# Practice problems for the video course on Pattern Recognition

**Bayes Classifier and related issues** (Modules 1-2)

1. Consider a 2-class PR problem with feature vectors in $\Re^2$. The class conditional density for class-I is uniform over $[1,\ 3] \times [1,\ 3]$ and that for class-II is uniform over $[2,\ 4] \times [2,\ 4]$. Suppose the prior probabilities are equal. Which is Bayes classifier? (For this problem, assume that we are using 0-1 loss function and hence Bayes classifier minimizes probability of misclassification). Is Bayes Classifier unique for this problem? What would be the probability of misclassification by Bayes Classifier? Consider a hyperplane given by $x + y = 5$ in $\Re^2$. Is this a Bayes (optimal) clasifier? Suppose the prior probabilities are changed to $p_1 = 0.4$ and $p_2 = 0.6$. What is a Bayes classifier now? How good would the earlier hyperplane be now?

2. Suppose that density for class-I is same as in the previous problem but the density for class-II is changed to uniform density over $[2,\ 8] \times [2,\ 8]$ and suppose that the prior probabilities are equal. What would be the Bayes classifier now? How good would the earlier hyperplane classifer be now?

3. Consider a 2-class PR problem with feature space $\Re$. Let $p_1$ and $p_2$ be the prior probabilities. Let the class conditional density for Class-1 be exponential with parameter $\lambda$ and that for Class-2 be normal with mean $\mu$ and variance $\sigma^2$. Derive the Bayes classifier for the 0–1 loss function. Specify any one special case when this Bayes classifier would be a linear discriminant function.

4. Consider a 2-class problem with $d$ Boolean features. Let $p_{ij} = \text{Prob}[X_i = 1 | X \in C_j]$, $i = 1, \cdots, d$ and $j = 0, 1$, where $X = [X_1\ \cdots\ X_d]^T$ is the feature vector and $C_0, C_1$ are the two classes. We assume that different features are stochastically independent. Suppose $p_{i0} = p$ and $p_{i1} = (1-p),\ i = 1, \cdots, d$ where $p > 0.5$. The prior probabilities of the

two classes are equal. Assume that $d$ is odd. Show that the minimum probability of error classifier is the decision rule: "Decide $X \in C_0$ if $\sum_{i=1}^{d} x_i > \frac{d}{2}$; else decide $X \in C_1$".

5. Consider a 2-class problem with feature space $\Re$ and equal prior probabilities. Let the class conditional densities be given by

$$f_i(x) = \frac{1}{\pi b} \frac{1}{1 + \left(\frac{x-a_i}{b}\right)^2} \quad i = 1, 2$$

where $a_1, a_2, b$ are parameters of the density functions. Assume $a_1 < a_2$. Find the Bayes classifier (under 0–1 loss function). Show that the minimum probability of error is given by

$$P(\text{error}) = \frac{1}{2} - \frac{1}{\pi} \tan^{-1} \left| \frac{a_2 - a_1}{2b} \right|$$

6. Consider a 2-class PR problem with feature space $\Re$. Let $p_1$ and $p_2$ be the prior probabilities and let the two class conditional densities be exponential with parameters $\lambda_1$ and $\lambda_2$ respectively. Derive the Bayes classifier for the 0–1 loss function. For $\lambda_1 = 2$, $\lambda_2 = 1$ and $p_1 = p_2$, derive an expression for the Bayes error.

7. Consider a 2-class PR problem with feature space $\Re$. Suppose the class conditional densities are normal with equal variances. Specify some conditions under which the Neymann-Pearson classifier would be same as the Bayes classifier under 0-1 loss function.

**Density Estimation** (Modules 3-5)

8. Suppose we have $n$ *iid* samples from a geometric distribution. Find the maximum likelihood estimator for the parameter $p$. (If $X$ is geometrically distributed, its probability mass function is: $f_X(x) = (1-p)^{x-1} p, \; x = 1, 2, \cdots$). For the same problem, suppose we want to use Bayesian estimation. What would be the conjugate prior? What is the MAP estimate for $p$ in this case?

