CS60050: Machine Learning

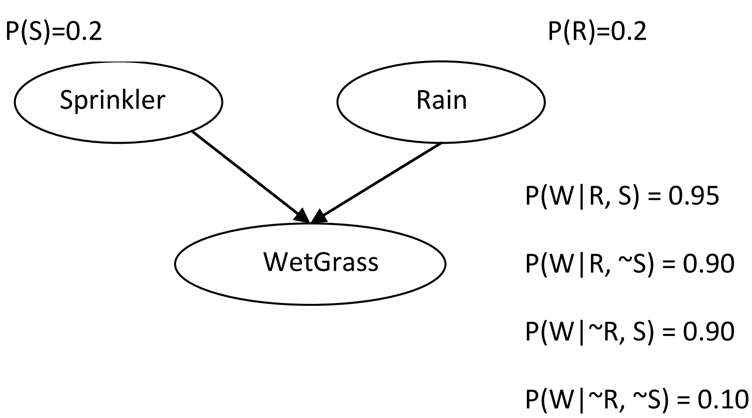
End-semester Examination, Autumn 2017

Time= 3 hrs. Marks: 100. Answer all FOUR questions. Make suitable assumptions if required.

1.(a). Construct the complete dendogram of the following eight points in one dimension using the single-linkage clustering algorithm: *{-5.5, -4.1, -3.0, -2.6, 10.1, 11.9, 12.3, 13.6}*. Show the steps. Using the dendogram also find the number of natural clusters in the data. Justify your answer. **[15]**

(b). A set of n points is partitioned into c disjoint clusters $D_1, D_2, \dots, D_c$, the mean m_i for a cluster D_i is defined as $m_i = \frac{1}{|D_i|}\sum_{x \in D_i} x$. The sum-squared error is defined as: $J_e = \sum_{i=1}^{c} \sum_{x \in D_i} \|x - m_i\|^2$. Consider a set of $n = 2k + 1$ one-dimensional points, k of which coincide at $x = -2$, k at $x = 0$, and one at $x = a > 0$. Show that the two-cluster partitioning that minimizes J_e groups k points at $x = 0$ with the one at $x = a$ if $a^2 < 2(k + 1)$. What is the optimal grouping if $a^2 > 2(k + 1)$? **[10]**

2.(a). Considering the following Bayesian network, calculate P(R | W), P(R | W, S), and P(R | W, ~S). **[15]**



(b). We use the notation $a \perp b \mid c$ to denote that a is *conditionally independent* of b given c. Formally, show that $a \perp b, c \mid d$ implies $a \perp b \mid d$. **[10]**

3.(a). Consider a multilayered perceptron with single hidden layer. There are two input nodes, one hidden layer node, and one output node. The hidden layer node activation function is given by a sigmoid function of the form, $\sigma(\mathrm{x}) = \frac{1}{1+e^{-x}}$. Show that there exists an equivalent network which computes exactly the same function, but with hidden layer node activation function given by a tanh function of the form, $\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$. The equivalent network can have different weight values and structure. **[10]**

(b). Suppose we have a multilayered perceptron where the output squared error is represented by the function $J(W)$, W being the weight values. The network updates weights by the usual squared error gradient descent backpropagation rule with learning rate η. Now, we add an additional update factor for weight decay. The additional update factor is of the form $W_{i,j}^{new} = W_{i,j}^{old}(1 - \varepsilon)$. Show that this amounts to performing gradient descent on the modified error function $J_m = J(W) + \frac{2\varepsilon}{\eta} W'W$. **[15]** **P.T.O**

4.(a). Draw a two-class two-dimensional data such that (i) PCA and LDA find the same direction, and (ii) PCA and LDA find totally different directions. **[10]**

(b). Define when a concept class is denoted as PAC learnable. **[5]**

(c). Show that the VC-dimension of axis aligned rectangles in two dimensional plane is 4. **[10]**

---------- BEST WISHES ---------