Name: _____ **Score:** _____

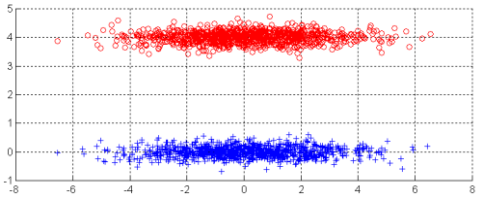
Midterm – CAP 5638: Pattern Recognition
Department of Computer Science, Florida State University, Fall 2015

Problem 1 (55 points, 5 points each) Short answers

1) Explain what would happen if we use Algorithm 4 (in Sect. 5.5.2 in the textbook) on a two-class training set that is not linearly separable. Then specify three ways that can (potentially) solve the problem.

2) Suppose D_1={2.52, 1.98, 2.46} is the training set for ω_1 and D_2={-0.36, 3.55} is the training set for ω_2, we like to estimate the posterior probabilities using Parzen window estimation with window function $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$. Compute the decision regions using window width of 1. If we like to minimize the Bayes error using the estimated posterior probability distributions for classification, should we use a smaller or a larger window width than 1? Briefly justify your answer.

3) For the following dataset of two classes (empty circles and pluses), suppose that we like to reduce the dimension to one. What would be the area under the resulting ROC curve if we use the principal component direction? What would be the area under the resulting ROC curve if we use the direction given by Fisher linear discriminant analysis? Briefly justify your answer.



4) For a three-class classification problem based on features x_1 and x_2, suppose that we use one against the rest and we obtain the following two-class linear discriminant functions: $g_1(x_1, x_2) = -x_1 + x_2 - 2$, $g_2(x_1, x_2) = x_1 + x_2 - 2$, and $g_3(x_1, x_2) = -x_1 - 1$, where g_i is the linear discriminant function for class i against the rest of the classes. Show the decision regions for the three classes and ambiguous regions. Briefly justify your answer.

5) Show the full training examples using the Kesler's construction to learn linear discriminant functions for the following three-class classification problem.

ω_1		ω_2		ω_3	
x_1	x_2	x_1	x_2	x_1	x_2
-5.0	7.6	4.3	-4.6	5.7	4.8

6) Suppose that there are 10 support vectors when a support vector machine is trained on the entire training set consisting of 1000 training samples, what is the maximal 10-fold cross validation error using support vector machines on the given training set for each of the cases: 1) there is one support vector in each of the 10 folds, 2) all the support vectors are in one fold and the other nine folds have no support vectors. Briefly justify your answer.

7) This question and the next one are about the real-time face detector using decision trees. In the process of selecting the first optimal weak classifier, for a particular feature, its values for all the face training images are 0.566 0.595 0.387 0.536 0.369 0.495 0.573 0.727 0.643 0.060 and its values for all the non-face training images are -0.209 0.161 0.054 0.577 0.146 0.604 0.020 0.128 0.237 0.525 0.326 0.063 -0.327 0.211 -0.233. Describe how to efficiently compute the optimal weak classifier for this feature and give the parameters for the resulting optimal weaker classifier and its weighted error.

8) Suppose the feature from the previous question gives the lowest error among all the features, specify the weights for all the samples to be used for computing the second weak classifier. You need to justify your answers.

9) Plot the decision boundary and label the decision regions resulting from the (3, 2)-nearest neighbor rule for the following two-dimensional dataset: (5, 0), (0, -5), (2, -1) are from class 1 and (3, 5), (0, 3) (3, 3) are from class 2. You need to briefly describe your steps.

10) Given a training set D={x_1, …, x_n}, derive the maximum likelihood estimate for the following model:

$$p(x \mid \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x+2)^2}{2\sigma^2}\right\}.$$

11) Design a neural network so that its output will be 1 for points within a unit square centered at the origin (that is, $-0.5 \le x_1, x_2 \le 0.5$) and -1 otherwise. All the neurons in the network are required to use the following transfer function $f(n) = \begin{cases} 1 & \text{for } n \ge 0; \\ -1 & \textit{otherwise.} \end{cases}$.

Problem 2 (**18 points**) For the following one-dimensional training data sampled from two categories, D_1 = {7.5, 4.5, 3.5, 6.0, 7.5} for class 1 and D_2 = {2.5, 3.5, 4.0} for class 2. We assume that the underlying probability distribution for class 1 is normal with unknown mean and variance and is uniform (i.e., the probability is $1/\theta$ for $0 \leq x \leq \theta$ and 0 otherwise) for class 2.

