Machine Learning and Computational Statistics, Spring 2015
Homework 4: Kernels and Duals

Due: Tuesday, March 3, 2015, at 4pm (Submit via NYU Classes)

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. You may include your code inline or submit it as a separate file. You may either scan hand-written work or, preferably, write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or MathJax via iPython).

1 Introduction

The problem set begins with a review of some important linear algebra concepts that we routinely use in machine learning and statistics. The solutions to each of these problems is at most a few lines long, and we've tried to give helpful hints. Everything leads up to proving a basic and important property of positive semidefinite matrices. These aren't meant to be challenging problems – just the opposite, in fact – we'd like this material to be second nature to you. We next have a couple problems on kernel methods, both of which should be quite easy if you understand the context. The last problem is both an introduction to "novelty detection" algorithms, as well as an exercise in the machinery of Lagrangian duality. This is the only problem that has a lengthy solution.

2 Positive Semidefinite Matrices

In statistics and machine learning, we use positive semidefinite matrices a lot. Let's recall some definitions from linear algebra that will be useful here:

Definition. A set of vectors $\{x_1, \ldots, x_n\}$ is **orthonormal** if $\langle x_i, x_i \rangle = 1$ for any $i \in \{1, \ldots, n\}$ (i.e. x_i has unit norm), and for any $i, j \in \{1, \ldots, n\}$ with $i \neq j$ we have $\langle x_i, x_j \rangle = 0$ (i.e. x_i and x_j are orthogonal).

Note that if the vectors are column vectors in a Euclidean space, we can write this as $x_i^T x_j = 1(i \neq j)$ for all $i, j \in \{1, \ldots, n\}$.

Definition. A matrix is **orthogonal** if it is a square matrix with orthonormal columns.

It follows from the definition that if a matrix $M \in \mathbf{R}^{n \times n}$ is orthogonal, then $M^T M = I$, where I is the $n \times n$ identity matrix. Thus $M^T = M^{-1}$, and so $MM^T = I$ as well.

Definition. A matrix M is **symmetric** if $M = M^T$.

Definition. For a square matrix M, if $Mv = \lambda v$ for some column vector v and scalar λ, then v is called an **eigenvector** of M and v is the corresponding **eigenvalue**.

Theorem (Spectral Theorem)**.** *A real, symmetric matrix $M \in \mathbf{R}^{n\times n}$ can be diagonalized as $M = Q\Sigma Q^T$, where $Q \in \mathbf{R}^{n\times n}$ is an orthogonal matrix whose columns are a set of orthonormal eigenvectors of M, and Σ is a diagonal matrix of the corresponding eigenvalues.*

Definition. A real, symmetric matrix $M \in \mathbf{R}^{n\times n}$ is **positive semidefinite (psd)** if for any $x \in \mathbf{R}^n$,

$$x^T Mx \geq 0.$$

Note that unless otherwise specified, when a matrix is described as positive semidefinite, we are implicitly assuming it is real and symmetric (or complex and Hermitian in certain contexts, though not here).

As an exercise in matrix multiplication, note that for any matrix A with columns $a_1, \ldots, a_d$, that is

$$A = \begin{pmatrix} | & & | \\ a_1 & \cdots & a_d \\ | & & | \end{pmatrix} \in \mathbf{R}^{n\times d},$$

we have

$$A^T MA = \begin{pmatrix} a_1^T Ma_1 & a_1^T Ma_2 & \cdots & a_1^T Ma_d \\ a_2^T Ma_1 & a_2^T Ma_2 & \cdots & a_2^T Ma_d \\ \vdots & \vdots & \cdots & \vdots \\ a_d^T Ma_1 & a_d^T Ma_2 & \cdots & a_d^T Ma_d \end{pmatrix}.$$

So M is psd if and only if for any $A \in \mathbf{R}^{n\times d}$, we have $\text{diag}(A^T MA) = \left(a_1^T Ma_1, \ldots, a_d^T Ma_d\right)^T \succeq 0$, where $\succeq$ is elementwise inequality, and 0 is a $d \times 1$ column vector of 0's .

1. Give an example of an orthogonal matrix that is not symmetric. (Hint: You can use a 2×2 matrix with only 0's and 1's.)
   Answer: $\begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$

2. Use the definition of a psd matrix and the spectral theorem to showthat all eigenvalues of a positive semidefinite matrix are non-negative. [Hint: In the definition of psd, make a good choice for x.]
   Answer: Suppose M is a psd matrix. So it is symmetric and can be diagnalized as $M = Q\Sigma Q^T$, where Q is an orthogonal matrix. Let $X = Q^T$, then we have

   $$XMX^T = Q^T Q\Sigma Q^T Q = \Sigma \succeq 0$$

   So, all eigenvalues of M is non-negative.

