HW Assignment 5 (Due by 1:30 pm on Oct 24)

1 Theory (110 points)

1. [**Properties of Linear Discriminants**, 20 points]
We have proven in class that the distance between origin and the decision hyperplane $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$ is equal with $-w_0/\left\|\mathbf{w}\right\|$. Prove that the margin between a point $\mathbf{x}$ and the same decision hyperplane is equal with $h(\mathbf{x})/\left\|\mathbf{w}\right\|$.

2. [**Bonus**, 20 points]
Prove the two properties above for the general n-dimensional case.

3. [**Fisher Criterion and Least Squares**, 30 points]
Show that the Fisher criterion can be written in the vectorized form shown below:
$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$$

4. [**Fisher Criterion (*)**, 20 points]
Reference the PRML Chapter 4 material available on Blackboard under Content.
Using the definitions of the between-class and within-class covariance matrices given by (4.27) and (4.28), respectively, together with (4.34) and (4.36) and the choice of target values described in Section 4.1.5, show that the expression (4.33) that minimizes the sum-of-squares error function can be written in the form (4.37).

5. [**Perceptrons**, 40 points]
Consider a training set that contains the following 8 examples:

$\mathbf{x}$	x_1	x_2	x_3	$t(x)$
$\mathbf{x}^{(1)}$	0	0	0	+1
$\mathbf{x}^{(2)}$	0	1	0	+1
$\mathbf{x}^{(3)}$	1.5	0	-1.5	+1
$\mathbf{x}^{(4)}$	1.5	1	-1.5	+1
$\mathbf{x}^{(5)}$	1.5	0	0	–1
$\mathbf{x}^{(6)}$	1.5	1	0	–1
$\mathbf{x}^{(7)}$	0	0	-1.5	–1
$\mathbf{x}^{(8)}$	0	1	-1.5	–1

   (a) Prove that the perceptron algorithm does not converge on this dataset. Do not forget to include the bias.

   (b) Consider a kernel perceptron that uses a polynomial kernel $k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T\mathbf{y})^d$. What is the smallest degree d for which the kernel perceptron would converge on this dataset? Provide a proof of your answer.

6. [**Perceptrons**, 10 points]
A kernel perceptron for binary classification is run for a number of epochs E on a training dataset containing N examples, resulting in the dual parameters α_1, α_2, ..., α_N. What is the total number of mistakes that are made during training?

7. [**Matrix Computations**, 10 points]
   Let $U \in R_{k \times m}$ and $X \in R_{n \times m}$. Let u_i and x_i be the i-th columns of U and X, respectively, for $1 \leq i \leq m$. Prove that $UX^T = \sum_{i=1}^{m} u_i x_i^T$.

2 Submission

Turn in a hard copy of your homework report at the beginning of class on the due date. On this theory assignment, **clear and complete explanations and proofs of your results are as important as getting the right answer**.