.*

Exercise 15 (source: CMU 10-701, ML, Edmunds, Xing, Gormley)**.** *When considering real world data, because of measurement errors, or approximations, it can happen that a dataset that should be linearly separable gets corrupted by a couple of (negligible) samples which prevent perfect linear separation. In this case, to keep the linear assumption and neglect those outliers, we consider the following extension of SVM, known as "soft margin" SVM,*

$$\min_{\boldsymbol{\beta}} \quad \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + C\sum_{i=1}^{N} \xi, \quad \textit{where } \xi \geq 0 \tag{28}$$

$$\textit{subject to} \quad y^{(i)}(\boldsymbol{\beta}^T \boldsymbol{x}^{(i)} + \beta_0) \geq 1 - \xi_i \tag{29}$$

The ξ_i *represents the slack for each data point* i*, which allows misclassification of datapoints which make the dataset non linearly separable. SVM without the addition of such slack variables are known as "hard " SVM.*

- *Intuitively, where does a data point lie relative to where the margin is when* $\xi = 0$*? Is this data point classified correctly?*
- *Intuitively, where does a data point lie relative to where the margin is when* $0 < \xi_i \leq 1$*? Is this data point classified correctly?*
- *Intuitively, where does a data point lie relative to where the margin is when* $\xi > 1$*? Is this data point classified correctly?*

Exercise 16 (source: CMU 10-701, ML, Brynn, Xing, Gormley)**.** *Recall that support vector machine can be defined through the minimization program*

$$\min \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{30}$$

$$\textit{subject to} \quad y^{(i)}(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b) \geq 1 \tag{31}$$

As for regression, it is possible to write the formulation of this problem on the kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{x}'$. *The resulting classifier is then given by*

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i (\boldsymbol{x}_i^T \boldsymbol{x}) + b \tag{32}$$

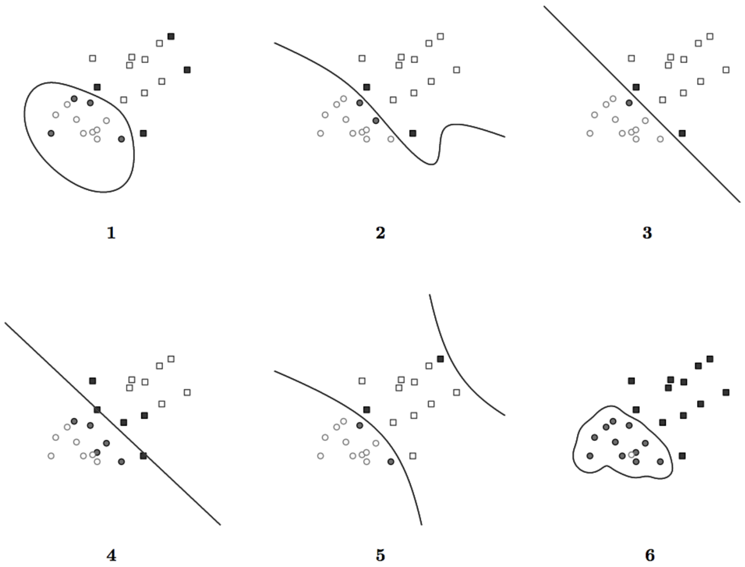

Figure 5. Material for exercise 16 (source: CMU 10-701, Machine Learning, Edmunds, Xing, Gormley)

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Table 2. Material for exercise 12

Or, in the case of a generic kernel,

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{33}$$

Consider the data/classifiers examples shown in Fig. 5. Support vectors are represented by solid circle and the rest of the data is labeled as $+1/-1$ and respectively represented by circles and squares. Each of the figures corresponds to one of the scenarios listed below. Make the connections and justify your choices.

- *A soft margin SVM with $C = 0.02$*
- *A soft margin SVM with $C = 20$*
- *A hard margin kernel SVM with $k(\boldsymbol{u}, \boldsymbol{v}) = \langle \boldsymbol{u}, \boldsymbol{v} \rangle + \langle \boldsymbol{u}, \boldsymbol{v} \rangle^2$*
- *A hard margin SVM with $k(\boldsymbol{u}, \boldsymbol{v}) = \exp(-5\|\boldsymbol{u} - \boldsymbol{v}\|^2)$*
- *A hard margin SVM with $k(\boldsymbol{u}, \boldsymbol{v}) = \exp(-\frac{1}{5}\|\boldsymbol{u} - \boldsymbol{v}\|^2)$*