Homework 6

The questions below are due on Sunday October 22, 2017; 11:00:00 PM.

You are not logged in.

If you are a current student, please Log In (https://introml.mit.edu/fall17/homework/hw06?loginaction=login) for full access to this page.

1) MAKING A PREDICTION

We are given the following data set `X, Y`:

```
X = np.array([[1, 2, -1, 0],
              [3, -2, 1, 4]])
Y = np.array([[-1, 1, 1, 1]])
```

For the kernel $k(x, z) = \sum_{i=1}^{d}(x_i + 1)(z_i + 1)$ and $\alpha = [1, 2, 0, 0]$, what is the predicted output for a new $x = (1, -3)$? Refer to Eq. (11) in the notes.

Enter a label: 1

Ask for Help

Write a Python function to make these predictions: `predict(X, Y, alpha, x, k)` where

- `X` is a dxn data matrix
- `Y` is a 1xn matrix of labels
- `alpha` is a 1xn array of coefficients
- `x` is a dx1 column vector
- `k` is a kernel function

It should return a label, 1 or -1 (you may need to convert it to a number from a 1x1 array by using `float`).

Please look at the documentation for `np.apply_along_axis`, this should come in handy. Here's a neat trick for turning a one dimensional array `v` into a column vector: `v.reshape(-1,1)` -- the -1 means "unknown dimension", which in this case is however many elements there are in `v`, so `np.array([1,2]).reshape(-1,1)` `=> np.array([[1],[2]])`.

Also write the `k1` kernel function corresponding to the kernel defined above; this will be passed as the kernel function to `predict`. `k1` should take two column vectors and return a number (you may need to convert it to a number from a 1x1 array by using `float`).

```
def predict(X, Y, alpha, x, k):
    kernel_x = np.apply_along_axis(k, 0, X, x) + 1 #1Xn array
    result = np.sum(np.multiply(np.multiply(Y, alpha), kernel_x))
    return np.sign(result)

k1 = lambda x,z: np.dot((x+1).T, z+1):
```

Ask for Help

2) COLONEL VECTOR

Consider the kernel

$$K(x, z) = x \cdot z + 4\,(x \cdot z)^2$$

where the vectors x and z are 2-dimensional. This kernel is equal to an inner product $\phi(x) \cdot \phi(z)$ for some definition of ϕ. What is the function ϕ?

Assume that $\phi(x)$ has the following form :

$$[c_0, c_1 x_1, c_2 x_2, c_3 x_1^2, c_4 x_1 x_2, c_5 x_2^2, c_6 x_1 x_2^2, c_7 x_1^2 x_2, c_8 x_1^2 x_2^2]$$

Provide the coefficients c_i:

Enter a list of 9 numbers or Python numerical expressions (involving `+,-,*,/,**`) corresponding to the c_i:

```
[0, 1, ,1, 2, 8**0.5, 2, 0, 0, 0]
```

Ask for Help

3) CLASSIFICATION WITH 3 POINTS

Consider a simple classification problem (of the kind that you could only encounter in an exam). The training data consist of only three labeled points in one dimension.

$$(x^{(1)} = -1, y^{(1)} = +1), (x^{(2)} = 0, y^{(2)} = -1), (x^{(3)} = +1, y^{(3)} = +1)$$

which we will try to separate with a linear classifier through origin in the feature space by looking for a separator of the form $\theta \cdot \phi(x)$.

1) We decided to solve the problem using kernel $K(x, z) = 1 + |xz|$ where $|\cdot|$ is the absolute value. Check each of the feature mappings $\phi(x)$ that correspond to this kernel?

$\phi(x) =$

☐	$(1, xz)^T$
☐	$(1, \|xz\|)^T$
☐	$(1, x)^T$
☑	$(1, \|x\|)^T$
☐	$(x, 1)^T$
☑	$(\|x\|, 1)^T$

Ask for Help

2) What are the transformed points $\phi(x^{(i)})$?

Enter a list of tuples, each of the form ((a,b), c), with a,b the feature values and c the label. Enter them in the same order as the points given above:

`[((1, 1), 1), ((1, 0), -1), ((1, 1), 1)]`

Ask for Help

3) Using this kernel (feature mapping), are the three training examples linearly separable through origin in the feature space? If so, specify a separator (in terms of values for $\alpha_1, \alpha_2, \alpha_3$) that correctly classifies them, else enter `'None'`.

Enter `None` or a list of three numbers $\alpha_1, \alpha_2, \alpha_3$:

`[2, 3, 0]`

Ask for Help

4) KERNELS OF WISDOM

4.1) Mystery Kernel

1) Let $x, x' \in \mathbb{R}$ and consider the kernel function $K(x, x') = e^{x+x'}$. Is this a valid kernel? Justify your answer using rules 1 -- 4 from section 7.1.3 in the notes.

a)

Indicate below which rules you need to use:

☑	Combination rule 1
☑	Combination rule 2
☐	Combination rule 3
☑	Combination rule 4

Ask for Help

If you were using rule (2) to construct the $e^{x+x'}$ kernel, what function f and more elementary kernel K would you use to compose it. You can use arithmetic operations, constants, the symbols `x` and `z` and `e**(w)` works for exponentials. If you don't need them, enter 0.

b)

The function $f(x) =$ `e**(x)`

Ask for Help

c)

The more elementary kernel $K(x, x') =$ `1`

Ask for Help

2) What is the feature mapping associated with that kernel? There is only one feature, enter an expression for it.

