MLCS HW4

2.1 Let $P = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$

$PP^T = I$, so P is Orthogonal

2.2 Since M is psd

$$x^T M x \geq 0, \text{ for all } x \in R^n$$

Using spectral theorem, M can be diagonalized

$$M = Q \Sigma Q^T$$

$$\Rightarrow \Sigma = Q^T M Q$$

Since Q is the matrix of eigenvectors of M, The eigenvalue of psd matrix M is

$$\text{diag}(Q^T M Q) = (q_1^T M q_1, q_2^T M q_1 \ldots, q_n^T M q_n)$$

So, $\forall q_1, q_2, q_3 \ldots q_n \in R_n$, the eigenvalues of M are non-negative.

2.3 $M = BB^T$, $\forall x \in R^n$,

$$x^T M x = x^T B B^T x$$

$$= (B^T x)^T B^T x \geq 0$$

so M is psd

Otherwise, if M is psd, we use spectral theorem

$$M = Q \Sigma Q^T$$

Since, all eigenvalues are non-negative.

let $\Sigma^{1/2} = \text{diag}(\sigma_1^{1/2}, \ldots \sigma_d^{1/2})$

$$\therefore M = Q(\Sigma^{1/2})^T \Sigma^{1/2} Q^T$$

$$= Q(\Sigma^{1/2} Q^T)^T \Sigma^{1/2} Q^T$$

$$= BB^T \text{ (since } M \text{ is symmetric)}$$

3.1 Using Spectral Theolem,

$\Sigma = Q^T M Q$, where Q is the matrix of M's eigenvectors

$$Q^T M Q = \begin{pmatrix} q_1^T M q_1 & q_1^T M q_2 & \cdots & q_1^T M q_n \\ \vdots & & & \vdots \\ q_n^T M q_1 & \cdots & \cdots & q_n^T M q_n \end{pmatrix}$$

$$= \begin{pmatrix} q_1^T M q_1 & 0 & \cdots & 0 \\ 0 & q_2^T M q_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & q_n^T M q_n \end{pmatrix}$$

So, if M is positive definite for all $i = 1, \ldots n$

$q_i^T M q_i > 0$ $\forall$ i from 1 to n

3.2

Let $M' = Q \Sigma^{-1} Q^T$

$$M'M = Q \Sigma Q^T Q \Sigma^{-1} Q^T$$

$$= Q \Sigma \Sigma^{-1} Q^T$$

$$= Q Q^T$$

$$= I$$

$\therefore$, $M = Q \Sigma^{-1} Q^T$ is the inverse of M

3.3

For all non-zero $x \in R^n$,

$$x^T(M + \lambda I)x = x^T M x + \lambda x^T x$$

Since M is psd,

$$x^T M x \geq 0 \quad \& \quad x^x > 0$$

$$x^T(M + \lambda I)x > 0$$

Therefore, $M + \lambda I$ is symmetric positive definite.

$$M + \lambda I = Q(\Sigma + \lambda I)Q^T$$

So $Q(\Sigma + \lambda I)^{-1}Q^T$ will be the inverse of $M + \lambda I$

3.4 For $x \in R^n$

$$x^T(M+N)x = x^TMx + x^TNx$$

If M is psd & N is positive definite

$$x^TMx \geq 0 \quad \& \quad x^TNx > 0$$

$$\Rightarrow x^T(M+N)x > 0$$

$\therefore$, $M+N$ is symmetric positive definite

4.1

$$K = XX^T$$

$$= \begin{pmatrix} x_1^T x_1 & x_1^T x_2 & \cdots & x_1^T x_m \\ \vdots & & & \vdots \\ x_m^T x_1 & x_m^T x_2 & \cdots & x_m^T x_m \end{pmatrix}$$

For a general x_i, x_j the distance can be expressed as

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

$$= \sqrt{x_i^T x_i + x_j^T x_j - 2 x_i^T x_j}$$

$$= \sqrt{K_{ii} + K_{jj} + (-2 K_{ij})}$$

where K_{ii}, K_{jj} & K_{ij} are elements of Gram Matrix.

5.1 $J(w) = \|Xw - y\|^2 + \lambda \|w\|^2$

$\partial J(w) = 2X^TXw - 2X^Ty + \lambda I w$

To minimize $J(w)$, $\partial J(w) = 0$

$\Rightarrow 2X^TXw - 2X^Ty + \lambda I w = 0$

$\Rightarrow X^TXw + \lambda I w = X^Ty$

$\Rightarrow w = (X^TX + \lambda I)^{-1} X^Ty$

Let $\alpha \in R^n$, $\alpha \neq 0$

$\alpha^T (X^TX + \lambda I) \alpha$

$= \alpha^T X^TX \alpha + \lambda \alpha^T \alpha$

Since, $\alpha^T X^TX \alpha \geq 0$ & $\lambda \alpha^T \alpha > 0$

$\Rightarrow X^TX + \lambda I$ is invertible for $\lambda > 0$

5.2

$$w = \frac{1}{\lambda}\left(X^T y - X^T X w\right)$$

$$\Rightarrow w = X^T \left[\frac{1}{\lambda}(y - Xw)\right]$$

$$= X^T \alpha$$

where $\alpha = \frac{1}{\lambda}(y - Xw)$

5.3

Since, $w = X^T \alpha$ it means that w is the linear combinations of the columns of X and so is in the span of range of X.

5.4

$$\alpha = \frac{1}{\lambda}(y - X\omega)$$

$$\lambda I \alpha = y - XX^T\alpha$$

$$\Rightarrow \alpha = (\lambda I + XX^T)^{-1} y$$

5.5

$$Xw = XX^T\alpha$$

$$\text{Since } \alpha = (\lambda I + XX^T)^{-1} y$$

$$Xw = XX^T(\lambda I + XX^T)^{-1} y$$

$$= K(\lambda I + K)^{-1} y$$

where, $K = XX^T$

5.6

$$w^* = X^T(\lambda I + XX^T)^{-1}y$$

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

$$f(x) = x^T w^*$$

$$= x^T X^T (\lambda I + XX^T)^{-1} y$$

$$= k_x^T (\lambda I + XX^T)^{-1} y$$

7.1 $y_j \langle w^{(t)}, x_j \rangle$

$= y_j \langle x_j, X^T \alpha^{(t)} \rangle \quad \{ w^{(t)} = X \alpha^{(t)} \}$

