Solution to Quiz 2 (ECE7252 Spring 2008)

1. Consider the following kernel function: $K(x_i, x_j) = (< x_i, x_j >)^2$, where $< x, y >= x^T y$ denotes the inner product of vectors x and y. Verify that for each of the following two feature mappings $\phi(x)$, it holds that $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$. Note in R^2, each vector $x_i = (x_{i1}, x_{i2})^T$. Show your calculation.

   (a) With $\phi(x) = \frac{1}{\sqrt{2}}(x_1^2 + x_2^2, 2x_1x_2, x_1^2 - x_2^2)^T$,

   $$< \phi(x_i), \phi(x_j) >= \frac{1}{2}\{[x_{i1}^2 + x_{i2}^2][x_{j1}^2 + x_{j2}^2] + [2x_{i1}x_{i2}][2x_{j1}x_{j2}] + [x_{i1}^2 - x_{i2}^2][x_{j1}^2 - x_{j2}^2]\}; \quad (1)$$

   or

   $$< \phi(x_i), \phi(x_j) >= [x_{i1}x_{j1} + x_{i2}x_{j2}]^2 = (< x_i, x_j >)^2 = K(x_i, x_j). \quad (2)$$

   (b) With $\phi(x) = (x_1^2, x_1x_2, x_1x_2, x_2^2)^T$,

   $$< \phi(x_i), \phi(x_j) >= [x_{i1}^2 x_{j1}^2] + 2[x_{i1}x_{i2}x_{j1}x_{j2}] + [x_{i2}^2 x_{j2}^2]; \quad (3)$$

   or

   $$< \phi(x_i), \phi(x_j) >= [x_{i1}x_{j1} + x_{i2}x_{j2}]^2 = (< x_i, x_j >)^2 = K(x_i, x_j). \quad (4)$$

2. It is easy to show that $J(w) = \frac{1}{2}(|w^T x| - w^T x) = 0$ for correctly classified samples in the training set when $sign(w^T x) > 0$, and $J(w) = -w^T x$ for misclassified samples. So by finding w to minimize $J^*(w) = \sum_i J(w) = \sum_i w^T x_i$ over all x_i in the set C^* of incorrectly classified data in the training phase with:

   $$\nabla J^*(w) = \sum_{i \in C^*} (-x_i). \quad (5)$$

   This is similar to the Rosenblatt perceptron learning algorithm.

3. Consider a simple linear SVM classifier, $(w_1 x + w_0)$, and a nonlinear SVM classifier, $(w_1\phi(x) + w_0)$, where $\phi(x) = (x, x^2)^T$. (a) If $x_1 = (-1, -1)^T$ and $x_2 = (1, 1)^T$ belong to Class 1, and $x_3 = (0, 0)^T$ belongs to Class 2, then they are not linearly separable. Now with $\phi(x) = (x, x^2)^T$, we have transformed $x_1 = (-1, 1)^T$ and $x_2 = (1, 1)^T$, and $x_3 = (0, 0)^T$. They are now linearly separable. (b) The line $y = \frac{1}{2}$ linearly separates the two classes in the transformed space.

4. In the SVM formulation we try to minimize $||w||^2$ such that $y_i(w^T x_i + b) \geq 1, i = 1, ..., n$ for the set of training examples denoted by $D = \{(x_i, y_i), i = 1, ..., n | y_i = -1 \text{ or } +1\}$. We can also consider from the viewpoint of robust optimization that every observed example, x_i, is corrupted by noise, and the true example, x_i^*, is within a radius, ρ, of the observed sample, x_i, or $||x_i^* - x_i||^2 \leq \rho^2$. Since we are not sure which example within the above sphere is a true example, we require every example in the sphere to be classified correctly, or for all $||x_i^* - x_i||^2 \leq \rho^2$, $y_i(w^T x_i^* + b) \geq 0$. We now have an alternative optimization problem by searching for the decision boundary with the largest ρ, i.e. to maximize ρ, such that $||x_i^* - x_i||^2 \leq \rho^2$, $y_i(w^T x_i^* + b) \geq 0, i = 1, ..., n$.

   By considering the case of a robust optimization setting, if we maximize the radius surrounding each data point until one of the neighborhood spheres touches the decision boundary hyperplane, it is clear that the first such data point has to be one of the support vectors in the original SVM optimization setting. It can also be argued that only the spheres surrounding the support vectors will touch the decision hyperplane first, and it can even be concluded that the maximum radius is $\hat{\rho} = \frac{1}{||w||}$.

5. The following table listed the number of outcomes to finish series in 4-7 games

	Play 4 games	Play 5 games	Play 6 games	Play 7 games
A wins	1	4	10	20
B wins	1	4	10	20

P(outcome in 4 games) =0.5 ^4 =0.0625
P(outcome in 5 games) =0.5^5=0.03125
P(outcome in 6 games) =0.5^6=0.015625
P(outcome in 7 games) =0.5^7=0.0078125

P(A wins in 4 games)= 1(0.5 ^4) =0.0625
P(B wins in 4 games)= 1(0.5 ^4) =0.0625
P(A wins in 5 games)= 4(0.5 ^5) =0.125
P(B wins in 5 games)= 4(0.5 ^5) =0.125
P(A wins in 6 games)= 10(0.5 ^6) =0.15625
P(B wins in 6 games)= 10(0.5 ^6) =0.15625
P(A wins in 7 games)= 20(0.5 ^7) =0.15625
P(B wins in 7 games)= 20(0.5 ^7) =0.15625

P(Y=4)=0.125
P(Y=5)=0.250
P(Y=6)=0.3125
P(Y=7)= 0.3125

Since X is a random variable describing outcomes (with log here with base 2)

H(X)= -2 x P(outcome in 4 games) x log P(outcome in 4 games) -8 x P(outcome in 5 games)x log P(outcome in 5 games) - 20 x P(outcome in 6 games) x log P(outcome in 6 games) – 40 x P(outcome in 7 games) x log P(outcome in 7 games)
H(X) = -2(0.0625)log(0.0625) – 8 (0.03125)log(0.03125) -20(0.015625)log(0.015625) -40(0.0078125)log(0.0078125)
H(X) = 5.8125

Y is the number of games played
H(Y) = -P(Y=4)log P(Y=4)- P(Y=5)log P(Y=5) -P(Y=6)log P(Y=6) -P(Y=7)log P(Y=7)
H(Y)= -0.125 log(0.125) – 0.250 log (0.250) -0.3125 log (0.3125) -0.3125 log (0.3125)
H(Y) = 1.924

Since Y provides no surprise if we already know X then
H(Y/X) = 0 ;

Since H(X)+H(Y/X) = H(Y) +H(X/Y)
Therefore
H(X/Y) = 5.8125- 1.924 =3.8885

I(X, Y)=H(X)-H(X/Y) = 5.8125-3.8885 =1.924=H(Y)-H(Y|X)