You can use arithmetic operations, constants, the symbols `x` and `z` and `e**(w)` works for exponentials.

$\phi(x) =$ e**(x)

Ask for Help

4.2) Polynomial Kernel

1) Let $x, z \in \mathbb{R}^2$ be two feature vectors, and let $K(x, z) = (x^T z + 1)^2$. This is often known as a polynomial kernel (of order 2). It's simple to compute: you just take the dot product between two feature vectors, add one, and then square the result. But what kind of feature mapping does this kernel implicitly use?

Assuming we can write $K(x, z) = \phi(x)^T \phi(z)$ give an expression for $\phi(x)$ assuming that x are two-dimensional feature vectors Note that the order of the polynomial and the dimension of the feature vectors are unrelated. We are illustrating the second-order polynomial kernel on two-dimensional vectors, but we could have demonstrated it on one-dimensional vectors or three-dimensional vectors or whatever (as we'll see later).

Assuming the feature mapping $\phi(x)$ has the following form:

$$[c_0, c_1 x_1, c_2 x_2, c_3 x_1^2, c_4 x_1 x_2, c_5 x_2^2, c_6 x_1 x_2^2, c_7 x_1^2 x_2, c_8 x_1^2 x_2^2]$$

Provide the coefficients c_i:

Enter a list of 9 numbers or Python numerical expressions (involving `+`,`-`,`*`,`/`,`**`) corresponding to the c_i:

[1, 2**0.5, 2**0.5, 1, 2**0.5, 1, 0, 0, 0]

Ask for Help

2) As a simple example that uses this kernel, imagine that our input feature vectors were bag of words vectors with a possible vocabulary of 10000 words.

How long would the feature vectors be? Hint: $\binom{d+k}{k}$ for k=order, d=words

501501

Ask for Help

3) How many addition and multiplication operations are required (without doing anything fancy) to compute $K(x, z)$ when the vectors are 10000 long?

Operations= about 1e15

Ask for Help

4) Write an expression for the entry in the above feature vector that is non-zero if and only if two different words j and l appear simultanously in a document?

You can use the symbols `x_s` for any subscript `s`, constants, numerical operators and `sqrt(a)` for any expression a.

Feature is ??

Ask for Help

5) What order polynomial kernel would allow us to find documents simultaneously containing the words: `6.036 best class evar`?

Order is 4??

Ask for Help

6) How long would the feature vectors be for the kernel from the previous question?

Length= about 100

Ask for Help

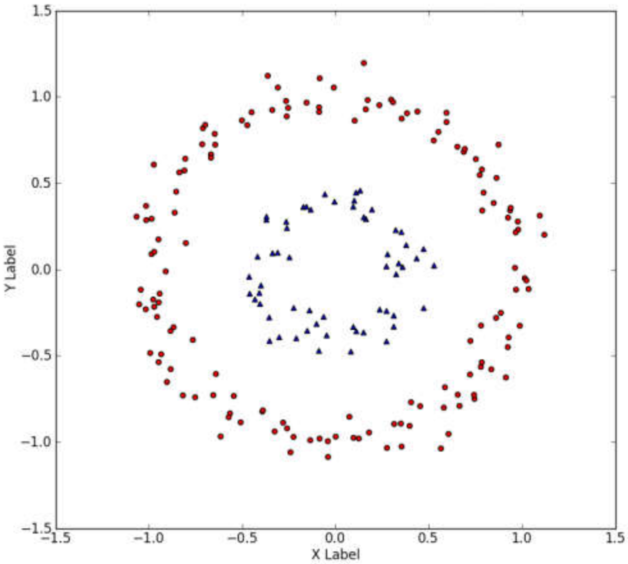
7) How many addition and multiplication operations are required (without doing anything fancy) to compute the corresponding kernel function when the vectors are 10000 long?

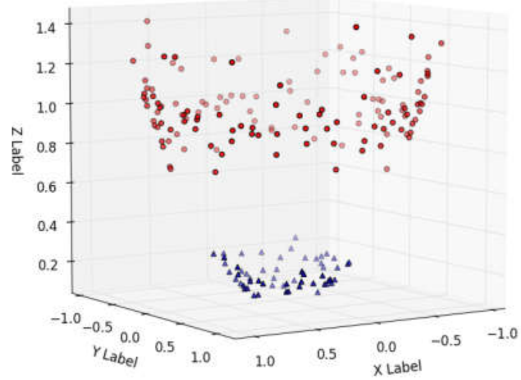
Operations= about 1e15

Ask for Help

5) POP OUT!