$= y_j (x_j)^T (X^T \alpha^{(t)})$

$= y_j K_j \alpha^{(t)}$

7.2

$$w^{(t)} = \sum_{i=1}^{n} \alpha_i^{(t)} x_i$$

Since, (x_j, y_i) doesn't have a malgin violation

$$w^{(t+1)} = (1 - \eta^{(t)} \lambda) w^{(t)}$$

$$= (1 - \eta^{(t)} \lambda) \sum_{i=1}^{n} \alpha_i^{(t)} \cdot x_i$$

$$\alpha^{(t+1)} = (1 - \eta_t \lambda) \alpha^{(t)}$$

7.3

In case of malgin violation,

$$w^{(t+1)} = (1 - \eta^{(t)}\lambda) w^{(t)} + \eta_t y_j x_j$$

$$= (1 - \eta^t \lambda) \sum_{i=1}^{n} \alpha_i^{(t)} x_i + \eta_t y_j x_j$$

$$\Rightarrow \alpha^{(t+1)} = (1 - \eta^t \lambda) \alpha_i^{(t)}, \quad i \neq j$$

$$(1 - \eta^t \lambda) \alpha_i^{(t)} + \eta_t y_j, \quad i = j$$

$\alpha^{(1)} = (0, \ldots 0) \in R^{\alpha}$

$t = 0$

repeat

$t = t + 1$

$\eta^t = 1/t\lambda$

randomly choose j in $1 \ldots n$

if $y_j K_j \alpha < 1$

$\alpha^{t+1} = (1 - \eta^t \lambda) \alpha^{(t)}$

$\alpha^{t+1}_{[j]} = \alpha^{(t+1)}[j] + \eta_t y_j$

else

$\alpha^{t+1} = (1 - \eta^{(t)} \lambda) \alpha^{(t)}$

solve $x^T \alpha^t$

6 [Optional] Pegasos and SSGD for ℓ_2-regularized ERM[1]

Consider the objective function

$$J(w) = \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \ell_i(w),$$

where $\ell_i(w)$ represents the loss on the ith training point (x_i, y_i). Suppose $\ell_i(w) : \mathbf{R}^d \to \mathbf{R}$ is a convex function. Let's write

$$J_i(w) = \frac{\lambda}{2}\|w\|_2^2 + \ell_i(w),$$

for the one-point approximation to $J(w)$ using the ith training point. $J_i(w)$ is probably a very poor approximation of $J(w)$. However, if we choose i uniformly at random from $1, \ldots, n$, then we do have $\mathbb{E}J_i(w) = J(w)$. We'll now show that subgradients of $J_i(w)$ are unbiased estimators of some subgradient of $J(w)$, which is our justification for using SSGD methods.

In the problems below, you may use the following facts about subdifferentials without proof (as in Homework #3): 1) If $f_1, \ldots, f_m : \mathbf{R}^d \to \mathbf{R}$ are convex functions and $f = f_1 + \cdots + f_m$, then $\partial f(x) = \partial f_1(x) + \cdots + \partial f_m(x)$ **[additivity]**. 2) For $\alpha \geq 0$, $\partial(\alpha f)(x) = \alpha \partial f(x)$ **[positive homogeneity]**.

1. [Optional] For each $i = 1, \ldots, n$, let $g_i(w)$ be a subgradient of $J_i(w)$ at $w \in \mathbf{R}^d$. Let $v_i(w)$ be a subgradient of $\ell_i(w)$ at w. Give an expression for $g_i(w)$ in terms of w and $v_i(w)$

2. [Optional] Show that $\mathbb{E}g_i(w) \in \partial J(w)$, where the expectation is over the randomly selected $i \in 1, \ldots, n$. (In words, the expectation of our subgradient of a randomly chosen $J_i(w)$ is in the subdifferential of J.)

3. [Optional] Now suppose we are carrying out SSGD with the Pegasos step-size $\eta^{(t)} = 1/(\lambda t)$, $t = 1, 2, \ldots$. In the t'th step, suppose we select the ith point and thus take the step $w^{(t+1)} = w^{(t)} - \eta^{(t)} g_i(w^{(t)})$. Let's write $v^{(t)} = v_i(w^{(t)})$, which is the subgradient of the loss part of $J_i(w^{(t)})$ that is used in step t. Show that

$$w^{(t+1)} = -\frac{1}{\lambda t}\sum_{\tau=1}^{t} v^{(\tau)}$$

[Hint: One approach is proof by induction. First show it's true for $w^{(2)}$. Then assume it's true for $w^{(t)}$ and prove it's true for $w^{(t+1)}$. This will prove that it's true for all $t = 2, 3, \ldots$ by induction.

[1] This problem is based on Shalev-Shwartz and Ben-David's book Understanding Machine Learning: From Theory to Algorithms, Sections 14.5.3, 15.5, and 16.3).

4. [Optional] We can use the previous result to get a nice equivent formulation of Pegasos. Let $\theta^{(t)} = \sum_{\tau=1}^{t-1} v^{(t)}$. Then $w^{(t+1)} = -\frac{1}{\lambda t}\theta^{(t+1)}$ Then Pegasos from Homework #3 is equivalent to Algorithm 1. Similar to the $w = sW$ decomposition from homework #3, this decomposition

Algorithm 1: Pegasos Algorithm Reformulation

input: Training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbf{R}^d \times \{-1, 1\}$ and $\lambda > 0$.
$\theta^{(1)} = (0, \ldots, 0) \in \mathbf{R}^d$
$w^{(1)} = (0, \ldots, 0) \in \mathbf{R}^d$
$t = 1$ # step number
repeat
  randomly choose j in $1, \ldots, n$
  if $y_j \left\langle w^{(t)}, x_j \right\rangle < 1$
    $\theta^{(t+1)} = \theta^{(t)} + y_j x_j$
  else
    $\theta^{(t+1)} = \theta^{(t)}$
  endif
  $w^{(t+1)} = -\frac{1}{\lambda t}\theta^{(t+1)}$ # need not be explicitly computed
  $t = t + 1$
until bored
return $w^{(t)} = -\frac{1}{\lambda(t-1)}\theta^{(t)}$

gives the opportunity for significant speedup. Explain how Algorithm 1 can be implemented so that, if x_j has s nonzero entries, then we only need to do $O(s)$ accesses in every pass through the loop.

7 Kernelized Pegasos

Recall the SVM objective function

$$\min_{w \in \mathbf{R}^n} \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max\left\{0, 1 - y_i w^T x_i\right\}$$

and the Pegasos algorithm on the training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbf{R}^d \times \{-1, 1\}$ (Algorithm 2).