9. Suppose a class conditional density is uniform over [0, $\theta$] with $\theta$ as the unknown parameter. We want Maximum Likelihood estimate of $\theta$ based on $n$ *iid* samples. (i). Suppose $n = 4$ and the four samples are: 0.3, 0.1, 0.7 and 0.5. Sketch the graph of the likelihood function $L(\theta|\mathcal{D})$ versus $\theta$ for the range $0 \leq \theta \leq 1$. (ii). Show that the maximum likelihood estimate of $\theta$ is given by $\hat{\theta} = \max_i\{X_i\}$.

10. We want to estimate the probability, $p$, of getting a head for a given coin. Our data is a sample of three tosses all of which have turned up heads. Choose a suitable conjugate prior and derive the Bayesian estimate for $p$. Explain how the parameters of the prior affect our final estimate for $p$. What would be the maximum likelihood estimate of $p$ given this data.

11. Let a class conditional density be a (one dimensional) normal distribution with mean $\mu_0$ and variance 10000. Suppose we assume a density model as a normal density with mean $\mu$ & variance 1, and then estimate $\mu$ from the data. Suppose we have a large amount of training data. Will the estimated $\mu$ be close to $\mu_0$? If we use the estimated density as the class conditional density, how good would be the performance of the classifier?

12. Consider a two class problem with one dimensional feature space. Suppose we have six training samples: $x_1, x_2, x_3$ from one class and $x_4, x_5, x_6$ from the other class. Suppose we want to estimate the class conditional densities nonparametrically using a kernel density estimate with Gaussian window with width parameter $\sigma$. Write an expression for the Bayes classifier (under 0–1 loss function) which uses these estimated densities.

13. Suppose we have $N$ data points each for the two classes. Consider the two scenarios. In one case we assume that the class conditional densities are normal, use maximum likelihood estimation to learn the densities and implement Bayes classifier with the learnt densities. In the other case, we use a non-parametric method using Gaussian window or kernel function to estimate the class conditional densities and then implement the Bayes classifier with these estimated densities. Compare the relative computational effort needed to classify a new pattern

using the classifiers in the two cases. (You can assume that the prior probabilities are equal and that we are using a 0-1 loss function).

14. Consider the parzen window estimate of a density given by

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n} \phi\left(\frac{x - x_i}{h_n}\right)$$

Let the function $\phi$ be given by $\phi(x) = \exp(-x)$ for $x > 0$ and it is zero for $x \leq 0$. Suppose the true density (from which samples are drawn) is uniform over $[0,\ a]$. Show that the expectation of the parzen window estimate is given by

$$E\hat{f}_n(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{a}\left(1 - e^{-x/h_n}\right) & \text{for } 0 \leq x \leq a \\ \frac{1}{a}\left(e^{a/h_n} - 1\right) e^{-x/h_n} & \text{for } x \geq a \end{cases}$$

Is this a good approximation to uniform density? Explain.

**Linear models** (Module 6)

15. Consider 2-class PR problems with $n$ Boolean features. Consider two specific classification tasks specified by the following: (i) a feature vector $X$ should be in Class-I if the integer represented by it is divisible by 4, otherwise it should be in Class-II; (ii) a feature vector $X$ should be in Class-I if it has odd number of 1's in it, otherwise it is in Class-II. In each of these two cases, state whether the classifier can be represented by a Perceptron; and, if so, show the Perceptron corresponding to it; if not, give reasons why it cannot be represented by a Perceptron.

16. Consider 2-class PR problems with feature vector in $\Re^d$. Consider the classification tasks specified by the following: (i) $X$ should be in class $C_0$ if all the components of $X$ are positive; otherwise it should be in class $C_1$; (ii). $X$ should be in class $C_0$ if a majority of the components of $X$ are positive, otherwise it should be in class $C_1$; (iii). $X$ should be in class $C_0$ if sum of the components of $X$ that are positive is more than the absolute value of the sum of components of $X$ that are negative;

otherwise it should be in class $C_1$. In each of these cases, state whether the classifier can be represented by a Perceptron; and, if so, show the Perceptron corresponding to it; if not, give reasons why it cannot be represented by a Perceptron.