In the figure below, a set of points in 2-D is shown on the left. On the right, the same points are shown mapped to a 3-D space via some transform $\phi(x)$, where x denotes a point in the 2-D space. Notice that $\phi(x)_1 = x_1$ and $\phi(x)_2 = x_2$, or in other words, the first and second coordinates are unchanged by the transformation.





1) Which of the following functions could have been used to compute the value of the 3rd coordinate,

$\phi(x)_3$ for each point:

○	$x_1 + x_2$
○	$x_1^2 + x_2^2$
◉	$x_1 x_2$
○	$x_1^2 - x_2^2$

Ask for Help

2) How would a linear decision boundary in the 3 dimensional space of the form ($\{\phi \in \mathbb{R}^3 : \phi_3 + \theta_0 = 0\}$) appear in the original 2 dimensional space?

circle

Ask for Help

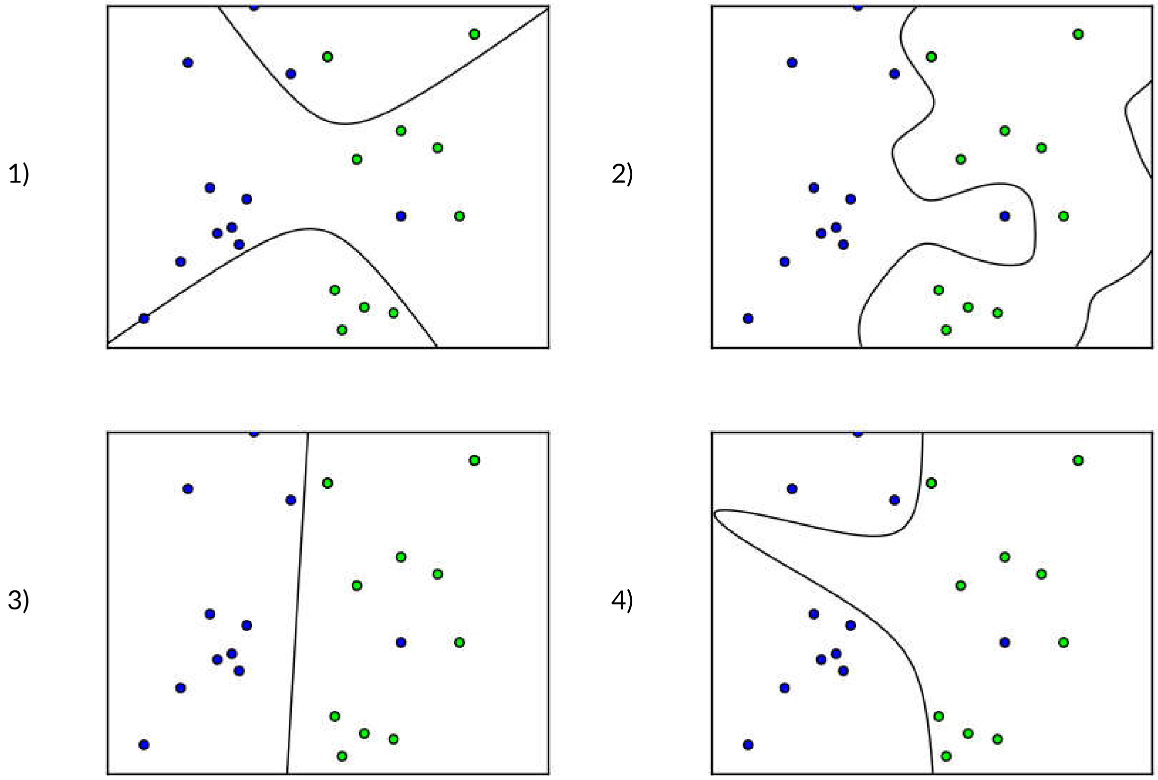
6) SVM LOSS FUNCTION, NOW WITH KERNELS

Consider fitting a kernelized SVM to a dataset $(x^{(i)}, y^{(i)})$ with $\forall i\ x^{(i)}, y^{(i)} \in \mathbb{R}$. To fit the parameters of this model, one computes θ and θ_0 to minimize the following objective:

$$L(x;\theta,\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_h(y^{(i)}(\theta \cdot \phi(x^{(i)}))) + \frac{\lambda}{2}||\theta||^2$$

where ϕ is the feature vector associated with the kernel function. Note that, in a kernel method, the optimization problem for training would be typically expressed solely in terms of the kernel function $K(x, x')$ rather than using the associated feature vectors $\phi(x)$ (primal). We use the primal only to highlight the classification problem solved.

The plots below show 4 different kernelized SVM models estimated from the same 11 data points. We used a different kernel to obtain each plot but got confused about which plot corresponds to which kernel. Help us out by assigning each plot (a-d) to one of the following models (i-iv).



Linear kernel: 3

Ask for Help

Quadratic kernel: 1

Ask for Help

3rd-order kernel: 4

Ask for Help

Gaussian radial basis kernel: 2

Ask for Help