CIS 520, Machine Learning, Fall 2015: Assignment 6A
Due: Friday, November 6th, 11:59pm
[54 points]

Arpit Panwar

Collaborators:

Sruthi Nair

Instructions. Please write up your responses to the following problems clearly and concisely. We encourage you to write up your responses using LaTeX; we have provided a LaTeX template, available on Canvas, to make this easier. **Submit your answers in PDF form to Canvas. We will not accept handwritten or paper copies of the homework**

Collaboration. You are allowed and encouraged to work together. You may discuss the homework to understand the problem and reach a solution in groups up to size **two students.** However, *each student must write down the solution independently, and without referring to written notes from the joint session.* **In addition, each student must write on the problem set the names of the people with whom you collaborated.** You must understand the solution well enough in order to reconstruct it by yourself. (This is for your own benefit: you have to take the exams alone.)

1 Mind Refreshing SVM [40 points]

1. **Finite Features**: You are given the data set D shown in in Figure 1 with data from a single feature X_1 in $\mathbb{R}^1$ and corresponding label $Y \in \{+, -\}$. The data set contains three positive examples at $X_1 = \{-3, -2, 3\}$ and three negative examples at $X_1 = \{-1, 0, 1\}$.

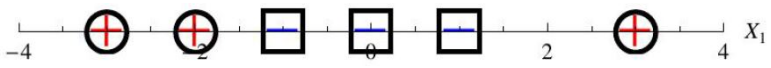

Figure 1: Dataset for SVM feature map task.

This data set, *in its current feature space*, cannot be perfectly separated using a linear separator. The data changes class twice along only one dimension; a linear classifier in one dimension can only represent a single split. In this problem, we'll investigate mapping the data to another feature space.

(a) **[2 points]** Lets define the simple feature map $\phi(u) = (u, u^2)$ which transforms points in $\mathbb{R}^1$ to points in $\mathbb{R}^2$. Apply ϕ to the data and plot the points in the new $\mathbb{R}^2$ feature space.

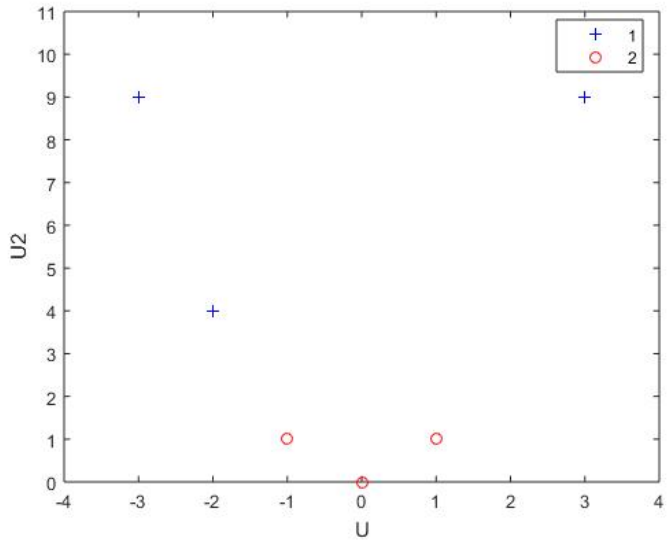

Figure 2: Transformed dataset for 1.b.

(b) **[2 points]** Can a linear separator perfectly separate the points in the new $\mathbb{R}^2$ features space induced by ϕ? If so, draw a linear separator that works. If not, explain why not.
Yes it can be linearly separated

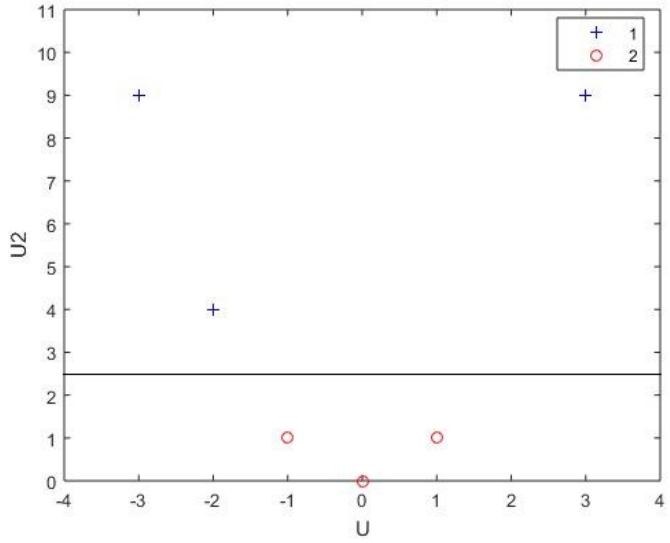

Figure 3: 1.c.