Note that in every step of Pegasos, we rescale $w^{(t)}$ by $\left(1 - \eta^{(t)}\lambda\right) = \left(1 - \frac{1}{t}\right) \in (0, 1)$. This "shrinks" the entries of $w^{(t)}$ towards 0, and it's due to the regularization term $\frac{\lambda}{2}\|w\|_2^2$ in the SVM objective function. Also note that if the example in a particular step, say (x_j, y_j), is not classified with the required margin (i.e. if we don't have margin $y_j w_t^T x_j \geq 1$), then we also add a multiple of x_j to $w^{(t)}$ to end up with $w^{(t+1)}$. This part of the adjustment comes from the empirical risk. Since

Algorithm 2: Pegasos Algorithm

input: Training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbf{R}^d \times \{-1, 1\}$ and $\lambda > 0$.
$w^{(1)} = (0, \ldots, 0) \in \mathbf{R}^d$
$t = 0$ # step number
repeat
  $t = t + 1$
  $\eta^{(t)} = 1/(t\lambda)$ # step multiplier
  randomly choose j in $1, \ldots, n$
  if $y_j \left\langle w^{(t)}, x_j \right\rangle < 1$
    $w^{(t+1)} = (1 - \eta^{(t)}\lambda) w^{(t)} + \eta_t y_j x_j$
  else
    $w^{(t+1)} = (1 - \eta^{(t)}\lambda) w^{(t)}$
until bored
return $w^{(t)}$

we initialize with $w^{(1)} = 0$, we are guaranteed that we can always write[2]

$$w^{(t)} = \sum_{i=1}^{n} \alpha_i^{(t)} x_i$$

after any number of steps t. When we kernelize Pegasos, we'll be tracking $\alpha^{(t)} = (\alpha_1^{(t)}, \ldots, \alpha_n^{(t)})^T$ directly, rather than w.

1. Kernelize the expression for the margin. That is, show that $y_j \left\langle w^{(t)}, x_j \right\rangle = y_j K_{j\cdot} \alpha^{(t)}$, where $k(x_i, x_j) = \langle x_i, x_j \rangle$ and $K_{j\cdot}$ denotes the jth row of the kernel matrix K corresponding to kernel k.

2. Suppose that $w^{(t)} = \sum_{i=1}^{n} \alpha_i^{(t)} x_i$ and for the next step we have selected a point (x_j, y_j) that does not have a margin violation. Give an update expression for $\alpha^{(t+1)}$ so that $w^{(t+1)} = \sum_{i=1}^{n} \alpha_i^{(t+1)} x_i$.

3. Repeat the previous problem, but for the case that (x_j, y_j) has a margin violation. Then give the full pseudocode for kernelized Pegasos. You may assume that you receive the kernel

[2]Note: This resembles the conclusion of the representer theorem, but it's saying something different. Here, we are saying that the $w^{(t)}$ after every step of the Pegasos algorithm lives in the span of the data. The representer theorem says that a mathematical minimimizer of the SVM objective function (i.e. what the Pegasos algorithm would converge to after infinitely many steps) lies in the span of the data. If, for example, we had chosen an initial $w^{(1)}$ that is NOT in the span of the data, then none of the $w^{(t)}$'s from Pegasos would be in the span of the data. However, we know Pegasos converges to a minimum of the SVM objective. Thus after a very large number of steps, $w^{(t)}$ would be very close to being in the span of the data. It's the gradient of the regularization term that pulls us back towards the span of the data. This is basically because the regularization is driving all components towards 0, while the empirical risk updates are only pushing things away from 0 in directions in the span of the data.

matrix K as input, along with the labels $y_1, \ldots, y_n \in \{-1, 1\}$

4. [Optional] Above we've directly kernelized . While the direct implementation of the original Pegasos required updating all entries of w in every step, a direct kernelization of Algorithm 2, as we have done above, leads to updating all entries of α in every step. Give a version of the kernelized Pegasos algorithm that does not suffer from this inefficiency. You may try splitting the scale and direction similar to the approach of the previous problem set, or you may use a decomposition based on Algorithm 1 from the optional problem 6 above.

8 Kernel Methods: Let's Implement

In this section you will get the opportunity to code kernel ridge regression and, optionally, kernelized SVM. To speed things along, we've written a great deal of support code for you, which you can find in the Jupyter notebooks in the homework zip file.

8.1 One more review of kernelization can't hurt (no problems)

Consider the following optimization problem on a data set $(x_1, y_1), \ldots (x_n, y_n) \in \mathbf{R}^d \times \mathcal{Y}$:

$$\min_{w \in \mathbf{R}^d} R\left(\sqrt{\langle w, w \rangle}\right) + L\left(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle\right),$$

where $w, x_1, \ldots, x_n \in \mathbf{R}^d$, and $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbf{R}^d$. The function $R : \mathbf{R}^{\geq 0} \to \mathbf{R}$ is nondecreasing and gives us our regularization term, while $L : \mathbf{R}^n \to \mathbf{R}$ is arbitrary[3] and gives us our loss term. We noted in lecture that this general form includes soft-margin SVM and ridge regression, though not lasso regression. Using the representer theorem, we showed if the optimization problem has a solution, there is always a solution of the form $w = \sum_{i=1}^n \alpha_i x_i$, for some $\alpha \in \mathbf{R}^n$. Plugging this into the our original problem, we get the following "kernelized" optimization problem:

$$\min_{\alpha \in \mathbf{R}^n} R\left(\sqrt{\alpha^T K \alpha}\right) + L\left(K\alpha\right),$$

where $K \in \mathbf{R}^{n \times n}$ is the Gram matrix (or "kernel matrix") defined by $K_{ij} = k(x_i, x_j) = \langle x_i, x_j \rangle$. Predictions are given by

$$f(x) = \sum_{i=1}^n \alpha_i k(x_i, x),$$

and we can recover the original $w \in \mathbf{R}^d$ by $w = \sum_{i=1}^n \alpha_i x_i$.

The "**kernel trick**" is to swap out occurrences of the kernel k (and the corresponding Gram matrix K) with another kernel. For example, we could replace $k(x_i, x_j) = \langle x_i, x_j \rangle$ by $k'(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle$ for an arbitrary feature mapping $\psi : \mathbf{R}^d \to \mathbf{R}^D$. In this case, the recovered $w \in \mathbf{R}^D$ would be $w = \sum_{i=1}^n \alpha_i \psi(x_i)$ and predictions would be $\langle w, \psi(x_i) \rangle$.

More interestingly, we can replace k by another kernel $k''(x_i, x_j)$ for which we do not even know or cannot explicitly write down a corresponding feature map ψ. Our main example of this is the RBF kernel

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right),$$

for which the corresponding feature map ψ is infinite dimensional. In this case, we cannot recover w since it would be infinite dimensional. Predictions must be done using $\alpha \in \mathbf{R}^n$, with $f(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$.

