COMS4771: Homework 2 Solution

October 7, 2015

Problem 1

(A)

Answer: The VC dimension of perceptron in R^d is d+1.

Proof. It will be proved in two steps:
(1) There exist d+1 points that perceptron can shatter.
(2) No $d+2$(or more) points can be shattered by H.

(1)

Suppose the perceptron $f(x)$:

$$f(x) = \begin{cases} 1 & \text{if } w^T x + b > 0, \\ -1 & \text{otherwise.} \end{cases}$$

Consider d+1 points $x^{(0)} = (0, ..., 0)^T, x^{(1)} = (1, 0, ..., 0)^T, x^{(2)} = (0, 1, ..., 0)^T, ..., x^{(d)} = (0, 0, ..., 1)^T$. After these d+1 points being arbitrarily labeled: $y = (y_0, y_1, ..., y_d)^T \in \{-1, 1\}^{d+1}$.

Let $b = 0.5 \cdot y_0$ and $w = (w_1, w_2, ...w_d)$ where $w_i = y_i, i \in \{1, 2, ..., d\}$. Thus $f(x)$ can label all these $d + 1$ points correctly.

So the VC dimension of perceptron is at least $d + 1$.

(2)

Expand $x \in R^d$ to $X \in R^{d+1}$, by letting $(X)^T = (x^T, 1)$, and let $W^T = (w^T, b)$.

Thus:

$$f(X) = \begin{cases} 1 & \text{if } W^T X > 0, \\ -1 & \text{otherwise.} \end{cases}$$

Assume there exist d+2 points that perceptron in R^d can shatter, namely $x^{(1)}, x^{(2)}, ..., x^{(d+2)} \in R^d$ corresponding to $X^{(1)}, X^{(2)}, ..., X^{(d+2)} \in R^{d+1}$.

Since $d + 2$ points in R^{d+1}, there exists certain i s.t.:

$$X^{(i)} = \sum_{j \neq i} a_j \cdot X^{(j)},$$

where at least one $a_j \neq 0$. Let $S = \{j | j \neq i, a_j \neq 0\}$.

$\forall j \in S$, we give $x^{(j)}$ the label $sign(a_j)$, and give $x(i)$ a label -1.

By our assumption, there exists W that make $f(X)$ label those $d+2$ points correctly.

So $\forall j \in S$, we have

$$a_j \cdot W^T X^{(j)} > 0,$$

and

$$W^T X^{(i)} \leq 0.$$

Also:

$$\begin{aligned} W^T X^{(i)} &= W^T(\sum_{j \neq i} a_j \cdot X^{(j)}) \\ &= W^T(\sum_{j \in S} a_j \cdot X^{(j)}) \\ &= \sum_{j \in S} a_j \cdot W^T X^{(j)} \\ &> 0. \end{aligned}$$

So, our assumption is false. The VC dimension of perceptron in R^d is at most d+1.

Combine (1) and (2), we can conclude that the VC dimension of perceptron in R^d is $d + 1$. □

(B)

Proof. Suppose a hypothesis space H whose VC-dimension $VC(H) = n$, so there exist n points that H can shatter. We can arbitrarily give 0 or 1 label to each of the points, so there are 2^n ways to label them. No matter how the n points are labeled, there exist a hypothesis $h \in H$ which can label them correctly. H must consists at least 2^n different hypothesises, that is $|H| \geq 2^n$. So $VC(H) = n \leq \log_2 |H|$. □

(C)

Answer: The VC dimension of C is 2.

Proof. First, $VCdim(C) \leq \log_2 |C| = \log_2 10$, so $VCdim(C) \leq 3$.

Assume there exist $X_1, X_2, X_3 \in X = \{1, 2, ..., 999\}$ that can be shattered by C, which means all kinds of labels can be realized by concept class C. So $\exists\{c_{i_1}, c_{i_2}, ..., c_{i_8}\} \subset C$, which can label X_1, X_2, X_3 as it's showed in Table 1.

	X_1	X_2	X_3
c_{i_1}	0	0	0
c_{i_2}	0	0	1
c_{i_3}	0	1	0
c_{i_4}	0	1	1
c_{i_5}	1	0	0
c_{i_6}	1	0	1
c_{i_7}	1	1	0
c_{i_8}	1	1	1

Table 1

As it can be seen from Table 1, there are at least 3 concepts $\in C$ that label X_1 to be 1, so X_1 must contains at least 4 different digits. However, $X_1 \leq 999$, thus it can contain at most 3 digits.The assumption is proved to be false.

So:

$$VCdim(C) \leq 2.$$

Consider the following Table 2.