(c) **[4 points]** Give the analytic form of the kernel that corresponds to the feature map ϕ in terms of only X_1 and X_1'. Specifically define $k(X_1, X_1')$.

We have $\phi_1(u) = u$ *and* $\phi_2(u) = u^2$

$$\begin{aligned} K(X_1, X_1') &= \sum_{j=1}^{2} \phi_j(x_1)\phi_j(x_1') \\ &= \phi_1(x_1)\phi_1(x_1') + \phi_2(x_1)\phi_2(x_1') \\ &= X_1X_1' + X_1^2(X_1')^2 \\ &= X_1X_1'(1 + X_1X_1') \end{aligned}$$

(d) **[7 points]** Construct a maximum-margin separating hyperplane. This hyperplane will be a line in $\mathbb{R}^2$, which can be parameterized by its normal equation, i.e. $w_1Y_1 + w_2Y_2 + c = 0$ for appropriate choices of w_1, w_2, c. Here, $(Y_1, Y_2) = \phi(X_1)$ is the result of applying the feature map ϕ to the original feature X_1. Give the values for w_1, w_2, c. Also, explicitly compute the margin for your hyperplane. You do not need to solve a quadratic program to find the maximum margin hyperplane. Instead, let your geometric intuition guide you.

Finding the midpoint of the points $(-1, 1)$ and $(-2, 4)$

$=(-1.5, 2.5)$

Slope of the line is $\dfrac{y_2 - y_1}{x_2 - x_1}$

$=\dfrac{1}{-3}$

Line is $y = \dfrac{-1}{3}x + c$

Putting in value of midpoint and solving for c we get

$$\begin{aligned} c &= 3 \\ 3y - x - 9 &= 0 \end{aligned}$$

(e) **[4 points]** On the plot of the transformed points (from part 3), plot the separating hyperplane and the margin, and circle the support vectors.

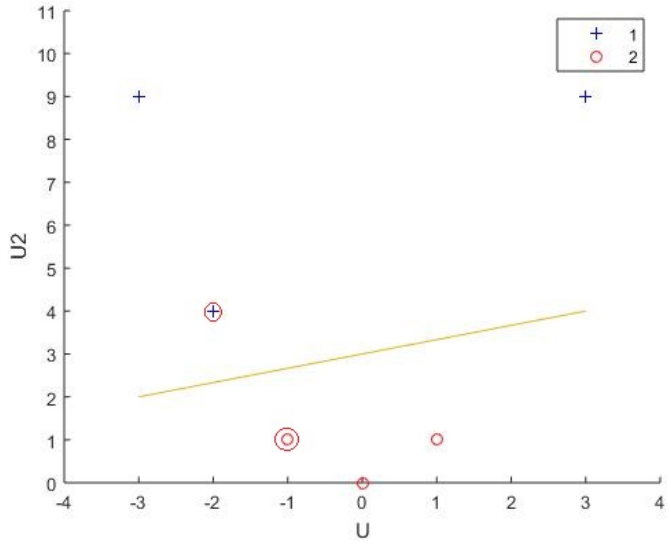

Figure 4: Transformed dataset for 1.d.

(f) **[2 points]** Draw the decision boundary separating of the separating hyperplane, in the original $\mathbb{R}^1$ feature space.

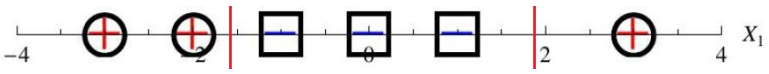

Figure 5: Transformed dataset for 1.e.

(g) **[4 points]** Compute the coefficients α and the constant b in Equation 1 for the kernel k and the support vectors $SV = \{u_1, u_2\}$ you chose in part 6. Be sure to explain how you obtained these coefficients.

$$y(x) = \text{sign}\left(\sum_{i=1}^{|SV|} \alpha_i y_i k(x, u_i) + b\right) \tag{1}$$

Think about the dual form of the quadratic program and the constraints placed on the α values.

