Machine Learning Midterm

This exam is open book. You may bring in your homework, class notes and text- books to help you. You will have 1 hour and 15 minutes. Write all answers in the blue books provided. Please make sure YOUR NAME is on each of your blue books. Square brackets [] denote the points for a question. ANSWER ALL FOUR QUESTIONS FOR FULL CREDIT

1. **PCA and ICA**
   (a) [5] For an $n \times n$ symmetric matrix A what can one say about its eigenvectors?
   (b) [5] For a symmetric positive definite matrix, what can one say about its eigenvalues?
   (c) [5] If a matrix's eigenvalues are real, does that imply the matrix is symmetric? Say why or why not.
   (d) [10] Compare and contrast PCA and ICA. What are the principal differences between them?

2. **Information Theory**
   (a) [10] Where x and y are discrete random variables with probability distribution p(x,y), their entropy is given by

$$H(x,y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$

   Where conditional entropy is given by $H(x|y) = H(x,y) - H(y)$ and mutual information is given by $I(x;y) = H(x) - H(x|y)$, show that

$$I(x;y) = \sum_x \sum_y p(x,y) \log\Big(\frac{p(x,y)}{p(x), p(y)}\Big)$$

   (b) [5] Is this formula the K-L divergence? Explain.
   (c) [10] Consider the following six data points
   (ie.g. from the table $< red, circle > +$ is a data point):

	circle	square	triangle
red	+	+	+
blue	+	-	-

   Write down, but do NOT evaluate, formulas for the information gain for TWO possible choices of feature.

3. **Support Vector Machines**

(a) [15] For a kernel function, and data points $\{\mathbf{x}^{(i)}, i = 1, \ldots, m\}$, we can define a matrix

$$\{K_{ij}\} = K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \phi(\mathbf{x}^{(i)})^T \phi(\mathbf{x}^{(j)})$$

Where $\phi_k(\mathbf{x})$ denotes the k-th coordinate of $\phi(\mathbf{x})$, show that for any vector $\mathbf{x}$,

$$\mathbf{x}^T K \mathbf{x} = \sum_i \sum_j \mathbf{x}_i K_{ij} \mathbf{x}_j = \sum_k \Big(\sum_i x_i \phi_k(\mathbf{x}) \Big)^2$$

(b) [5]Is the kernel matrix symmetric? Show why or why not.

(c) [5] *Mercer's Theorem* states that any kernel matrix K must be positive semidefinite i.e $\mathbf{x}^T K \mathbf{x} \geq 0$. Does the kernel in (a) have this property? Say why or why not.

4. **Learning Theory**

(a) [5] For an n-dimensional space, what is the VC dimension of a function set that consists of linear separation surfaces?

(b) [5]In two dimensions, lines cannot shatter arbitrary sets of four points. What about the function set that consists of *arbitrary two-dimensional ellipses*? Show explicitly whether or not this function set can shatter sets of four points. What does this say about its VC dimension?

(c) [15] A model neuron has m simultaneous inputs each representing a voltage spike $X = 1$ with probability ϕ. The neuron's job is to estimate ϕ using the formula (all weights are =1):

$$\hat{\phi} = \frac{1}{m} \sum_i X_i$$

If you need to know the value of ϕ to within 10%, with a probability of error less than 0.05, how many inputs does the neuron need?