	34	24
c_1	0	0
c_2	0	1
c_3	1	0
c_4	1	1

Table 2

As it show in Table 2, 34 and 24 can be shattered by concept class C. So:

$$VCdim(C) \geq 2.$$

Finally, we can conclude that the VC dimension of C is 2. □

Problem 2

(A)

Proof.

$$K_{ij} = k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

$$\begin{aligned} c^T K c &= \sum_i \sum_j c_i c_j K_{ij} \\ &= \sum_i \sum_j c_i c_j \phi(x_i) \cdot \phi(x_j) \\ &= (\sum_i c_i \phi(x_i)) \cdot (\sum_i c_i \phi(x_i)) \\ &= \|\sum_i c_i \phi(x_i)\|_2^2 \\ &\geq 0 \end{aligned}$$

□

a)

Proof.

$$\begin{aligned} k(x, y) &= \alpha k_1(x, y) + \beta k_2(x, y) \\ &= \langle \sqrt{\alpha}\phi_1(x), \sqrt{\alpha}\phi_1(y) \rangle + \langle \sqrt{\beta}\phi_2(x), \sqrt{\beta}\phi_2(y) \rangle \\ &= \langle [\sqrt{\alpha}\phi_1(x), \sqrt{\beta}\phi_2(x)], [\sqrt{\alpha}\phi_1(y), \sqrt{\beta}\phi_2(y)] \rangle \end{aligned}$$

□

b)

Proof. Let $f_i(x)$ be the ith feature value under the feature map ϕ_1, $g_i(x)$ be the ith feature value under the feature map ϕ_2.

Then:

$$\begin{aligned} k(x,y) &= k_1(x,y)k_2(x,y) \\ &= (\phi_1(x)\cdot\phi_1(y))(\phi_2(x)\cdot\phi_2(y)) \\ &= (\sum_{i=0}^{\infty} f_i(x)f_i(y))(\sum_{j=0}^{\infty} g_j(x)g_j(y)) \\ &= \sum_{i,j} f_i(x)f_i(y)g_j(x)g_j(y) \\ &= \sum_{i,j}(f_i(x)g_j(x))(f_i(y)g_j(y)) \\ &= \langle\phi_3(x),\phi_3(y)\rangle \end{aligned}$$

where ϕ_3 has feature $h_{i,j}(x) = f_i(x)g_j(x)$. □

c)
Since each polynomial term is a product of kernels with a positive coefficient, the proof follows from part *a* and *b*.

d)
By Taylor expansion:

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

Then the proof follows part c.

(B)

Proof. We wish to show that the kernel $k(\mathbf{x},\mathbf{y}) = \exp(-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2)$ can be written as an inner-product between some mapping ϕ on $\mathbf{x}$ and $\mathbf{y}$, in other words, $k(\mathbf{x},\mathbf{y}) = \langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle$. Assume that $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$. Consider the formula for $\phi_{\mathbf{z}}(\mathbf{x}) = (\pi/2)^{-d/4}\exp(-\|\mathbf{x}-\mathbf{z}\|^2)$ which is an infinite dimensional function over $\mathbf{z}\in\mathbb{R}^d$ (rather than a finite dimensional vector with $\mathbf{z}$ being a discrete index as we did in class). Similarly, we have $\phi_{\mathbf{z}}(\mathbf{y}) = (\pi/2)^{-d/4}\exp(-\|\mathbf{y}-\mathbf{z}\|^2)$. Then, we define the kernel as $k(\mathbf{x},\mathbf{y}) = \langle\phi_{\mathbf{z}}(\mathbf{x}),\phi_{\mathbf{z}}(\mathbf{y})\rangle = \int_{\mathbf{z}}\phi_{\mathbf{z}}(\mathbf{x})\times\phi_{\mathbf{z}}(\mathbf{y})\mathbf{dz}$. This integral evaluates to

$$\begin{aligned} k(\mathbf{x},\mathbf{y}) &= \int_{\mathbf{z}}(\pi/2)^{-d/4}\exp\left(-\|\mathbf{x}-\mathbf{z}\|^2\right)\times(\pi/2)^{-d/4}\exp\left(-\|\mathbf{y}-\mathbf{z}\|^2\right)\mathbf{dz} \\ &= (\pi/2)^{-d/2}\int_{\mathbf{z}}\exp\left(-\mathbf{x}^\top\mathbf{x}-\mathbf{z}^\top\mathbf{z}+2\mathbf{x}^\top\mathbf{z}\right)\exp\left(-\mathbf{y}^\top\mathbf{y}-\mathbf{z}^\top\mathbf{z}+2\mathbf{y}^\top\mathbf{z}\right)\mathbf{dz} \\ &= (\pi/2)^{-d/2}\exp\left(-\mathbf{x}^\top\mathbf{x}-\mathbf{y}^\top\mathbf{y}\right)\int_{\mathbf{z}}\exp\left(-2\mathbf{z}^\top\mathbf{z}+2\left(\mathbf{y}+\mathbf{x}\right)^\top\mathbf{z}\right)\mathbf{dz} \end{aligned}$$

