Machine Learning Midterm

This TWO-SIDED 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. **Eigenvalues**
   (a) [15] For the matrix A given by

   $$A = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix}$$

   find its eigenvalues and eigenvectors.
   The detreminant is

   $$(3-\lambda)(2-\lambda) - 2 = 0$$

   which be factored to get $\lambda = 4, 1$ with eigenvectors $\{2, 1\}$ and $\{-1, 1\}$ respectively.

   (b) [5]Could this matix be a covariance matrix? Say why or why not.
   Cannot because its not symmetric.

   (c) [5] Could this matrix be positive definite? Say why or why not Is Positive Definite because the eigenvalues are both positive.

2. **Information Theory**
   (a) [5] Consider the discrete distribution $\{p_k | k = 1, \ldots, N\}$. What is an expression for the entropy of this distribution?
   The equation is given by

   $$H = -\sum_{k=1}^{N} p_k \log p_k$$

   (b) [5] Now consider the distribution that maximizes the entropy measure. Sketch what should it look like. The entropy $p_k = \frac{1}{N}, \forall k$

   (c) [15] Confirm your answer by*maximizing* entropy *subject to the constraint* $\sum_{k=1}^{N} p_k = 1$.
   Use a Lagrange multiplier

   $$\max J = -\sum_{k=1}^{N} p_k \log p_k + \lambda \sum_{k=1}^{N} p_k - 1$$

$$J_{p_k} = -\log p_k - 1 + \lambda = 0$$

Since this holds for all p_k, they must all be equal. Hence $p_k = \frac{1}{N}$ in order to sum to one.

3. **Decision Trees**

   (a) [10] You are building a decision tree to classify (x, y), points in two dimensions. Nodes in the tree test one of the two coordinates against arbitrary constants e.g. $(x \leq c?)$ where c is a constant. For d nodes in the tree, what is the VC-dimension of the decision tree? Illustrate your answer for a small number of points.

   The trick is to divide the points into seperate rectangles. This takes $d-1$ nodes. Hence that is the VC dimension.

Decision Trees Cont.

   (b) [15]In a general decision tree, with a node n and question q, let $n_+(n, q)$ be the denote the right child of n after a node split and n_- be the left child. Futhermore let $\eta(n, q)$ be the fraction of points in n that are moved to $n_-(n, q)$. Your friend thinks that in building the decision tree, you do not have to use *information gain*, but that just *counting the number of missclassifications* will work. For any node n and class $l \in [1, k]$ $p_l(n)$ denotes the fraction of points at n that belong to class l. Then measure the number of missclassifications $F(n)$ as

$$F(n) = 1 - \max_{l \in [1,k]} p_l(n)$$

   Your friend says that for each node we can to pick the question q that maximizes

$$F(n) - [\eta(n, q)F(n_-(n, q) + (1 - \eta(n, q))F(n_+(n, q))].$$

   Compare two features from the attatached data set to show how this would work.

4. **Support Vector Machines**

   (a) [10] By appropriately differentiating the primal problem, one finds that:

$$\mathbf{w} = \sum_{j=1}^{m} \alpha_i d_i \mathbf{x}_i \tag{1}$$

   and also that

$$\sum_{j=1}^{m} \alpha_i d_i = 0$$

   A given support vector obeys

$$\mathbf{w}^T \mathbf{x}_i + b = d_i$$

Use Eq. 1 to show that

$$b = d_i - \sum_{j=1}^{m} \alpha_j d_j \mathbf{x}_j^T \mathbf{x}_i \tag{2}$$

Eq. 2 follows from the previous Eqs. by substitution.