Your implementation of kernelized methods below should not make any reference to w or to a feature map ψ. Your "learning" routine should return α, rather than w, and your prediction function should also use α rather than w. This will allow us to work with kernels that correspond to infinite-dimensional feature vectors.

[3] You may be wondering "Where are the y_i's?". They're built into the function L. For example, a square loss on a training set of size 3 could be represented as $L(p_1, p_2, p_3) = \frac{1}{3}\left[(p_1 - y_1)^2 + (p_2 - y_2)^2 + (p_3 - y_3)^3\right]$, where each p_i stands for the ith prediction $\langle w, x_i \rangle$.

8.2 Kernels and Kernel Machines

There are many different families of kernels. So far we've spoken about linear kernels, RBF/Gaussian kernels, and polynomial kernels. The last two kernel types have parameters. In this section we'll implement these kernels in a way that will be convenient for implementing our kernelized ML methods later on. For simplicity, and because it is by far the most common situation[4], we will assume that our input space is $\mathcal{X} = \mathbf{R}^d$. This allows us to represent a collection of n inputs in a matrix $X \in \mathbf{R}^{n \times d}$, as usual.

1. Write functions that compute the RBF kernel $k_{\text{RBF}(\sigma)}(x, x') = \exp\left(-\|x - x'\|^2 / \left(2\sigma^2\right)\right)$ and the polynomial kernel $k_{\text{poly}(a,d)}(x, x') = (a + \langle x, x' \rangle)^d$. The linear kernel $k_{\text{linear}}(x, x') = \langle x, x' \rangle$, has been done for you in the support code. Your functions should take as input two matrices $W \in \mathbf{R}^{n_1 \times d}$ and $X \in \mathbf{R}^{n_2 \times d}$ and should return a matrix $M \in \mathbf{R}^{n_1 \times n_2}$ where $M_{ij} = k(W_{i\cdot}, X_{j\cdot})$. In words, the (i, j)'th entry of M should be kernel evaluation between w_i (the ith row of W) and x_j (the jth row of X). The matrix M could be called the "cross-kernel" matrix, by analogy to the cross-covariance matrix. For the RBF kernel, you may use the scipy function `cdist(X1,X2,'sqeuclidean')` in the package `scipy.spatial.distance` or (with some more work) write it in terms the linear kernel (Bauckhage's article on calculating Euclidean distance matrices may be helpful).

```
def RBF_kernel(W,X,sigma = 1):

    return(np.exp((-1) * distance.cdist(W, X, 'sqeuclidean')/(2*(sigma**2))))


def polynomial_kernel(W, X, offset, degree):

    return (np.dot(W,np.transpose(X)) + offset)**degree
```

[4] We are noting this because one interesting aspect of kernel methods is that they can act directly on an arbitrary input space $\mathcal{X}$ (e.g. text files, music files, etc.), so long as you can define a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$. But we'll not consider that case here.

2. Use the linear kernel function defined in the code to compute the kernel matrix on the set of points $x_0 \in \mathcal{D}_X = \{-4, -1, 0, 2\}$. Include both the code and the output.

```
x = np.matrix([-4,-1,0,2])
print(linear_kernel(np.transpose(x),np.transpose(x)))
```

$$\begin{pmatrix} 16 & 4 & 0 & -8 \\ 4 & 1 & 0 & 4 \\ 0 & 0 & 0 & 0 \\ -8 & -2 & 0 & 4 \end{pmatrix}.$$

3. Suppose we have the data set $\mathcal{D} = \{(-4, 2), (-1, 0), (0, 3), (2, 5)\}$. Then by the representer theorem, the final prediction function will be in the span of the functions $x \mapsto k(x_0, x)$ for $x_0 \in \mathcal{D}_X = \{-4, -1, 0, 2\}$. This set of functions will look quite different depending on the kernel function we use.

   (a) Plot the set of functions $x \mapsto k_{\text{linear}}(x_0, x)$ for $x_0 \in \mathcal{D}_X$ and for $x \in [-6, 6]$.

   (b) Plot the set of functions $x \mapsto k_{\text{poly}(1,3)}(x_0, x)$ for $x_0 \in \mathcal{D}_X$ and for $x \in [-6, 6]$.

   (c) Plot the set of functions $x \mapsto k_{\text{RBF}(1)}(x_0, x)$ for $x_0 \in \mathcal{D}_X$ and for $x \in [-6, 6]$.

