Machine Learning and Computational Statistics, Spring 2016
Homework 4: Kernels, Duals, and Trees

Due: Tuesday, March 22, 2016, at 6pm (Submit via NYU Classes)

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single file, either HTML or PDF. 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

This problem set is entirely written – we'll return to coding problems on the next problem set. 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. These aren't meant to be very 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: the first explores what geometric information about the data is stored in the kernel matrix, and the second revisits kernel ridge regression with a direct approach, rather than using the Representer Theorem. The last required problem has some exercises related to decision trees. We also have three optional problems this week. The first completes the proof of the Representer Theorem that we discussed in lecture. The second applies Lagrangian duality to show the equivalence of Tikhonov and Ivanov regularization. And the third introduces an approach to "novelty" or "anomoly" detection as an exercise in the machinery of Lagrangian duality.

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 $x \in \mathbf{R}^n$,

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

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

$$A^T M A = \begin{pmatrix} a_1^T M a_1 & a_1^T M a_2 & \cdots & a_1^T M a_d \\ a_2^T M a_1 & a_2^T M a_2 & \cdots & a_2^T M a_d \\ \vdots & \vdots & \cdots & \vdots \\ a_d^T M a_1 & a_d^T M a_2 & \cdots & a_d^T M a_d \end{pmatrix}.$$

So M is psd if and only if for any $A \in \mathbf{R}^{n\times d}$, we have $\text{diag}(A^T M A) = \left(a_1^T M a_1, \ldots, a_d^T M a_d\right)^T \succeq 0$, 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.)

2. Use the definition of a psd matrix and the spectral theorem to show that all eigenvalues of a positive semidefinite matrix M are non-negative. [Hint: By Spectral theorem, $\Sigma = Q^T M Q$ for some Q. What if you take $A = Q$ in the "exercise in matrix multiplication" described above?]

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.]

3 Positive Definite Matices

Definition. A real, symmetric matrix $M \in \mathbf{R}^{n\times n}$ is **positive definite** (spd) if for any $x \in \mathbf{R}^n$ with $x \neq 0$,

$$x^T M x > 0.$$

1. Show that all eigenvalues of a symmetric positive definite matrix are positive. [Hint: You can use the same method as you used for psd matrices above.]

2. Let M be a symmetric positve definite matrix. By the spectral theorem, $M = Q\Sigma Q^T$, where Σ is a diagonal matrix of the eigenvalues of M. By the previous problem, all diagonal entries of Σ are positive. If $\Sigma = \text{diag}\,(\sigma_1, \ldots, \sigma_n)$, then $\Sigma^{-1} = \text{diag}\left(\sigma_1^{-1}, \ldots, \sigma_n^{-1}\right)$. Show that the matrix $Q\Sigma^{-1}Q^T$ is the inverse of M.

3. Since positive semidefinite matrices may have eigenvalues that are zero, we see by the previous problem that not all psd matrices are invertible. Show that if M is a psd matrix and I is the identity matrix, then $M + \lambda I$ is symmetric positive definite for any $\lambda > 0$, and give an expression for the inverse of $M + \lambda I$.

4. Let M and N be symmetric matrices, with M positive semidefinite and N positive definite. Use the definitions of psd and spd to show that $M + N$ is symmetric positive definite. Thus $M + N$ is invertible. (Hint: For any $x \neq 0$, show that $x^T(M + N)x > 0$. Also note that $x^T(M + N)x = x^T Mx + x^T Nx$.)

4 Kernel Matrices

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 $\|x\| = \sqrt{\langle x, x\rangle} = \sqrt{x^T x}$.]

5 Kernel Ridge Regression

In lecture, we discussed how to kernelize ridge regression using the representer theorem. He we pursue a bare-hands approach.

Suppose our input space is $\mathcal{X}$=$\mathbf{R}^d$ and our output space is $\mathcal{Y} = \mathbf{R}$. Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set from $\mathcal{X} \times \mathcal{Y}$. We'll use the "design matrix" $X \in \mathbf{R}^{n \times d}$, which has the input vectors as rows:

$$X = \begin{pmatrix} -x_1- \\ \vdots \\ -x_n- \end{pmatrix}.$$

Recall the ridge regression objective function:

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

for $\lambda > 0$.