17. Consider the incremental version of the Perceptron algorithm. If, at iteration $k$, we misclassified the next pattern then we correct the weight vector as: $W(k+1) = W(k)+X(k)$. However, this does not necessarily ensure that $W(k+1)$ will classify $X(k)$ correctly. Suppose we change the algorithm so that when we misclassify a pattern, we change the weight vector by an amount that ensures that after the correction the weight vector correctly classifies this pattern. Will this new version of the Perceptro algorithm also converge in finite iterations if the classes are linearly separable? Explain.

18. Consider a 3-class problem with a linear classifier specified through three discriminant functions: $g_i(X) = W_i^T X + w_{i0},\ i = 1, 2, 3$. The classifier will assign $X$ to class $i$ if $g_i(X) \geq g_j(X),\ \forall j$. Show that the decision regions of such a linear classifier are convex. Suppose $X \in \Re^2$ and $w_{i0} = 0, \forall i$. Draw a sketch showing the three vectors $W_i$ and the decision regions of the three classes.

19. Consider a classification problem with $K$ classes: $C_1, \cdots, C_K$. We say that a training set of examples is linearly separable if there are $K$ functions, $g_j(X) = W_j^T X + w_{j0},\ j = 1, 2, \cdots, K$ such that $g_i(X) > g_j(X),\ \forall j \neq i$ whenever the training pattern $X$ belongs to $C_i$. A training set of examples is said to be totally linearly separable if the examples of each class can be separted from all other examples by a hyperplane. Show that totally linearly separable implies linearly separable; but the converse need not be true.

20. Consider a two class problem where class conditional densities are normal with equal covariance matrices. Suppose we have a large amount of training data. Can you say something about the relationship between a classifier based on Fisher linear discriminant and the Bayes classifier that minimizes probability of misclassification.

21. Let $X \in \Re^d$ denote feature vector and let $y \in \{-1,\ +1\}$ denote class labels. Let $\Sigma$ denote the covariance matrix of $X$. Let $\mu_1 = E[X|y =$

$+1]$ and $\mu_2 = E[X|y = -1]$. Let $p_1 = P[y = +1]$ and $p_2 = P[y = -1]$. Suppose $p_1 = p_2$. Suppose we want to find $W \in \Re^d$ and $w_0 \in \Re$ to minimize the mean square error given by $E[W^T X + w_0 - y]^2$. Show that $W \propto \Sigma^{-1}(\mu_1 - \mu_2)$. Based on this, what can you say about the quality of a linear classifier in a 2-class problem learnt through least squares approach if the class conditional densities are normal with equal covariance matrices.

22. Consider a one dimensional regression problem with $X \in \Re$ as the feature and $Y \in \Re$ as the target. Suppose that the joint distribution of $X, Y$ is Gaussian. Show that optimal regression function (which minimizes mean squared error) would be linear.

**Statistical Learning Theory** (Module 7)

23. Consider the following 'guess-the-number' game. The teacher picks some number, $c^*$, from the interval $[0, 1]$. The learner is given a set of examples of the form $\{(x_i, y_i), i = 1, \ldots, n\}$ where $x_i \in [0, 1]$ and $y_i = 1$ if $x_i \leq c^*$ and $y_i = 0$ if $x_i > c^*$. Suggest a PAC learning algorithm for this problem. That is, give an algorithm that takes as input $n$ examples and outputs a number $c$ in the interval $[0, 1]$; and then show that given any $\epsilon, \delta$, the difference between $c$ and $c^*$ would be less than $\epsilon$ with a probability greater than $(1 - \delta)$ if $n$ is sufficiently large.

24. Consider a variation of the game in the previous problem where, instead of being given the examples, the learner can choose his examples. That is, at each instant $i$, the learner can choose any $x_i \in [0, 1]$ and the teacher would respond with a $y_i$ which is 1 or 0 based on whether or not $x_i \leq c^*$. The learner can use all the information he has till $i$ to decide on $x_i$. Suggest a strategy for the learner to choose his examples and to learn $c^*$. What can you say about the number of examples needed (for a given accuracy of learning) in this case in comparison to the case in the previous problem?

