Assignment-2
MA6100: Maths Behind Machine Learning

1 Basics

1. Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. A least-squares solution/approximation of the matrix equation $Ax = b$ is $\hat{x} \in \mathbb{R}^n$ such that $dist(A\hat{x}, b) \leq dist(Ax, b)\ \forall\ x \in \mathbb{R}^n$.

   (a) Based on the above definition of least-squares, pose this into an optimization problem.

   (b) Suppose $m > n$ and $rank(A) = n$, then find its closed form solution.

   (c) Suppose $m < n$ and $rank(A) = m$, then find its closed form solution.

2. By now, from your MBML lectures, you are well familiar with this expression: $w^\top \phi(x)$, where $\phi : \mathcal{X} \to \mathcal{H}$, $\mathcal{X}$ is the input space[1] and $\mathcal{H}$ is known as feature[2] space[3].

   (a) Show that one can ignore the term b in the expression: $w^\top x + b$ by introducing a feature map.

   (b) Consider the equation of a circle in 2-D: $(x-a)^2 + (y-b)^2 = r^2$. Expand out the formula and show that every circular boundary is linear in some feature space[4]. (Hint: Try to write the equation as $w^\top \phi(x) = 0$)

2 Linear Regression

1. Consider the optimization problem in linear regression:

$$\min_{w \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} (w^\top \phi(x_i) - y_i)^2 \tag{1}$$

   with the training data: $\{(x_1, y_1), \ldots, (x_m, y_m)\}$, where $x_i \in \mathcal{X}$[5] and $y_i \in \mathbb{R}$.

   (a) Find a closed form expression/equation[6] to find the optimal w_L^* with the feature map $\phi : \mathcal{X} \to \mathbb{R}^n$.

[1] As an example think of some Hilbert space say $\mathbb{R}^m$

[2] This map is known as feature map

[3] Again as an example you may think of this as $\mathbb{R}^n$

[4] From this excercise one should observe the power of feature maps

[5] $\mathcal{X}$ is a finite dimensional Hilbert space. From functional analysis result $\mathcal{X} \simeq \mathbb{R}^m$

[6] Observe that the solution to this problem is similar to the solution of the least square problem: 1

(b) Now take the feature map as $\phi(x) = x$. What will be the optimal w^* in this case? Compare the results with the least-squares solution above (1).

(c) Consider the training data: $\{(1, 7), (2, 9)\}$. Find the optimal w^* for this training data.

2. Create a dataset[7] that represents the relationship between the number of hours a student studies (x) and the score they achieve on a test (y). Now consider the feature map $\phi(x) = (x, 1)$. Build a linear regression model to predict a student's test score based on the number of hours they study. (Take the cardinality of the dataset as 4)

3. Consider the Ridge regression[8] setup:

$$\min_{w \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} (w^\top \phi(x_i) - y_i)^2 + \frac{1}{2} \lambda \|w\|_2^2 \tag{2}$$

(a) Find a closed form expression/equation to find the optimal w_R^* with the feature map $\phi : \mathbb{R}^n \to \mathbb{R}^m$.

(b) Show that as $\lambda \to 0$, $w_R^* \to w_L^*$, where w_L^* is the optimal solution to the problem in (1).

4. Consider the optimization[9] problem for the l_2 regularized linear models[10] and a loss function[11], l:

$$\min_{w \in \mathbb{R}^n} C \sum_{i=1}^{m} l(w^\top \phi(x_i), y_i) + \frac{1}{2} \|w\|_2^2 \tag{3}$$

(a) Show that the optimal solution $\hat{w}$ of this problem is of the form: $\hat{w} = \sum_{i=1}^{m} \alpha_i \phi(x_i)$, i.e the optimal w can be written as a linear combination of the training datapoints in the feature space[12].

(b) Now using the result in 4(a), write the equivalent form of the optimization problem in (3).

3 PCA

1. Consider the following dataset: $\begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \begin{bmatrix} -3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$.

(a) Find the principal component for the given data.

(b) Transform the data $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$ using the principal component.

[7] Create a dataset as simple as possible so that the computation will be easier.

[8] Here l_2 regularization is used

[9] Also known as Empirical risk minimization (ERM)

[10] Linear models are the functions of the form $f = w^\top \phi(x)$

[11] Note that with the mean squared error loss it is the same as Ridge regression's ERM and with hinge loss ($l(f(x), y) = max(0, 1 - yf(x))$ it becomes SVM's ERM in eq (2)

[12] This is the celebrated **Representer Theorem**.

2. Consider the set-up of PCA where the dimensionality of the training data: $\mathcal{D} = \{x_1, \ldots, x_m\}$ wherer $x_i \in \mathbb{R}^n$ has to be reduced from n to 1.

   (a) Pose the PCA optimization problem.

   (b) Show that, without loss of generality, one can take the encoder and decoder[13] as same with unit norm, in the problem you posed.

   (c) Now perform algebraic simplifications to the problem and write the equivalent from.

   (d) Let $u_1, \ldots, u_n$ be the eigen vectors[14] corresponding to the eigenvalues of $A = \sum_{i=1}^{m} x_i x_i^\top$, which are $\lambda_1, \ldots, \lambda_n$ written in decreasing order. Now represent the encoding as $U\alpha$ where $U = [u_1 \ldots u_n]$. Now with this change of variables, further simplify the optimization problem.