1. Show that for w to be a minimizer of $J(w)$, we must have $X^TXw + \lambda Iw = X^Ty$. Show that the minimizer of $J(w)$ is $w = (X^TX + \lambda I)^{-1}X^Ty$. Justify that the matrix $X^TX + \lambda I$ is invertible, for $\lambda > 0$. (The last part should follow easily from the earlier exercises on psd and spd matrices.)

2. Rewrite $X^TXw + \lambda Iw = X^Ty$ as $w = \frac{1}{\lambda}(X^Ty - X^TXw)$. Based on this, show that we can write $w = X^T\alpha$ for some α, and give an expression for α.

3. Based on the fact that $w = X^T\alpha$, explain why we say w is "in the span of the data."

4. Show that $\alpha = (\lambda I + XX^T)^{-1}y$. Note that XX^T is the kernel matrix for the standard vector dot product. (Hint: Replace w by $X^T\alpha$ in the expression for α, and then solve for α.)

5. Give a kernelized expression for the Xw, the predicted values on the training points. (Hint: Replace w by $X^T\alpha$ and α by its expression in terms of the kernel matrix XX^T.

6. Give an expression for the prediction $f(x) = x^Tw^*$ 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

$$k_x = \begin{pmatrix} x^Tx_1 \\ \vdots \\ x^Tx_n \end{pmatrix}$$

to simplify the expression.]

6 Decision Trees

6.1 Building Trees by Hand[1]

In this problem we're going to be build a a small decision tree by hand for predicting whether or not a mushroom is poisonous. The training dataset is given below:

Poisonous	Size	Spots	Color
N	5	N	White
N	2	Y	White
N	2	N	Brown
N	3	Y	Brown
N	4	N	White
N	1	N	Brown
Y	5	Y	White
Y	4	Y	Brown
Y	4	Y	Brown
Y	1	Y	White
Y	1	Y	Brown

[1]Based on Homework #4 from David Sontag's DS-GA 1003, Spring 2014.

We're going to build a binary classification tree using the Gini index as the node impurity measure. The feature "Size" should be treated as numeric (i.e. we should find real-valued split points). For a given split, let R_1 and R_2 be the sets of data indices in each of the two regions of the split. Let $\hat{p}_1$ be the proportion of poisonous mushrooms in R_1, and let $\hat{p}_2$ be the proportion in R_2. Let N_1 and N_2 be the total number of training points in R_1 and R_2, respectively. Then the Gini index for the first region is $Q_1 = 2\hat{p}_1(1-\hat{p}_1)$ and $Q_2 = 2\hat{p}_2(1-\hat{p}_2)$ for the second region. When choosing our splitting variable and split point, we're looking to minimize the weighted impurity measure:

$$N_1 Q_1 + N_2 Q_2.$$

1. What is the first split for a binary classification tree on this data, using the Gini index? Work this out "by hand", and show your calculations. [Hint: This should only require calculating 6 weighted impurity measures.]

2. The first split partitions the data into two parts. Make another split so that the space is partitioned into 3 regions. Determine the predicted "probability of poisonous" for each of those regions.

3. Suppose we build a binary tree on the dataset given below using the Gini criterion and we build it so deep that all terminal nodes are either pure or cannot be split further. (To think about: How could we have a node that is not pure, but cannot be split further?) What would the training error be, given as a percentage? Why? [Hint: You can do this by inspection, without any significant calculations.]

Y	A	B	C
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	0
0	0	1	1
1	0	1	1
0	1	0	0
1	1	0	1
1	1	1	0
0	1	1	1
1	1	1	1

6.2 Investigating Impurity Measures[2]

1. Consider a data set with 400 data points from class C_1 and 400 data points from class C_2 . Suppose that a tree model A splits these into $(300, 100)$ at the first leaf node and $(100, 300)$ at the second leaf node, where (n, m) denotes that n points are assigned to C_1 and m points are assigned to C_2 . Similarly, suppose that a second tree model B splits them into $(200, 400)$ and $(200, 0)$. Show that the misclassification rates for the two trees are equal, but that the cross-entropy and Gini impurity measures are both lower for tree B than for tree A.