25. A monomial over Boolean variables is a conjunction of literals. A literal is a variable or its compliment. For example, $x_1$, $x_2 x_3$, $\bar{x}_1 x_2 x_3$ are all monomials over three Boolean variables. Here, $\bar{x}_1$ denotes the literal

which is the compliment of $x_1$. (It may be noted that the total number of monomials over $n$ variables is $3^n$). Consider a 2-class problem with $n$ Boolean features. Suppose we know that all patterns can be correctly classified by some monomial. (That is, the correct monomial would have value 1 on all feature vectors of $C_0$ and would have value 0 on all feature vectors from $C_1$). We want to learn the monomial given some examples. Consider a learning algorithm for this as given below. We start with the monomial $x_1\bar{x}_1x_2\bar{x}_2\ldots x_n\bar{x}_n$. (Note that this monomial classifies all patterns as $C_1$). At each iteration we modify the current monomial as follows. If the next example is from $C_1$ we do nothing. If the next example is from $C_0$, then, for each $i$, $1 \leq i \leq n$, if the example has value 1 for $i^{th}$ feature, then we delete the literal $\bar{x}_i$ (if present) from the current monomial; if the example has value 0 for $i^{th}$ feature, then we delete the literal $x_i$ (if present) from the current monomial. Show that this is a PAC-learning algorithm. That is, show that given any $\epsilon$ and $\delta$, we can find $n$ such that after $n$ random examples, the probability that the error of the classifier learnt by the algorithm is greater than $\epsilon$ is less than $\delta$.

**Nonlinear Classifiers** (Modules 8-9)

26. Suppose we have a 3-class classification problem. Consider two different architectures of single hidden layer feedforward networks with sigmoidal activation functions. Both have the same number of input and hidden nodes. In the first architecture we have only one output node and we are going to use values 0.1, 0.5 and 0.9 as the desired outputs for the three classes. In the second architecture we will have three output nodes and are going to use the three unit vectors (that is, vectors like $[1\ 0\ 0]^T$ etc) as the desired outputs for the three classes. Is one of these architectures better than the other from the point of view of approximating the Bayes classifier? If so, is there any 'extra cost' in using this better architecture?

27. Consider a 3-layer feedforward network with sigmoidal activation function. Given any training set, is it always possible to achieve zero error on training data by having sufficiently large number of hidden nodes?

Is it possible to achieve zero training error if we use RBF network instead? Explain.

28. Consider a specific 3-layer feedforward network with sigmoidal activation functions for all hidden nodes. Can we construct another 3-layer feedforward network (with same architecture) where hidden units use hyperbolic tangent as the activation function such that the two networks compute the same function?

29. For a linear SVM, let $W^*, b^*$ be the optimal hyperplane and let $\mu_i^*$ be the optimal Lagrange multipliers. Show that

$$(W^*)^T W^* = \sum_i \mu_i^*.$$

30. Suppose we are using a SVM with slack variables. (That is the one with primal objective function $0.5\, W^T W + C \sum \xi$). Suppose the given training data is linearly separable. Would the SVM always output a separating hyperplane?

31. Consider a pattern recognition problem in $\Re^2$, for an SVM, with the following training samples:

    Class +1: $(0.5, 0.5)$, $(0, 2)$, $(0, 1)$, $(2, 2)$
    Class -1: $(0, 0)$

    (a) By solving the (primal) optimization problem, show that the optimal hyperplane is given by $W^* = [2\ 2]^T$ and $b^* = -1$.

    (b) Suppose we want to solve this problem using the slack variables, $\xi_i$. That is, we want to maximize $0.5 W^T W + C \sum \xi_i$ subject to constraints $1 - y_i[W^T x_i + b] - \xi_i \leq 0$ and $\xi_i \geq 0$. Take $C = 1$. For the $W, b$ obtained in part (a), find the smallest values for $\xi_i$ so that all constraints are satisfied. At these $W, b, \xi_i$ what is the value of the objective function being minimized. Now consider another hyperplane given by $W = [0\ 1]^T$ and $b = -1$. (That is, the hyperplane is a line parallel to $x$-axis and passing through (0,1)). For these $W, b$ find the smallest possible values for $\xi_i$ so that all constraints are satisfied. What is the value of the objective

function at these $W, b, \xi_i$. Based on all this, can you say whether the optimal separating hyperplane found in part (a) would also be the optimal solution to the current optimization problem (which includes the slack variables $\xi_i$).

(c) Can you suggest values for $C$ so that the optimal separating hyperplane found in part (a) would be the optimal solution to the optimization problem with the slack variables. Explain.