3. In this problem we show that a psd matrix is a matrix version of a non-negative scalar, in that they both have a "square root". Show that a symmetric matrix M can be expressed as $M = BB^T$ for some matrix B, if and only if M is psd. [Hint: To show $M = BB^T$ implies M is psd, use the fact that for any vector v, $v^T v \geq 0$. To show that M psd implies $M = BB^T$ for some B, use the Spectral Theorem.]
   Answers: M is a psd matrix, so it is symmetric and can be diagnalized as $M = Q\Sigma Q^T$, where Q is an orthogonal matrix.], and Σ is a diagonal matrix of the corresponding eigenvalues.

   $$\Sigma = diag\{e_1, e_2, ...e_k, 0, 0..., 0\}$$

$diag\{\}$denotes the diagnoal matrix and $e_i, i = 1...k$ are the eigenvalues of Σ. Then we have:

$$\begin{aligned}\Sigma &= diag\{e_1, e_2, ...e_k, 0, 0..., 0\} \\ &= diag\{\sqrt{e_1}, \sqrt{e_2}, ..., \sqrt{e_k}, 0, ..., 0\} * diag\{\sqrt{e_1}, \sqrt{e_2}, ..., \sqrt{e_k}, 0, ..., 0\} \\ &= \Sigma' * \Sigma'^T\end{aligned}$$

then

$$M = Q\Sigma Q^T = Q\Sigma'\Sigma'^T Q^T = (Q\Sigma')(Q\Sigma')^T = BB^T$$

3 Kernel Matrices

(Problem from Michael Jordan's Stat 241b Problem Set #1, Spring 2004)

The following problem will gives us some additional insight into what information is encoded in the kernel matrix.

1. Consider a set of vectors $S = \{x_1, \ldots, x_m\}$. Let X denote the matrix whose rows are these vectors. Form the Gram matrix $K = XX^T$. Show that knowing K is equivalent to knowing the set of pairwise distances among the vectors in S as well as the vector lengths. [Hint: The distance between x and y is given by $d(x, y) = \|x - y\|$, and the norm of a vector x is defined as $\|x\| = \sqrt{\langle x, x\rangle} = \sqrt{x^T x}$.]
Answer: $X = [x_1, x_2, .., x_m]^T$

$$K = XX^T = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix} * [x_1, x_2, .., x_m] \tag{1}$$

$$= \begin{bmatrix} x_1^T x_1 & x_1^T x_2 & \cdots & x_1^T x_m \\ x_2^T x_1 & x_2^T x_2 & \cdots & x_2^T x_m \\ \vdots & \vdots & \ddots & \vdots \\ x_m^T x_1 & x_m^T x_2 & \cdots & x_m^T x_m \end{bmatrix} \tag{2}$$

This means if we know K, we know every $x_i^T x_j$. and

$$d(x_i, x_j) = ||x_i - x_j|| = \sqrt{(x_i - x_j)^T (x_i - x_j)} \tag{3}$$

$$= \sqrt{x_i^T x_i - x_i^T x_j - x_j^T x_i + x_j^T x_j} \tag{4}$$

So, we will konw every $d(x_i, x_j)$.

4 Kernel Ridge Regression

In this problem, we complete the kernelization of ridge regression that we discussed in Lab (see `https://davidrosenberg.github.io/ml2015/docs/4.Lab.kernelizations.pdf`). To recap, the ridge regression objective function is given by

$$J(w) = ||Xw - y||^2 + \lambda||w||^2,$$

where $X \in \mathbf{R}^{n \times d}$ is the design matrix, and $\lambda > 0$ is the regularization parameter. We showed that the minimizer is given by $w^* = X^T \alpha^*$, where $\alpha^* = (\lambda I + XX^T)^{-1} y$. We can replace the Gram matrix XX^T by the data kernel matrix corresponding to any kernel satisfying Mercer's theorem. We also showed that the vector of predictions on the training points is given by $Xw = \left(XX^T\right)\left(\lambda I + XX^T\right)^{-1} y$, which is also "kernelized".

1. Give an expression for the prediction $f(x) = x^T w^*$ for a new point x, not in the training set. The expression should only involve x via inner products with other x's. [Hint: It is often convenient to define the column vector

$$k_x = \begin{pmatrix} x^T x_1 \\ \vdots \\ x^T x_n \end{pmatrix}$$

to simplify the expression.]
Answer:

$$\begin{aligned} f(x) &= x^T w^* = x^T X^T \alpha^* \\ &= x^T [x_1, x_2, ...x_n] \alpha^* \\ &= [x^T x_1, x^T x_2, ...x^T x_n] \alpha^* \\ &= [< x, x_1 >, < x, x_2 >, ... < x, x_n >] \alpha^* \end{aligned}$$

5 Novelty Detection

(Problem derived from Michael Jordan's Stat 241b Problem Set #2, Spring 2004)

A novelty detection algorithm can be based on an algorithm that finds the smallest possible sphere containing the data in feature space.