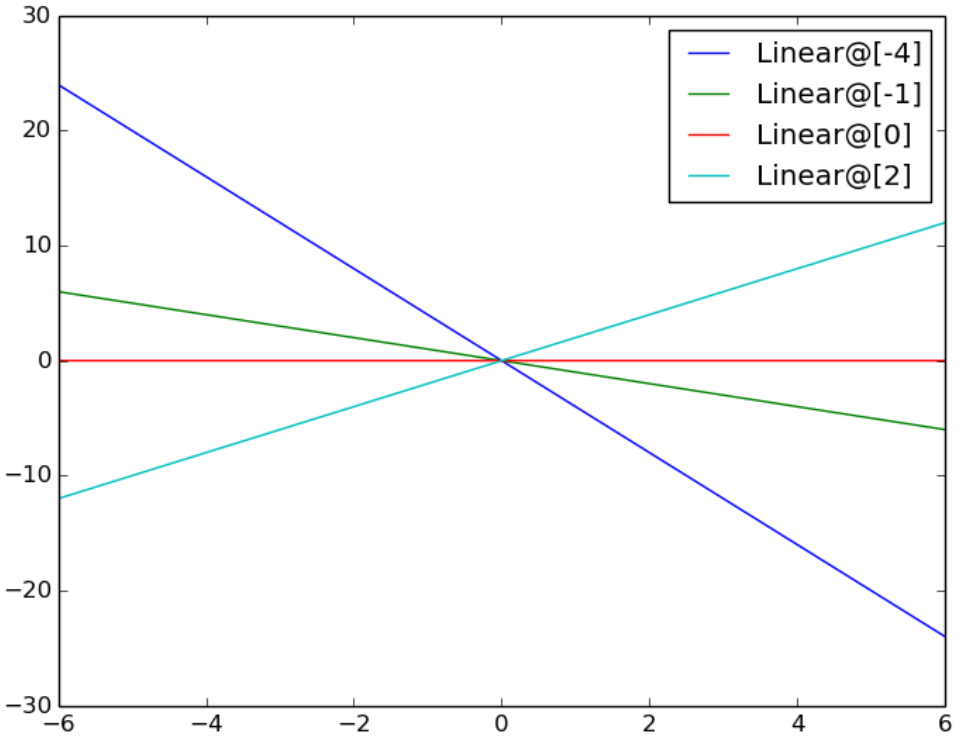

Figure 1: Linear set of functions

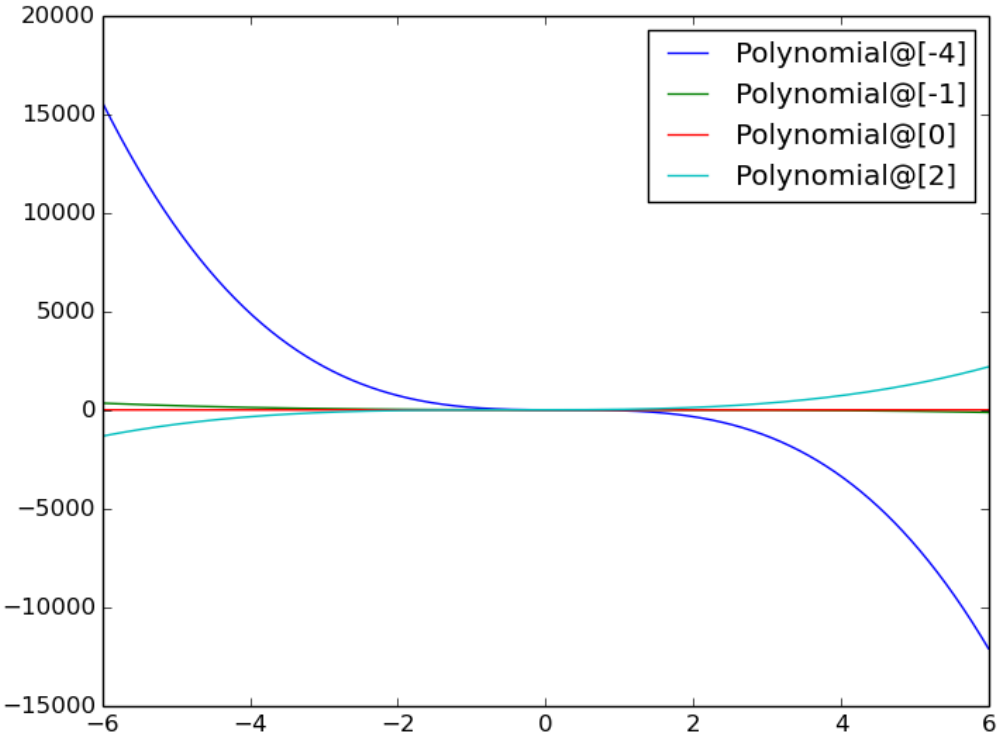

Figure 2: Polynomial set of functions

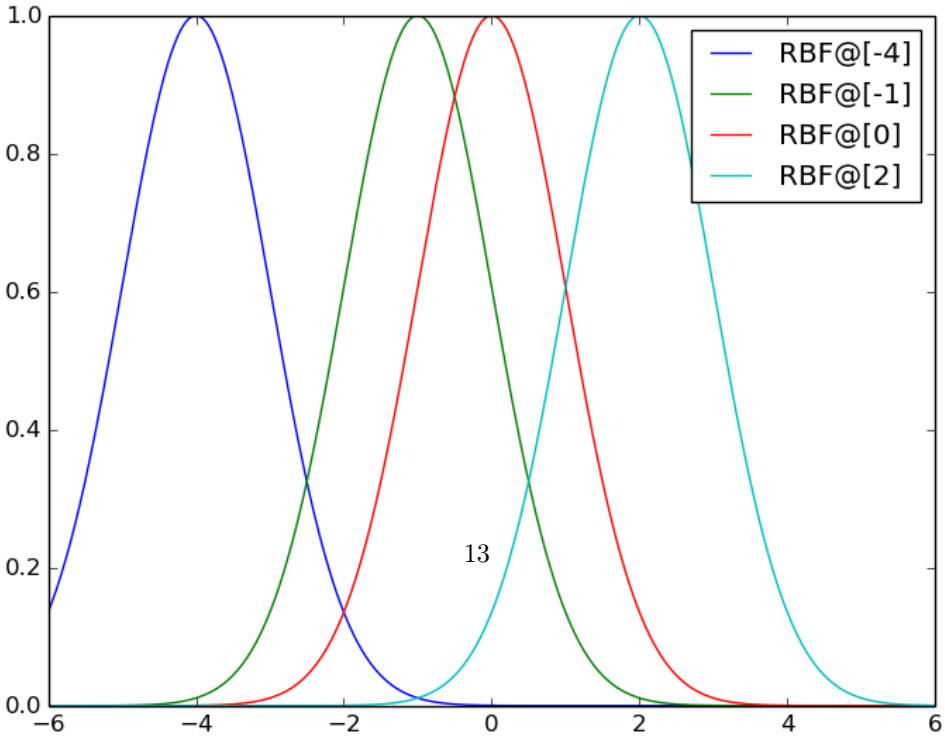

Figure 3: RBF set of functions

4. By the representer theorem, the final prediction function will be of the form $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$, where $x_1, \ldots, x_n \in \mathcal{X}$ are the inputs in the training set. This is a special case of what is sometimes called a **kernel machine**, which is a function of the form $f(x) = \sum_{i=1}^{r} \alpha_i k(\mu_i, x)$, where $\mu_1, \ldots, \mu_r \in \mathcal{X}$ are called **prototypes** or **centroids** (Murphy's book Section 14.3.1.). In the special case that the kernel is an RBF kernel, we get what's called an **RBF Network** (proposed by Broomhead and Lowe in 1988). We can see that the prediction functions we get from our kernel methods will be kernel machines in which every input point $x_1, \ldots, x_n$ serves as a prototype point. Complete the `predict` function of the class `Kernel_Machine` in the skeleton code. Construct a `Kernel_Machine` object with the RBF kernel (sigma=1), with prototype points at $-1, 0, 1$ and corresponding weights $1, -1, 1$. Plot the resulting function.