Define $\mathbf{r} = (\mathbf{y} + \mathbf{x})/2$ for short-hand and write...

$$
\begin{aligned}
k(\mathbf{x}, \mathbf{y}) &= (\pi/2)^{-d/2} \exp\left(-\mathbf{x}^\top \mathbf{x} - \mathbf{y}^\top \mathbf{y}\right) \int_{\mathbf{z}} \exp\left(-2\mathbf{z}^\top \mathbf{z} + 4\mathbf{r}^\top \mathbf{z}\right) \mathbf{dz} \\
&= (\pi/2)^{-d/2} \exp\left(-\mathbf{x}^\top \mathbf{x} - \mathbf{y}^\top \mathbf{y}\right) \exp\left(2\mathbf{r}^\top \mathbf{r}\right) \int_{\mathbf{z}} \exp\left(-2\mathbf{z}^\top \mathbf{z} + 4\mathbf{r}^\top \mathbf{z} - 2\mathbf{r}^\top \mathbf{r}\right) \mathbf{dz} \\
&= (\pi/2)^{-d/2} \exp\left(-\mathbf{x}^\top \mathbf{x} - \mathbf{y}^\top \mathbf{y}\right) \exp\left(2\mathbf{r}^\top \mathbf{r}\right) \int_{\mathbf{z}} \exp\left(-2\|\mathbf{z} - \mathbf{r}\|^2\right) \mathbf{dz} \\
&= (\pi/2)^{-d/2} \exp\left(-\mathbf{x}^\top \mathbf{x} - \mathbf{y}^\top \mathbf{y}\right) \exp\left(2\mathbf{r}^\top \mathbf{r}\right) (\pi/2)^{d/2} \\
&= \exp\left(-\mathbf{x}^\top \mathbf{x} - \mathbf{y}^\top \mathbf{y}\right) \exp\left(\frac{1}{2}\mathbf{x}^\top \mathbf{x} + \frac{1}{2}\mathbf{y}^\top \mathbf{y} + \mathbf{x}^\top \mathbf{y}\right) \\
&= \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{x} - \frac{1}{2}\mathbf{y}^\top \mathbf{y} + \mathbf{x}^\top \mathbf{y}\right) \\
&= \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2\right)
\end{aligned}
$$

□

Problem 3

a)

By applying Lagrangian, we can write the primal as:

$$\min_{\xi,w} \max_{\alpha} \ \mathcal{L}(\xi, w, \alpha)$$

where

$$\mathcal{L}(\xi, w, \alpha) = \sum_{i=1}^{n} \xi_i^2 + \lambda w^\top w - \sum_{i=1}^{n} \alpha_i (w^\top x_i - y_i - \xi_i),$$
$$\alpha \in \mathbb{R},$$

Then we write the dual:

$$\max_{\alpha} \min_{\xi,w} \ \mathcal{L}(\xi, w, \alpha)$$

Take partial derivatives over ξ, w.

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 2\xi_i + \alpha_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial w} = 2\lambda w - \sum_{i=1}^{n} \alpha_i x_i = 0$$

Therefore,

$$\xi = -\frac{\alpha}{2}, \ \ w = \frac{1}{2\lambda} X\alpha.$$

Plug in ξ, w,

$$\begin{aligned} \mathcal{L}_D(\alpha) &= \frac{1}{4}\alpha^\top \alpha + \frac{1}{4\lambda}\alpha^\top X^\top X\alpha - \frac{1}{2\lambda}\alpha^\top X^\top X\alpha + \alpha^\top y - \frac{1}{2}\alpha^\top \alpha \\ &= -\frac{1}{4}\alpha^\top \alpha - \frac{1}{4\lambda}\alpha^\top X^\top X\alpha + \alpha^\top y \end{aligned}$$

The dual problem is:

$$\max_{\alpha} \ \mathcal{L}_D(\alpha)$$

Take partial derivative over α, we have:

$$\frac{\partial \mathcal{L}_D}{\partial \alpha} = -\frac{1}{2}\alpha - \frac{1}{2\lambda} X^\top X\alpha + y = 0$$

The solution to the dual problem:

$$\alpha = 2\lambda (X^\top X + \lambda I)^{-1} y$$

b)