1. Let $\phi : \mathcal{X} \to F$ be our feature map, mapping elements of the input space into our "feature space" F, which is equipped with an inner product. Formulate the novelty detection algorithm described above as an optimization problem.
Answer: Let r denote the radius of the sphere and y_0 denotes the center of the sphere. Use $<,>$ to represent the inner product of two vector in the feature space. Then the problem can be expressed as:

$$\min r$$
$$s.t. \quad < \phi(X_i) - y_0, \phi(X_i) - y_0 > \leq r, \quad i = 1..n$$

2. Give the Lagrangian for this problem, and write an equivalent, unconstrained "inf sup" version of the optimization problem.
Answer: The Lagrangian function:

$$L(r, y, \alpha_i) = r + \sum_{i=1}^{n} \alpha_i (< \phi(X_i) - y_0, \phi(X_i) - y_0 > - r)$$

Then the problem convert to:

$$\inf_{r,y} \sup_{\alpha_i > 0} L(r, y, \alpha_i) = \inf_{r,y} \sup_{\alpha_i > 0} [r + \sum_{i=1}^{n} \alpha_i (< \phi(X_i) - y_0, \phi(X_i) - y_0 > - r)]$$

3. Show that we have strong duality and thus we will have an equivalent optimization problem if we swap the inf and the sup. [Hint: Use Slater's qualification conditions.]
Answer: $[r = max(< \phi(X_i), \phi(X_i) >), y = 0]$ is a feasible point under the constraints. And the Slater's condition shows that we have strong duality so long as the problem is feasible. Thus we will have an equivalent optimization problem if we swap the inf and sup.

4. Solve the inner minimization problem and give the dual optimization problem. [Note: You may find it convenient to define the kernel function $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ and to write your final problem in terms of the corresponding kernel matrix K to simplify notation.]
Answer: The inner minimization problem is:

$$\min_{y,r} L(r, y, \alpha) = min_{y,r}(r + \sum_{i=1}^{n} \alpha_i (< \phi(X_i) - y_0, \phi(X_i) - y_0 > -r))$$

Let its partial derivatives to zeros , we have:

$$\frac{\partial L}{\partial r} = 1 - \sum_{i=1}^{n} \alpha_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i = 1$$

$$\frac{\partial L}{\partial y_0} = 2y_0 - 2\sum_{i=1}^{n} \alpha_i \phi(X_i) = 0 \Rightarrow y_0 = \sum_{i=1}^{n} \alpha_i \phi(X_i)$$

subsitute y_0 and r in the function L with the fomulas above:

$$\begin{aligned} L(\alpha) &= \sum_{i=1}^{n} \alpha_i < \phi(X_i) - y_0, \phi(X_i) - y_0 > \\ &= \sum_{i=1}^{n} \alpha_i^2 < \phi(X_i), \phi(X_i) > -2 * \sum_{i=1}^{n} \alpha_i < \phi(X_i), \sum_{j=1}^{n} \alpha_j \phi(X_j) > + \\ &\quad <, \sum_{i=1}^{n} \alpha_i \phi(X_i), \sum_{j=1}^{n} \alpha_j \phi(X_j) > \\ &= \sum_{i=1}^{n} \alpha^2 k(X_i, X_i) - \sum_{i=1,j=1}^{n} \alpha_i \alpha_j k(X_i, X_j) \end{aligned}$$

So, the dual optimization problem is:

$$\begin{aligned} &obj. && \max [\sum_{i=1}^{n} \alpha_i^2 k(X_i, X_i) - \sum_{i=1,j=1}^{n} \alpha_i \alpha_j k(X_i, X_j)] \\ &s.t. && \alpha_i > 0 \\ & && \sum_{i=1}^{n} \alpha_i = 1 \end{aligned}$$

5. Write an expression for the optimal sphere in terms of the solution to the dual problem.
Answer:

$$\begin{aligned} y_0 &= \sum_{i=1}^{n} \alpha_i \phi(X_i) \\ r &= \max_{i=1,..,n} [< \phi(X_i) - y_0, \phi(X_i) - y_0 >] \end{aligned}$$

6. Write down the complementary slackness conditions for this problem, and characterize the points that are the "support vectors".
Answer: The complementary slackness conditions for this problem is :

$$\alpha_i^*(||\phi(X_i) - y_0^*||^2 - r^*) = 0$$

The support vectors are the X_i, which subject to $||\phi(X_i) - y_0^*||^2 = r^*$. With the complementary slackness condition, it is equivalent to the X_i whose corresponding $\alpha_i > 0$

7. Briefly explain how you would apply this algorithm in practice to detect "novel" instances.
Answer: For any new point X, compute $||\phi(X) - y_0^*||^2$, compare the result with r^*. If it greater than r^*, we know the new point is a "novel" instance.

8. [Optional] Redo this problem allowing some of the data to lie outside of the sphere, where the number of points outside the sphere can be increased or decreased by adjusting a parameter. (Hint: Use slack variables).

6 Feedback (not graded)

1. Approximately how long did it take to complete this assignment?

2. Any other feedback?