   (e) Find the optimal solution of this problem.

   (f) Find the optimal encoder.

4 SVM

1. Consider the set-up of hard-margin SVM:

$$\begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \tfrac{1}{2}\|w\|_2^2, \\ \text{s.t.} \quad & y_i w^\top \phi(x_i) \geq 1 \ \forall \ i = 1, \ldots, m. \end{aligned} \tag{4}$$

   Now using[15] $w = \sum_{i=1}^{m} \alpha_i \phi(x_i)$ in equation (4) show that the equivalent form is:

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^m} \quad & \tfrac{1}{2}\alpha^\top G \alpha, \\ \text{s.t.} \quad & \sum_{j=1}^{m} \alpha_j \phi(x_j)^\top \phi(x_i) \geq 1 \ \forall \ i = 1, \ldots, m. \end{aligned} \tag{5}$$

   Where G is the gram matrix with $G_{ij} = < \phi(x_i), \phi(x_j) >$

2. Consider the training data for binary classification with inputs in $\mathbb{R}^2$:

$$\left\{ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, +1 \right), \left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, -1 \right) \right\}.$$

   (a) Take the feature map as: $\phi(x) = x$. What is the optimal solution of the problem with hard-margin SVM in (5)?

   (b) Now take the feature map as: $\phi(x) = (x, 1)$. What is the optimal solution of the problem with hard-margin SVM in (5)?

[13] Recall that for dimensionality reduction, encoder is the function which brings the dimensionality down and decoder is the function which brings back to the original dimension and in PCA, linear transformation is used for encoder and decoder.

[14] From spectral theorem they can be chosen to be orthonormal.

[15] Result in 4a

3. Consider the following binary classfication training data:

   $\left\{\left(\begin{bmatrix}1\\0\end{bmatrix}, +1\right), \left(\begin{bmatrix}-1\\0\end{bmatrix}, +1\right), \left(\begin{bmatrix}0\\1\end{bmatrix}, \text{-1}\right), \left(\begin{bmatrix}0\\-1\end{bmatrix}, \text{-1}\right)\right\}$ and the homogeneous quadratic model, which is the set of all functions of the form: $g(x) = w^\top \phi(x)$, where $w = \begin{bmatrix}w_1\\w_2\\w_3\end{bmatrix}$ is the model parameter and $\phi(x) = \begin{bmatrix}x_1^2\\x_2^2\\x_1x_2\end{bmatrix}$.

   (a) Pose the optimization problem (with expressions involving w alone) corresponding to the hard-margin SVM for choosing the optimal (homogeneous) quadratic discriminator.

   (b) Solve the above optimization problem for the optimal w.

   (c) Find the equation (Note that your expression should not involve ϕ or x or w etc) for the discriminating quadratic surface with this optimal w.

4. Show that the margin/distance between two supporting hyperplanes $w^\top \phi(x) = 1$ and $w^\top \phi(x) = -1$ is $\frac{2}{\|w\|_2}$. Observe that the prime goal of SVM is to maximize this margin between the supporting hyperplane.

5. Consider the 1-D training data for binary classification: $\left\{\begin{pmatrix}3\\+1\end{pmatrix}, \begin{pmatrix}1\\+1\end{pmatrix}, \begin{pmatrix}-1\\-1\end{pmatrix}\right\}$.

   Using the hard-margin SVM find the value of w and b and find the classifier. Predict the class label of 2 using the classifier.

6. Consider a dataset for binary classification in 1-D $\left\{\begin{pmatrix}0\\+1\end{pmatrix}, \begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}1\\-1\end{pmatrix}\right\}$

   (a) Is this dataset linearly separable? If yes find the classifier in hard-margin SVM set-up.

   (b) Take the feature map $\phi : \mathbb{R} \to \mathbb{R}^3$ as: $\phi(x) = (1, \sqrt{2}x, x^2)$. Now is this new dataset linearly separable? If yes find the classifier in hard-margin SVM set-up.

5 Kernels

1. Recall the famous **Mercer's theorem** i.e if $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a valid kernel then $\exists$ a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that $k(x, y) =< \phi(x), \phi(y) >$. Find the feature maps for the following kernels:

   (a) xy

   (b) $1 + xy + (xy)^2 + (xy)^3$

   (c) $x^\top y$

   (d) $e^{x^\top y}$

   (e) $(1 + xy)^2$

(f) $x^\top Ay$, $A > 0$

(g) $ck_1(x, y)$, $c > 0$

(h) $k_1(x, y) + k_2(x, y)$

(i) $k_1(x, y) \cdot k_2(x, y)$

(j) $e^{-\frac{1}{2}\|x-y\|}$

2. Prove that the gram matrix induced by a kernel is always positive definite.

3. Refer to the equation (3). Using **Representer Theorem** you wrote the equivalent form of the optimization problem in Q4 part (b). Now write the same equivalent form using kernels[16].

4. Write the optimization problem set-ups for the following algorithms in terms of kernels:

   (a) Kernelized Ridge Regression

   (b) Kernelized Logistic Regression

   (c) Kernelized SVM

[16]Note that when you have a feature map ϕ, then automatically you have a kernel in hand which is $k(x, y) =< \phi(x), \phi(y) >$.