International Institute of Information Technology, Hyderabad

CSE471 (Spring 2015)FINAL EXAM

STATISTICAL METHODS IN AI

Time: 3 Hours **Max. Marks: 150**

INSTRUCTIONS

1. Scientific Calculators are Allowed. Please be as concise as possible in your answers.
2. Please start each question on a separate page, indicating the question and sub-part numbers clearly.
3. When a question is *ambiguous*, state so. Make reasonable assumptions and write them clearly before you answer the question.

PART A: **(14 × 5 = 70 Marks)**

Write Brief (3-4 line) Answers!

1. In Support Vector Machines (SVM), what are the "support vectors"?
2. Give an example (show graphically) where Principal Component Analysis (PCA) and Fisher Linear Disriminant (FLD) give different projections for the same data set and another example where they give the same projection.
3. Will the Perceptron learning algorithm based on Minimum Squared-Error Procedure converge, if the data set is not linearly separable? If not, why not? If so, what will it converge to?
4. What property of the sigmoid activation function is important for the backpropagation (BP) learning algorithm? What other activation function has this property? A multilayer network of linear nodes can create a nonlinear decision boundary – TRUE OR FALSE? Briefly justify your answers.
5. What are basic assumptions on the distribution of the data in k-Means clustering algorithm? What parameters does the user specify, if any, in this algorithm?
6. In decision tree learning, when is a node "pure"? The attributes nearer to the leaf nodes in a decision tree are considered important for the given classification problem. TRUE OR FALSE? Briefly justify your answer.

7. Suppose we want to build a classifier with data sets having multiple features / attributes. Can you suggest two ways of handling data items with missing feature values?

8. Of the two algorithms, k-Nearest Neighbour and k-Means, which one is supervised and which is unsupervised? Why?

9. Assume there are c classes $w_1, \ldots, w_c$, and one feature vector $\overrightarrow{x}$, give the Bayes Decision Rule for classification in terms of *a priori* probabilities of the classes and class-conditional probability densities of $\overrightarrow{x}$. If you know prior information about the problem, how can you incorporate that in a Bayes Decision Rule?

10. Assume that the feature vector $\overrightarrow{x}$ for a given class ω_i follows Normal Density and that the features are statistically independent but with unknown mean and each feature has the same variance, σ^2 for the two classes. What is the the shape of the decision boundary (give a sketch) and what is the impact of prior probabilities on the location of the boundary?

11. How are Maximum Likelihood Estimate (MLE) and Maximum A Posteriori (MAP) estimates related to each other and when is the MAP estimate equivalent to MLE?

12. You are given a 1-D dataset $D = \{0, 1, 1, 1, 2, 2, 2, 2, 3, 4, 4, 4, 5\}$. Assuming that the data come from a Gaussian density, compute the maximum likelihood estimate of the Gaussians parameters (Mean and Variance – both biased and unbiased estimate values, as applicable).

13. Give a Multilayer Neural Network solution for the *XOR* problem. Verify your solution. What is the role of the hidden layer in this problem?

14. What is the average out-sample error if the number of support vectors is 10 and the number of training samples is 1000? What do you expect will happen to the out-sample error, if the number of support vectors increases by 10-times?

PART B: Answer **any** 4 out of 5 **(4 × 20 = 80 Marks)**

1. Suppose that there is a student who decides whether or not to go for classes on any given day based on the weather, sleeping time, and whether there is an interesting class to attend. The data collected from 13 days are as shown in the table. Build a decision tree based on these observations, using ID3 algorithm with entropy impurity (Remember: Entropy $E = -\sum_j P(w_j) \log_2 P(w_j)$).

   A. Show all the steps and the resulting tree.

   B. Classify the following sample using your algorithm: $< Slept = Late, Interesting = No, Weather = Rain >$.

   C. If for a new observation, one of the attributes is missing, is it possible to find the classification from the tree constructed in Question A above. For example: $< Slept = ?, Interesting = Yes, Weather = Sunny >$.

   D. If the criterion is to construct a tree with the least height and you consider "Day" attribute as well for constructing the tree what is the minimum height you can expect? Any comments on the resulting tree?

Day	Slept	Interesting	Weather	GoToClass
D1	Normal	No	Sunny	No
D2	Normal	No	Rain	No
D3	Early	No	Sunny	Yes
D4	Late	No	Sunny	Yes
D5	Late	Yes	Sunny	Yes
D6	Late	Yes	Rain	No
D7	Early	Yes	Rain	Yes
D8	Normal	No	Sunny	No
D9	Normal	Yes	Sunny	Yes
D10	Late	Yes	Sunny	Yes
D11	Normal	Yes	Rain	Yes
D12	Early	No	Rain	Yes
D13	Early	Yes	Sunny	Yes

2. Derive the following k-Means clustering algorithm update equations from the first principles using Expectation-Maximization framework. Clearly indicate all the assumptions and show all the steps.

   E-Step $$E[z_{ij}] = \frac{e^{\frac{-1}{2\sigma^2}(x_i - \mu_j)^2}}{\sum_{n=1}^{k} e^{\frac{-1}{2\sigma^2}(x_i - \mu_n)^2}}$$

   M-Step $$\mu_j = \frac{\sum_{i=1}^{m} E[z_{ij}] x_i}{\sum_{i=1}^{m} E[z_{ij}]}$$

3. Derive the following learning rule for the weights from *input-layer to hidden layer*: $\Delta w_{ji} = \eta\, f'(net_j)\, [\sum_{k=1}^{c} w_{kj} \delta_k]\, x_i$. What is the specific form of the rule, if *hyperbolic tangent function* ($f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$) is used as the activation function in the hidden nodes? Show the graph of this activation function and specify its domain and range.

4. Is the solution vector found by Perceptron Learning algorithm unique, Why or Why not? Now, given a hyperplane $g(\mathbf{x}) = 0$, where $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$, answer the following:

   A. Show that $\mathbf{w}$ is orthogonal to the hyperplane.

   B. Also show that the value of $g(\mathbf{x})$ for any point is a scaled version of its perpendicular distance from the hyperplane.

   C. What is the distance of origin from this hyperplane?

5. **A.** Derive the dual form of the objective function for computing the maximum margin linear classifier for a linearly separable set of training samples. Show all the steps.

   B. From the above, write down the expressions for the weight vector and bias terms of the linear discriminant solution.

   C. Also explain how you can identify support vectors from the formulation derived in Question A.

   D. For a cubic kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^3$ that maps $\mathbf{x}$ in 2-D input space to 4-D feature space, what is the computational saving when computing inner products in the input space (2-D) instead of the feature space (4-D)?