   Note: For this problem, and for other problems below, it may be helpful to use partial application on your kernel functions. For example, if your polynomial kernel function has signature `polynomial_kernel(W, X, offset, degree)`, you can write `k = functools.partial(polynomial_kernel, offset=2, degree=2)`, and then a call to `k(W,X)` is equivalent to `polynomial_kernel(W, X, offset=2, degree=2)`, the advantage being that the extra parameter settings are built into `k(W,X)`. This can be convenient so that you can have a function that just takes a kernel function `k(W,X)` and doesn't have to worry about the parameter settings for the kernel.

```
class Kernel_Machine(object):
    def __init__(self, kernel, prototype_points, weights):

        self.kernel = kernel
        self.prototype_points = prototype_points
        self.weights = weights

    def predict(self, X):

        return np.dot(np.transpose(self.kernel(self.prototype_points, X)),
        self.weights)


plot_step = .01
xpts = np.arange(-6.0, 6.0, plot_step).reshape(-1,1)
prototypes = np.array([-1,0,1]).reshape(-1,1)
weights = np.array([1,-1,1]).reshape(-1,1)

k = Kernel_Machine(RBF_kernel, prototypes, weights)
y = k.predict(xpts)
plt.plot(xpts, y)
plt.xlabel(Xpts")
plt.ylabel("Predicted value")
plt.show()
```

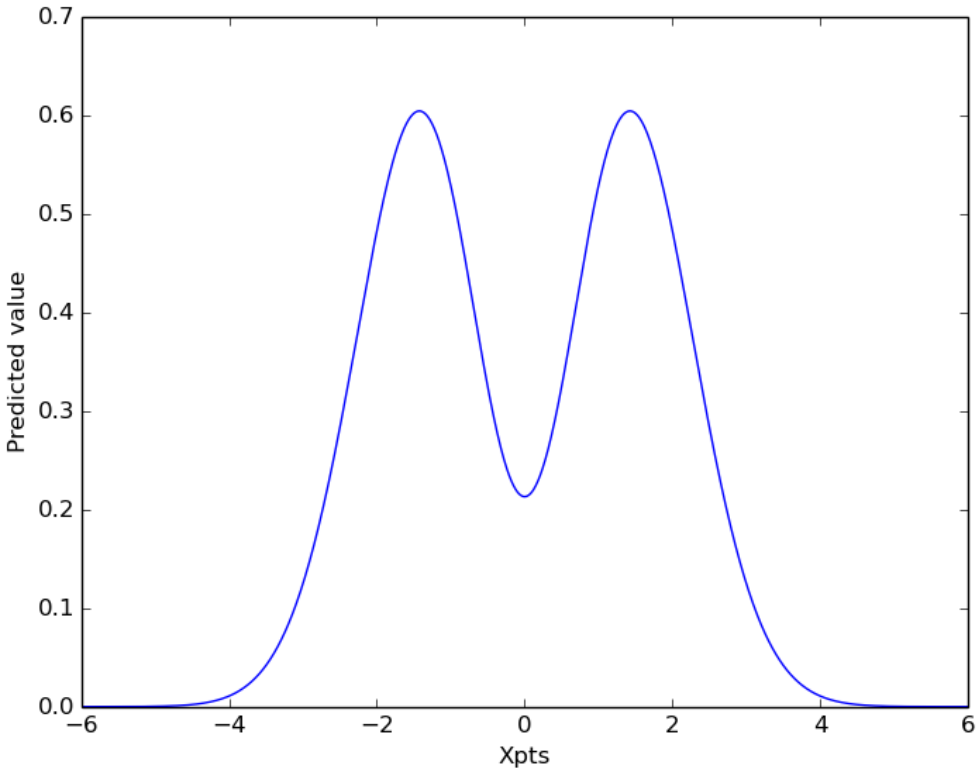

Figure 4: RBF Prediction Function

8.3 Kernel Ridge Regression

In the zip file for this assignment, you'll find a training and test set, along with some skeleton code. We're considering a one-dimensional regression problem, in which $\mathcal{X} = \mathcal{Y} = \mathcal{A} = \mathbf{R}$. We'll fit this data using kernelized ridge regression, and we'll compare the results using several different kernel functions. Because the input space is one-dimensional, we can easily visualize the results.

1. Plot the training data. You should note that while there is a clear relationship between x and y, the relationship is not linear.

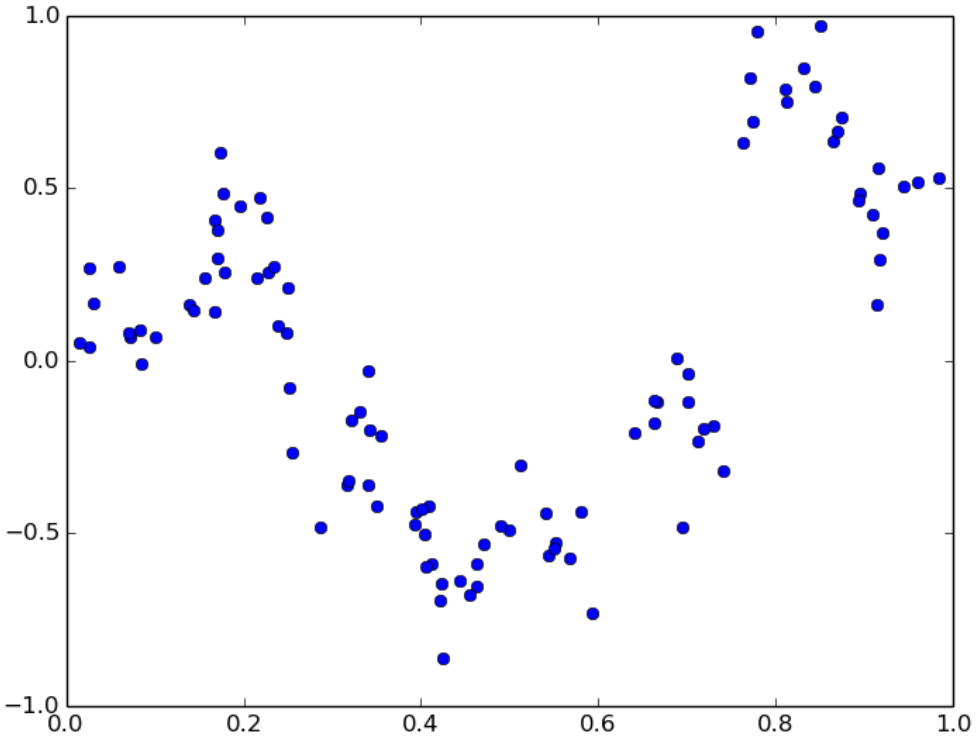

Figure 5: Plot of the test data

2. In a previous problem, we showed that in kernelized ridge regression, the final prediction function is $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$, where $\alpha = (\lambda I + K)^{-1} y$ and $K \in \mathbf{R}^{n \times n}$ is the Gram matrix of the training data: $K_{ij} = k(x_i, x_j)$, for $x_1, \ldots, x_n$. In terms of kernel machines, α_i is the weight on the kernel function evaluated at the prototype point x_i. Complete the function `train_kernel_ridge_regression` so that it performs kernel ridge regression and returns a `Kernel_Machine` object that can be used for predicting on new points.

```
def train_kernel_ridge_regression(X, y, kernel, l2reg):

    mat_krnl = kernel(X,X)
    alpha = np.dot(np.linalg.inv(l2reg * np.identity(X.shape[0]) + mat_krnl), y)

    return Kernel_Machine(kernel, X, alpha)
```

3. Use the code provided to plot your fits to the training data for the RBF kernel with a fixed regularization parameter of 0.0001 for 3 different values of sigma: 0.01, 0.1, and 1.0. With values of sigma do you think would be more likely to over fit, and which less?