[2]From Bishop's *Pattern Recognition and Machine Learning*, Problem 14.11

7 Representer Theorem [Optional]

Recall the following theorem from lecture:

Theorem (Representer Theorem)**.** *Let*

$$J(w) = R\left(\|w\|\right) + L\left(\langle w, \psi(x_1)\rangle, \ldots, \langle w, \psi(x_n)\rangle\right),$$

where $R : \mathbf{R}^{\geq 0} \to \mathbf{R}$ *is nondecreasing (the* ***regularization*** *term) and* $L : \mathbf{R}^n \to \mathbf{R}$ *is arbitrary (the* ***loss*** *term). If* $J(w)$ *has a minimizer, then it has a minimizer of the form*

$$w^* = \sum_{i=1}^{n} \alpha_i \psi(x_i).$$

Furthermore, if R *is strictly increasing, then all minimizers have this form.*

Note: There is nothing in this theorem that guarantees $J(w)$ has a minimizer at all. If there is no minimizer, then this theorem does not tell us anything.

In the lecture slides we proved the first part of the theorem. Now we will prove the part beginning with "Furthermore."

1. Let M be a closed subspace of a Hilbert space $\mathcal{H}$. F or any $x \in \mathcal{H}$, let $m_0 = \text{Proj}_M x$ be the projection of x onto M. By the Projection Theorem, we know that $x - m_0 \perp M$. Then by the Pythagorean Theorem, we know $\|x\|^2 = \|m_0\|^2 + \|x - m_0\|^2$. From this we concluded in lecture that $\|m_0\| \leq \|x\|$. Show that we have $\|m_0\| = \|x\|$ only when $m_0 = x$. (Hint: Use the postive-definiteness of the inner product: $\langle x, x\rangle \geq 0$ and $\langle x, x\rangle = 0 \iff x = 0$, and the fact that we're using the norm derived from such an inner product.)

2. Continue the proof of the Representer Theorem from the lecture slides to show that if R is strictly increasing, then all minimizers have this form. (Hint: Consider separately the cases that $\|w\| < \|w^*\|$ and the case $\|w\| = \|w^*\|$.)

3. Suppose that R and L are both convex in their arguments. Use properties of convex functions to show that L is a convex function of w, and then that J is also convex function of w. For simplicity, you may assume that our feature space is $\mathbf{R}^d$, rather than a generic Hilbert space. You may also use the fact that the composition of a convex function and an affine function is convex. That is:, suppose $f : \mathbf{R}^n \to \mathbf{R}$, $A \in \mathbf{R}^{n \times m}$ and $b \in \mathbf{R}^n$. Define $g : \mathbf{R}^m \to \mathbf{R}$ by $g(x) = f\left(Ax + b\right)$. Then if f is convex, then so is g. From this exercise, we can conclude that if L is convex, then J does have a minimizer of the form $w^* = \sum_{i=1}^{n} \alpha_i \psi(x_i)$, and if R is also strictly increasing, then all minimizers of J have this form.

8 Ivanov and Tikhonov Regularization [Optional]

In lecture there was a claim that the Ivanov and Tikhonov forms of ridge and lasso regression are equivalent. We will now prove a more general result.

8.1 Tikhonov optimal implies Ivanov optimal

Let $\phi : \mathcal{F} \to \mathbf{R}$ be any performance measure of $f \in \mathcal{F}$, and let $\Omega : \mathcal{F} \to \mathbf{R}$ be any complexity measure. For example, for ridge regression over the linear hypothesis space $\mathcal{F} = \left\{ f_w(x) = w^T x \mid w \in \mathbf{R}^d \right\}$, we would have $\phi(f_w) = \frac{1}{n} \sum_{i=1}^n \left(w^T x_i - y_i \right)^2$ and $\Omega(f_w) = w^T w$.