Using the results in part a, we have:

$$w = \frac{1}{2\lambda} X\alpha = X(X^\top X + \lambda I)^{-1} y$$

The linear regression can be written as:

$$\hat{y} = y^\top (X^\top X + \lambda I)^{-1} X^\top \hat{x} = \sum_{i=1}^{n} y^\top (X^\top X + \lambda I)^{-1} (x_i \cdot \hat{x})$$

Replace x with $\phi(x)$, and $X^\top X$ with Gram matrix K, we get:

$$\alpha = 2\lambda (K + \lambda I)^{-1} y$$

Kernel regression:

$$\hat{y} = \sum_{i=1}^{n} y^\top (K + \lambda I)^{-1} k(x_i, \hat{x}).$$

Problem 4

First, we normalize the data X, and then randomly split the dataset in to two half for cross validation. Then, we train SVM using polynomial kernel and RBF kernel with different parameters and costs. Testing errors are plotted in the following two graphs.

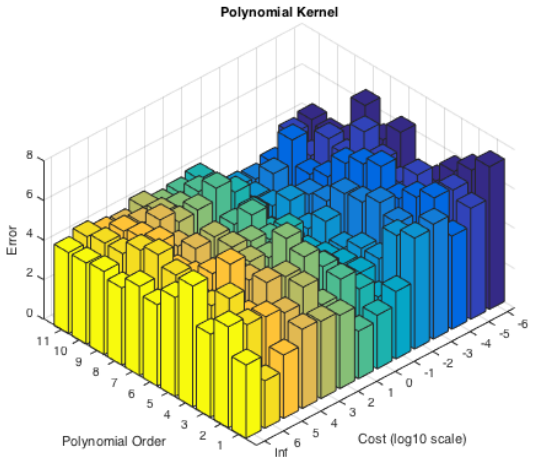

Figure 1: Testing errors using polynomial kernel with different orders and costs

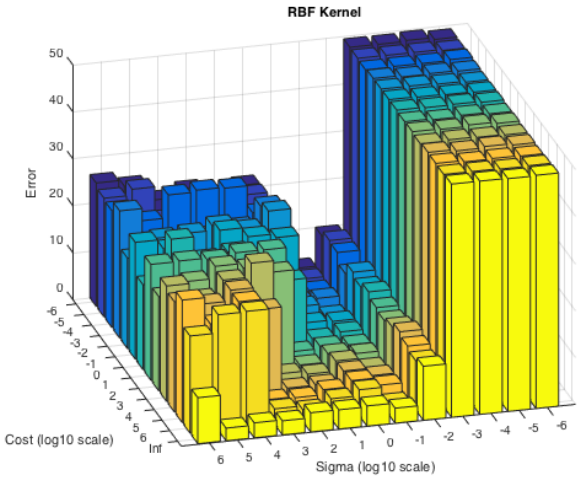

Figure 2: Testing errors using RBF kernel with different variances and costs

Problem 4

```
function question4

% load the dataset
load hw2-2015-dataset;

n = size(X,1);
idxtrain = randsample(n, n/2);

% split the dataset
idxtest  = setdiff(1:n, idxtrain);

trnX = X(idxtrain, :);
tstX  = X(idxtest,  :);

trnY = Y(idxtrain, :);
tstY  = Y(idxtest,  :);

% classify using polynomials
d_max = 50;
err_d = zeros(1, d_max);

global p1;
for d=1:d_max
    p1 = d;

    % train
    [nsv, alpha, bias] = svc(trnX,trnY,'poly',inf);

    % predict
    predictedY = svcoutput(trnX,trnY,tstX, 'poly',alpha,bias);

    % compute test error
    err_d(d)     = svcerror(trnX,trnY,tstX,tstY,'poly',alpha,bias);
end

% plot error vs polynomial degree
f = figure(1);
clf(f);
plot(1:d_max, err_d);
print(f, '-depsc', 'poly.eps');
```

```
% classify using rbfs
sigmas = .1:.1:2;
err_sigma = zeros(1, numel(sigmas));
for sigma_i=1:numel(sigmas)
    p1 = sigmas(sigma_i);

    % train
    [nsv, alpha, bias] = svc(trnX,trnY,'rbf',inf);

    % predict
    predictedY = svcoutput(trnX,trnY,tstX, 'rbf',alpha,bias);

    % compute test error
    err_sigma(sigma_i) = svcerror(trnX,trnY,tstX,tstY,'rbf',alpha,bias);
end

% plot error vs polynomial degree
f = figure(1);
clf(f);
plot(sigmas, err_sigma);
print(f, '-depsc', 'rbf.eps');
```