We know that the dual of SVM is given by

$$max_{\alpha \geq 0} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

such that $\sum_i \alpha_i y_i$

Substituing values for alphas

$$\alpha_1(1) + \alpha_2(-1) = 0$$

$$\alpha_1 = \alpha_2 = \alpha$$

Now

$$b = y_i - \sum_j \alpha_j y_j k(x_j, x_i) \text{ when } \alpha \geq 0$$

$$b = 1 - \sum_j \alpha y_j k(x_j, -2)$$

$$= 1 - (\alpha k(-2,-2) - \alpha k(-1,-2))$$

$$= 1 - 14\alpha$$

Again

$$b = -1 - \sum_j \alpha y_j k(x_j, -1)$$

$$= -1 - (\alpha k(-2,-1) - \alpha k(-1,-1))$$

$$= -1 - 4\alpha$$

From above we can say that

$$1 - 14\alpha = -1 - 4\alpha$$

$$\alpha = \frac{1}{5}$$

$$b = 1 - 14\frac{1}{5}$$

$$= \frac{-9}{5}$$

(h) **[2 points]** If we add another positive ($Y = +$) point to the training set at $X_1 = 5$ would the hyperplane or margin change? Why or why not? The hyperplane will not change because the point lies beyond the margin (SVM Bound).

2. **Infinite Features Spaces and Kernel Magic**: Lets define a new (infinitely) more complicated feature transformation $\phi_n : \mathbb{R}^1 \to \mathbb{R}^n$ given in Equation 2.

$$\phi_n(x) = \left\{ e^{-x^2/2}, e^{-x^2/2}x, \frac{e^{-x^2/2}x^2}{\sqrt{2}}, \ldots, \frac{e^{-x^2/2}x^i}{\sqrt{i!}} \ldots, \frac{e^{-x^2/2}x^n}{\sqrt{n!}} \right\} \tag{2}$$

Suppose we let $n \to \infty$ and define new feature transformation in Equation 3. You can think of this feature transformation as taking some finite feature vector and producing an infinite dimensional feature vector rather than the simple two dimensional feature vector used in the earlier part of this problem.

$$\phi_\infty(x) = \left\{ e^{-x^2/2}, e^{-x^2/2}x, \frac{e^{-x^2/2}x^2}{\sqrt{2}}, \ldots, \frac{e^{-x^2/2}x^i}{\sqrt{i!}} \ldots \right\} \tag{3}$$

(a) **[3 points]** Assuming a *consistent* dataset, is there a finite set of points that cannot be linearly separated in this feature space? If so, give an example dataset. If not, explain why not. Note: A consistent dataset is one where there aren't any points $\mathbf{x}_i = \mathbf{x}_j$ where $y_i \neq y_j$. That is, no point can be labeled multiple different things.
It is linearly separable as there are no points that map to the same set of features.
We can see from the set of features that there can be no case where 2 different dataset values have the same values in every dimension.

(b) **[4 points]** We know that we can express a linear classifier using only inner products of support vectors in the transformed feature space as seen in Equation 1. It would be great if we could some how use the feature space obtained by the feature transformation ϕ_∞. However, to do this we must be able to compute the inner product of examples in this infinite vector space. Lets define the inner product between two infinite vectors $a = \{a_1, \ldots, a_i, \ldots\}$ and $b = \{b_1, \ldots, b_i, \ldots\}$ as the infinite sum given in Equation 4.

$$k(a, b) = a \cdot b = \sum_{i=1}^{\infty} a_i b_i \tag{4}$$

We cannot explicitly compute $k(a, b)$ in this form since it contains an infinite sum. However, we can re-write $k(a, b)$ in a form that is efficiently computable. Derive the efficiently computable form. *Hint*: You may want to use the Taylor series expansion of e^x which is given in Equation 5.

$$e^x = \lim_{n \to \infty} \sum_{i=0}^{n} \frac{x^i}{i!} \tag{5}$$

We can generalize the phi as $\phi_j(x) = \frac{e^{-\frac{x^2}{2}} x^{j-1}}{\sqrt{!(j-1)}}$