1. Suppose that for some $\lambda > 0$ we have the Tikhonov regularization solution

$$f_* = \arg\min_{f \in \mathcal{F}} \left[\phi(f) + \lambda \Omega(f) \right]. \tag{1}$$

   Show that f_* is also an Ivanov solution. That is, $\exists r > 0$ such that

$$f_* = \arg\min_{f \in \mathcal{F}} \phi(f) \text{ subject to } \Omega(f) \leq r. \tag{2}$$

   (Hint: Start by figuring out what r should be. If you're stuck on this, ask for help. Then one approach is proof by contradiction: suppose f^* is not the optimum in (2) and show that contradicts the fact that f^* solves (1).)

8.2 Ivanov optimal implies Tikhonov optimal

For the converse, we will restrict our hypothesis space to a parametric set. That is,

$$\mathcal{F} = \left\{ f_w(x) : \mathcal{X} \to \mathbf{R} \mid w \in \mathbf{R}^d \right\}.$$

So we will now write ϕ and Ω as functions of $w \in \mathbf{R}^d$.

Let w^* be a solution to the following Ivanov optimization problem:

$$\begin{array}{ll} \text{minimize} & \phi(w) \\ \text{subject to} & \Omega(w) \leq r. \end{array}$$

Assume that strong duality holds for this optimization problem and that the dual solution is attained. Then we will show that there exists a $\lambda \geq 0$ such that $w_* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda \Omega(w) \right]$.

1. Write the Lagrangian $L(w, \lambda)$ for the Ivanov optimization problem.

2. Write the dual optimization problem in terms of the dual objective function $g(\lambda)$, and give an expression for $g(\lambda)$. [Writing $g(\lambda)$ as an optimization problem is expected - don't try to solve it.]

3. We assumed that the dual solution is attained, so let $\lambda^* = \arg\max_{\lambda \geq 0} g(\lambda)$. We also assumed strong duality, which implies $\phi(w^*) = g(\lambda^*)$. Show that the minimum in the expression for $g(\lambda^*)$ is attained at w^*. [Hint: You can use the same approach we used when we derived that strong duality implies complementary slackness [3].] **Conclude the proof** by showing that for the choice of $\lambda = \lambda^*$, we have $w_* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda \Omega(w) \right]$.

[3] See `https://davidrosenberg.github.io/ml2015/docs/3b.convex-optimization.pdf` slide 24.

4. [Optional] The conclusion of the previous problem allows $\lambda = 0$, which means we're not actually regularizing at all. To ensure we get a proper Ivanov regularization problem, we need an additional assumption. The one below is taken from [**?**]:

$$\inf_{w \in \mathbf{R}^d} \phi(w) < \inf_{\substack{w \in \mathbf{R}^d \\ \Omega(w) \leq r}} \phi(w)$$

Note that this is a rather intuitive condition: it is simply saying that we can fit the training data better [strictly better] if we don't use any regularization. With this additional condition, show that $w_* = \arg\min_{w \in \mathbf{R}^d} [\phi(w) + \lambda \Omega(w)]$ for some $\lambda > 0$.

8.3 Ivanov implies Tikhonov for Ridge Regression.

To show that Ivanov implies Tikhonov for the ridge regression problem (square loss with ℓ_2 regularization), we need to demonstrate strong duality and that the dual optimum is attained. Both of these things are implied by Slater's constraint qualifications.

1. Show that the Ivanov form of ridge regression is a convex optimization problem with a strictly feasible point.

9 Novelty Detection [Optional]

(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 $\phi : \mathcal{X} \to \mathcal{H}$ be our feature map, mapping elements of the input space into our "feature space" $\mathcal{H}$, which is a Hilbert space (and thus has an inner product). Formulate the novelty detection algorithm described above as an optimization problem.
2. Give the Lagrangian for this problem, and write an equivalent, unconstrained "inf sup" version of the optimization problem.
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.]
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.]
5. Write an expression for the optimal sphere in terms of the solution to the dual problem.
6. Write down the complementary slackness conditions for this problem, and characterize the points that are the "support vectors".
7. Briefly explain how you would apply this algorithm in practice to detect "novel" instances.
8. [More 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).