(b) [15] Use the above equations to show that the margin between separating lines,ρ, can be expressed in terms of an L_1 norm ($||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|$)

$$\frac{1}{\rho^2} = ||\alpha||_1$$

where $\alpha = \{\alpha_1, \ldots, \alpha_m\}$.
Hint: Multiply Eq. 2 by $\sum_{j=1}^{m} \alpha_i d_i$.
Taking the hint...

$$\sum_{j=1}^{m} \alpha_i d_i b = \sum_{j=1}^{m} \alpha_i d_i^2 - \sum_{j=1}^{m} \alpha_i d_i \sum_{j=1}^{m} \alpha_j d_j \mathbf{x}_j^T \mathbf{x}_i$$

The term on the LHS is zero and $d_i^2 = 1$, so:

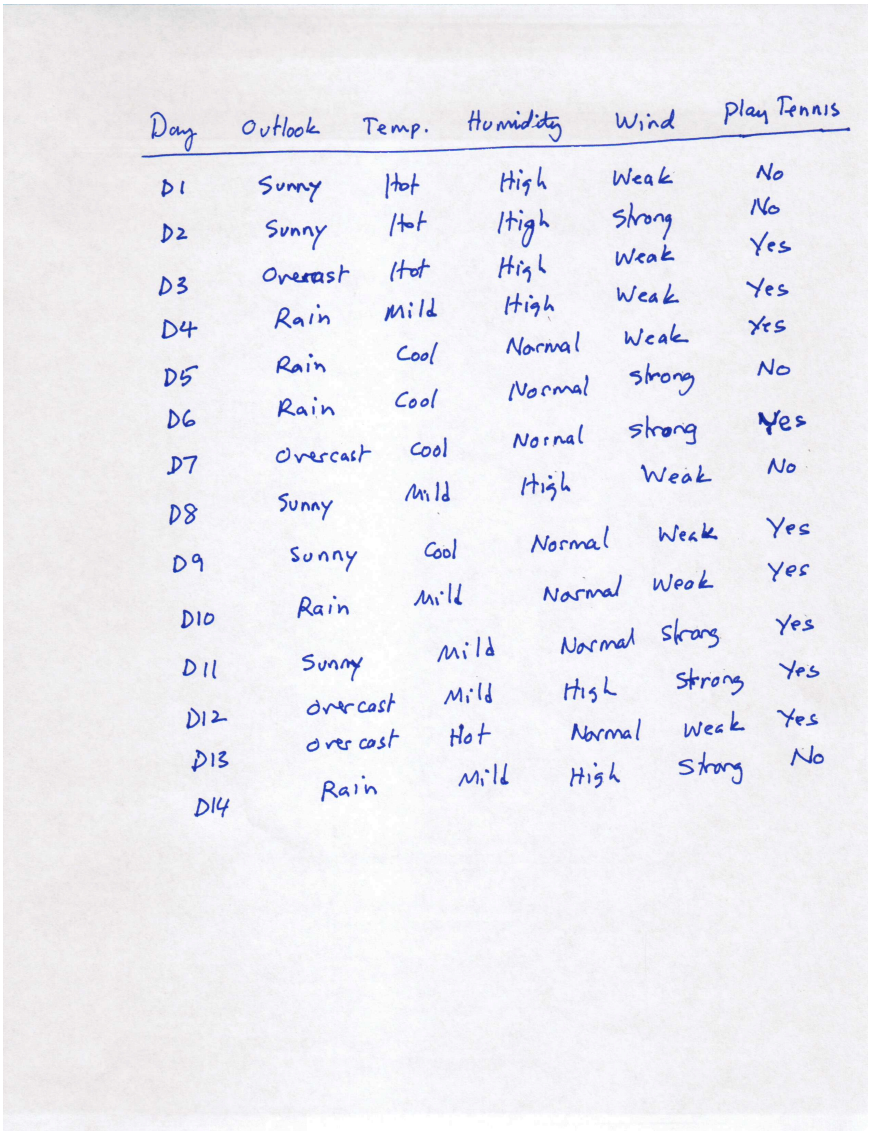
$$0 = \sum_{j=1}^{m} \alpha_i - \sum_{j=1}^{m} \alpha_i d_i \sum_{j=1}^{m} \alpha_j d_j \mathbf{x}_j^T \mathbf{x}_i$$

Now using Eq. 1:

$$0 = \sum_{j=1}^{m} \alpha_i - \mathbf{w}^T \mathbf{w}$$

But $\mathbf{w}^T \mathbf{w} = \frac{1}{\rho^2}$

Data for Decision Tree Question



Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	overcast	Mild	High	Strong	Yes
D13	overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No