Substituting it in the equation for kernel

$$\begin{aligned}
K(a, b) &= \sum_{j=1}^{\infty} \phi_j(a)\phi_j(b) \\
&= \sum_{j=1}^{\infty} \frac{e^{-\frac{a^2}{2}} x^{j-1}}{\sqrt{!(j-1)}} \frac{e^{-\frac{b^2}{2}} x^{j-1}}{\sqrt{!(j-1)}} \\
&= \sum_{i=0}^{\infty} \frac{e^{-\frac{a^2}{2}} x^{i}}{\sqrt{!(i)}} \frac{e^{-\frac{b^2}{2}} x^{i}}{\sqrt{!(i)}} \\
&= e^{\frac{-a^2-b^2}{2}} \sum_{i=0}^{\infty} \frac{a^i b^i}{\sqrt{!(i)}}
\end{aligned}$$

Using taylor series

$$= e^{\frac{-a^2-b^2}{2}} e^{ab}$$

Solving the powers

$$= e^{-\frac{(a-b)^2}{2}}$$

(c) **[6 points]** Prove or disprove each claim below.

i. Suppose we translate the inputs $x'_i = x_i + x_0$ for some arbitrary x_0 before using the kernel

above in an SVM. Will my kernel function change? (i.e., does $k(x_i, x_j) = k(x'_i, x'_j)$)?

We can see the kernel from above $e^{-\frac{(a-b)^2}{2}}$

Substituting the values for a and b

$$e^{-\frac{(x_1 + x_0 - x_2 - x_0)^2}{2}}$$

$$= e^{-\frac{(x_1 - x_2)^2}{2}}$$

Thus noting changes

ii. What if we negated all the inputs $x'_i = -x_i$?
Similar to above the on substituting a with -a nothing changes as we take the square of the values which is s

iii. What about rescaling $x'_i = ax_i$ for some positive scalar a?

Substituting the values for a and b

$$e^{-\frac{(ax_1 - ax_2)^2}{2}}$$

$$= e^{-a^2\frac{(x_1 - x_2)^2}{2}}$$

As we can see rescaling changes the kernel

2 Eigenvectors

- **[7 points]** Show that if X has rank p (all its columns are linearly independent) and $n > p$ then using the p-dimensional pseudo-inverse X^+ in $\hat{w} = X^+ y$ solves the least squares problem $\hat{w} = argmin \sum_i (y-$

$Xw)^2$.

$$\begin{aligned}
\hat{w} &= argmin \sum_i (y - Xw)^2 \\
&= (y - Xw)^T (y - Xw) \\
&= X^T(y - Xw) = 0 \\
\hat{w} &= (X^T X)^{-1} X^T y
\end{aligned}$$

Now solving the middle term

$$\begin{aligned}
(X^T X)^{-1} X^T &= ((U\Lambda V^T)^T U\Lambda V^T)^{-1} (U\Lambda V^T)^T \\
&= (V\Lambda^T U^T U\Lambda V^T)^{-1} (U\Lambda V^T)^T
\end{aligned}$$

Taking in the inverse and simplifying

$$= (V\Lambda^{-1}\Lambda^{-1}V^{-1})(V\Lambda^T U^T)$$

We know inverse and transpose of orthogonal vector are the same

$$= (V\Lambda^{-1}\Lambda^{-1}\Lambda^T U^T)$$

We know that transpose of Lambda and Lambda are same as it is a symmetric diagonal matrix

$$\begin{aligned}
&= (V\Lambda^{-1}\Lambda^{-1}\Lambda U^T) \\
&= (V\Lambda^{-1}U^T) \\
&= X^+
\end{aligned}$$

Thus pseudo-inverse solves the least squares problem

- **[7 points]** We want to efficiently find the largest eigenvectors of the matrix $X^T X$ where X is nxp with $p \gg n$. Show how to do this using the largest eigenvectors of XX^T

We know that XX^T and $X^T X$ have the same eigen values. Let us assume λ is the largest eien value

$$\begin{aligned}
XX^T a &= \lambda a \\
X^T(XX^T a) &= X^T \lambda a \\
X^T X(X^T a) &= \lambda(X^T a)
\end{aligned}$$

Since we know p greater than n, it is much easier to do computation using the equation above since the complexity wi $O(n^2)$ as opposed to $O(p^2)$ if we were to directly compute $X^T X$