   Solution:

   There is overfitting for sigma = 0.01 and underfitting for sigma = 1

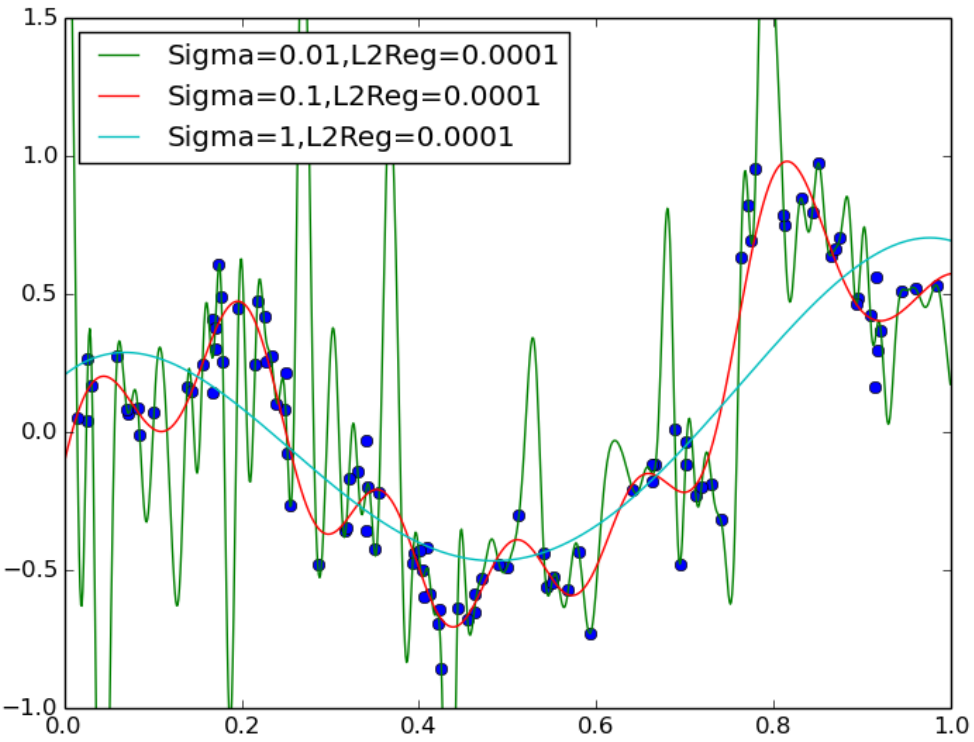

Figure 6: Prediction using different value of sigma

4. Use the code provided to plot your fits to the training data for the RBF kernel with a fixed sigma of 0.02 and 4 different values of the regularization parameter λ: 0.0001, 0.01, 0.1, and 2.0. What happens to the prediction function as $\lambda \to \infty$?

   Solution:

   The value of the prediction function approaches 0 as the regularization increases. If $\lambda = \infty$, then the value would be always 0.

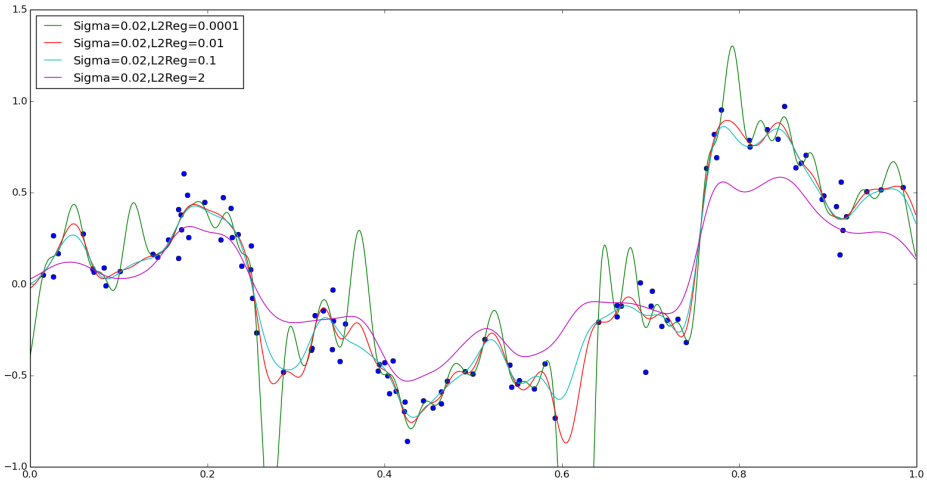

Figure 7: Prediction using different value of lambda

5. Find the best hyperparameter settings (including kernel parameters and the regularization parameter) for each of the kernel types. Summarize your results in a table, which gives training error and test error for each setting. Include in your table the best settings for each kernel type, as well as nearby settings that show that making small change in any one of the hyperparameters in either direction will cause the performance to get worse. You should use average square loss on the test set to rank the parameter settings. To make things easier for you, we have provided an sklearn wrapper for the kernel ridge regression function we have created so that you can us sklearn's GridSearchCV. Note: Because of the small dataset size, these models can be fit extremely fast, so there is no excuse for not doing extensive hyperparameter tuning.

   For RBF, the optimal values are: sigma = 0.07, $\lambda = 0.064$

Sigma	Lambda	Train Loss	Test Loss
0.07	0.064	0.0146	0.0138
0.07	0.054	0.0131	0.0147
0.06	0.064	0.0173	0.0155
0.07	0.074	0.0142	0.0149
0.08	0.064	0.0168	0.0169

   For Polynomial, the optimal values are : degree = 30, offset = -3, $\lambda = 44.3$

Degree	Offset	Lambda	Train Loss	Test Loss
30	-3	44.3	0.022	0.015
35	-3	44.3	0.024	0.023
25	-3	44.3	0.027	0.032
30	-2	44.3	0.025	0.019
30	-4	44.3	0.021	0.028
30	-3	50.3	0.026	0.021
30	-3	38.3	0.025	0.032

   For linear, the change is very subtle, optimal value for $\lambda = 3.9$

Lambda	Train Loss	Test Loss
3.9	0.20655	0.01654095
3.8	0.20656	0.01654096
4.0	0.20656	0.01654097

6. Plot your best fitting prediction functions using the polynomial kernel and the RBF kernel. Use the domain $x \in (-0.5, 1.5)$. Comment on the results.

