DUE: MONDAY MIDNIGHT MARCH 3

CS 391L Machine Learning Assignment 3

This problem set is to be done by yourself. You can discuss the questions in generality with each other, but you should not collaborate on solving the individual problem details.

1. **Linear Algebra**
   (a) [10]Where the covariance of $\mathbf{x}$ is given by Σ, and
   $$\mathbf{y} = A\mathbf{x} + \mathbf{b}$$
   show that the covariance of $\mathbf{y}$ is given by $A\Sigma A^T$.
   (b) [5] Show that for any interger k
   $$A^k\mathbf{x} = \lambda^k\mathbf{x}$$
2. **Information Theory** X and Y are independent random variables that are identically distributed so that H(X)=H(Y), but they are not necessarily independent. Define$r$ by
   $$r = 1 - \frac{H(Y|X)}{H(X)}$$
   (a) [5] Show that $r = \frac{I(X,Y)}{H(X)}$.
   (b) [5] Show that $0 \leq r \leq 1$.
   (c) [5] When is $r = 0$ and when is $r = 1$?

3. **Neural networks**
   The standard nonlinear function used in neural networks is the sigmoid function $g(v) = \frac{1}{1+e^{-v}}$. But sometimes one uses $g(v) = \tanh(v)$
   (a) [5] What would be the advantage of the $\tanh$ function?
   (b) [15] The Backpropagation algorithm uses an objective function H given by
   $$H = \frac{1}{2}\|\mathbf{x}^K - \mathbf{d}\|^2 + \sum_{k=0}^{k=K-1} (\lambda^{k+1})^T[-\mathbf{x}^{k+1} + g(W^k\mathbf{x}^k)]$$
   where the first term measures the cost of errors reproducing the pattern and the second term constrains the system to follow its dynamic equations.
   By using a 2×2 W matrix, show that for $k < K$, the partial derivative $\frac{\partial H}{\partial w_{ij}}$ is given by
   $$\frac{\partial H}{\partial w_{ij}^k} = \lambda_i^{k+1} x_j^k g'(\mathbf{w}_i^k \cdot \mathbf{x}^k) \quad (1)$$

Color	Size	Noise	Like
red	med	loud	yes
red	medium	quiet	yes
red	small	quiet	no
blue	small	loud	yes
red	large	loud	no
blue	small	quiet	yes
blue	medium	loud	yes
red	medium	loud	no
blue	small	quiet	yes
red	large	quiet	no
red	small	quiet	yes
blue	large	loud	no

4. **Decision Trees** Likes and dislikes of toys are captured in the following table.
   (a) [10] Build the decision tree based on ID3's *information gain* to model this data.
   (b) [5] Suppose some data records have missing values. How would you handle this in building the tree?

5. **Lagrange multipliers** A conical container of radius r and height h has volume $V = \frac{1}{3}\pi r^2 h$ and surface area equal to $\pi r s + \pi r^2$ where $s^2 = r^2 + h^2$.
   (a) [15] For a fixed area A, find the radius and height that maximize the conical container's volume.

6. **VC Dimension** [15] Show that the VC dimension of a decision tree with n nodes in dimension M is O(m log M). Hint: How many different classifiers are there?

7. **Support Vector Machines** [15] The Mercer kernel used to solve the XOR problem is given by
$$k(\mathbf{x}, \mathbf{x_i}) = (\mathbf{1} + \mathbf{x^T x_i})^{\mathbf{p}}$$
What is the smallest positive integer p for which the XOR problem is solved? What is the result of using a value pf p larger than the minimum?