1) (**2 points**) Give the estimated prior for each class using **maximum likelihood** estimation.

2) (**4 points**) Estimate the class conditional for class 1 using the maximum likelihood. You need to write down the equations used and specify all the constants.

3) (**4 points**) Estimate the class conditional for class 2 using the Bayesian parameter estimation, where the prior probability distribution for the parameter is uniform between 2 and 10. You need to show intermediate steps and specify all the constants.

4) (**4 points**) Using your estimated priors and class-conditional distributions to classify -0.5, 1.5, 5.5, and 8.5.

5) (**4 points**) Sketch the ROC curve using the estimated class conditional distributions. You need to specify how the false alarm rate and hit rate are computed so that the area under curve is larger than 0.5.

Problem 3 (16 points) Answer the following questions regarding constructing a decision tree from the training set given below. Here you can only use feature x_1 or x_2 at each node.

ω_1			ω_2		
Sample label	x_1	x_2	Sample label	x_1	x_2
S1	**-5.0**	**2.7**	S3	**2.0**	**2.0**
S2	**4.0**	**-3.0**	S4	**3.5**	**1.0**
			S5	**5.5**	**0.5**

1) (**4 points**) Using the entropy impurity, show a threshold and the impurity reduction for each possible distinctive branching at the root node. A branching is considered to be distinctive when the left subset and right subset are different from the existing ones for a given feature. You need to show the entropy calculation.

2) (**6 points**) Using the entropy impurity, construct the complete decision tree (i.e., until all the leaf nodes contain only samples from one class) for the given classification problem. You need to specify the feature and the corresponding threshold at each non-leaf node.

3) (**4 points**) Specify the classification rules for ω_1 and ω_2 respectively based on the decision tree you obtained for part 2).

4) (**2 points**) Would your decision tree obtained from part 2) change if the entropy impurity gain ratio is used instead? Briefly justify your answer.

Problem 4 (**13 points**) Answer the following questions regarding the following training set.

ω_1		ω_2		ω_3		ω_4	
x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
-9.5	-12.5	-7.5	12.0	4.8	-8.0	2.0	3.0
-16.5	-13.0	-7.0	6.0	5.0	-10.0	3.0	2.0

1) (**4 points**) Show the decision boundaries and decision regions of the nearest neighbor rule for the given training set.

2) (**6 points**) Construct a k-d tree, where a splitting dimension is chosen to be the one that has the largest variance and the threshold with the largest margin with the branches to be as balanced as possible. You need to specify the threshold and the splitting dimension of each internal node and the data point at each leaf node (note that there should be only one data point at each leaf node). You also need to label the nodes.

3) (**3 points**) Show the nodes that need to be checked in order to classify $(-3,-2)$ using the nearest neighbor rule. Note that in order to get full credit, you need to check all the nodes that are necessary to get the true nearest neighbor in the entire training set. Show your steps.

Problem 5 (13 points) A three-layer neural network (with 2 inputs, one hidden neuron, and one output neuron), whose transfer function for the hidden layer is:

$$y = w_{11}\exp\left\{-\frac{(x_1 + x_2 - b)^2}{2}\right\} + w_{12}\exp\left\{-\frac{(x_1 - x_2 - b)^2}{2}\right\} + w_{13}x_1 + w_{14}x_2 + w_{15}x_1x_2\,,$$

and the transfer function for the output layer is

$$z = \frac{1}{1+e^{w_2 y}}$$

where x_1 and x_2 are the inputs, and w_{11}, w_{12}, w_{13}, w_{14}, w_{15}, w_2, and b are network parameters.

1) (**6 points**) Find a learning rule for the network parameters if we use the training error $\frac{1}{2}(t-z)^2$ as the criterion function, where t is the target output. (Hint: using a gradient descent algorithm.)

2) (**4 points**) Find a learning rule for the network parameters if we use cross-entropy $t\ln\frac{t}{z}$ as the criterion function, where t is the target output.

3) (**3 points**) Can the given neural network be trained to solve the XOR problem (with ±1 as inputs and 0/1 as output; that is, XOR(-1, -1)=0, XOR(-1, 1)=1, XOR(1, -1)=1, and XOR(1, 1) = 0)? If yes, then give one solution (note that you do **NOT** have to obtain your solution by applying the backpropagation algorithm); if no, give a brief justification. Here an output from the neural network that is less than or equal to 0.1 is considered to be 0 and an output that is larger than or equal to 0.9 is considered to be 1.

End of Questions