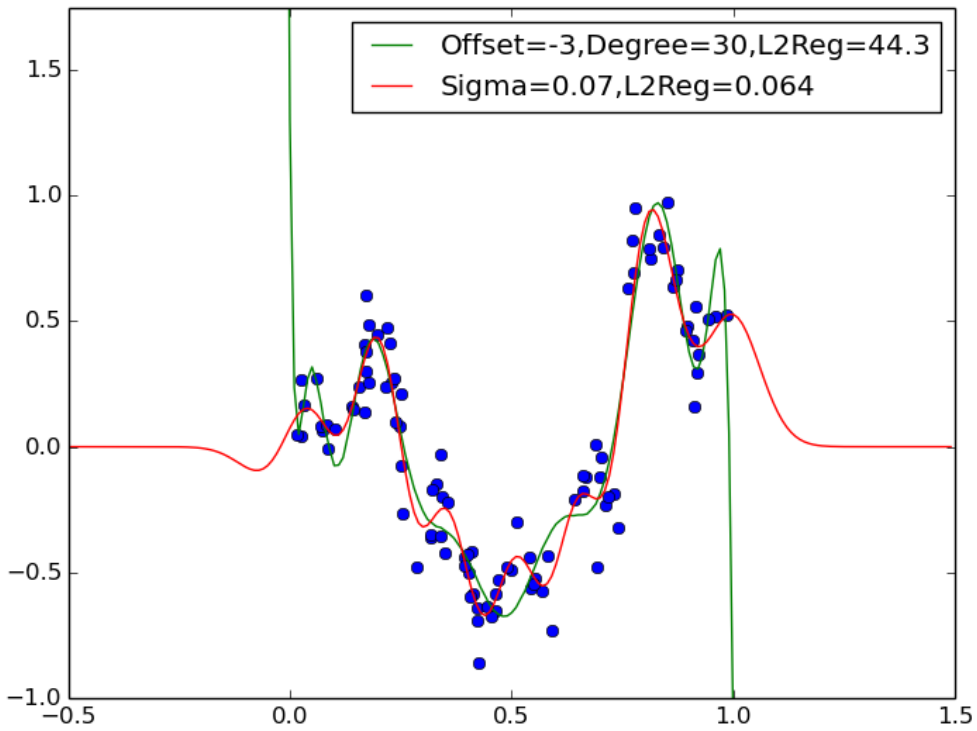

Figure 8: Plot using the best parameters obtained for RBF and Polynomial

7. The data for this problem was generated as follows: A function $f : \mathbf{R} \to \mathbf{R}$ was chosen. Then to generate a point (x, y), we sampled x uniformly from $(0, 1)$ and we sampled $\varepsilon \sim \mathcal{N}(0, 0.1)$ (so $\mathrm{Var}(\varepsilon) = 0.1$). The final point is $(x, f(x) + \varepsilon)$. What is the Bayes decision function and the Bayes risk for the loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

   Solution:

   Bayes function is given by $\mathbf{E}(Y|X) = \mathbf{E}((f(x) + \varepsilon)|X)$

   $= \mathbf{E}(f(x)|X) + \mathbf{E}(\varepsilon|X)$

   Therefore, Bayes Decision Function $f^*(x) = f(x)$

   Bayes risk is given by $\mathrm{Var}(Y|X) = \mathrm{Var}(\varepsilon) = 0.1$

8. [Optional] Attempt to improve performance by using different kernel functions. Chapter 4 from Rasmussen and Williams' book *Gaussian Processes for Machine Learning* describes many kernel functions, though they are called

covariance functions in that book (but they have exactly the same definition). Note that you may also create a kernel function by first explicitly creating feature vectors, if you are so inspire.

9. [Optional] Use any machine learning model you like to get the best performance you can.

8.4 [Optional] Kernelized Support Vector Machines with Kernelized Pegasos

1. Load the SVM training and test data from the zip file. Plot the training data using the code supplied. Is the data linearly separable? Quadratically seperable? What if we used some RBF kernel?

2. Unlike for kernel ridge regression, there is no closed-form solution for SVM classification (kernelized or not). Implement kernelized Pegasos. Because we are not using a sparse representation for this data, you will probably not see much gain by implementing the "optimized" versions described in the problems above.

3. Find the best hyperparameter settings (including kernel parameters and the regularization parameter) for each of the kernel types. Summarize your results in a table, which gives training error and test error (i.e. average 0/1 loss) for each setting. Include in your table the best settings for each kernel type, as well as nearby settings that show that making small change in any one of the hyperparameters in either direction will cause the performance to get worse. You should use the 0/1 loss on the test set to rank the parameter settings.

4. Plot your best fitting prediction functions using the linear, polynomial, and the RBF kernel. The code provided may help.

9 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 this problem, we will prove the part of the Representer theorem for the case that R is strictly increasing.

1. Let M be a closed subspace of a Hilbert space $\mathcal{H}$. For 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. Give the proof of the Representer Theorem in the case that R is strictly increasing. That is, show that if R is strictly increasing, then all minimizers have this form claimed. (Hint: Consider separately the cases that $\|w\| < \|w^*\|$ and the case $\|w\| = \|w^*\|$.)

3. Suppose that $R : \mathbf{R}^{\geq 0} \to \mathbf{R}$ and $L : \mathbf{R}^n \to \mathbf{R}$ are both convex functions. Use properties of convex functions to show that $w \mapsto L\left(\langle w, \psi(x_1)\rangle, \ldots, \langle w, \psi(x_n)\rangle\right)$ is a convex function of w, and then that $J(w)$ 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 and R are 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.

10 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.

10.1 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).)

10.2 Ivanov optimal implies Tikhonov optimal (when we have Strong Duality)

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:

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

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[5].] **Conclude the proof** by showing that for the choice of $\lambda = \lambda^*$, we have $w_* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda\Omega(w)\right]$.

4. 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 $w_* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda\Omega(w)\right]$ for some $\lambda > 0$.

10.3 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.

[5] See https://davidrosenberg.github.io/mlcourse/Archive/2016/Lectures/3b.convex-optimization.pdf#page=30 slide 30.

11 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).