

E1 213 Pattern Recognition and Neural networks
Test I

Time: 90 minutes
20 feb 2017

Answer ALL Questions

1.
 - a. Consider a 2-class problem in \mathbb{R} . The density of class-0 patterns is exponential with parameter λ and that of class-1 is normal with mean μ and variance σ^2 . The prior probabilities are p_0 and p_1 . Find the Bayes classifier (under 0-1 loss function). Specify any special case under which the Bayes classifier is linear.
 - b. Consider a 2-class problem with $X \in \mathbb{R}$. Suppose both class conditional densities are normal. Specify a special case when the Bayes classifier (under 0-1 loss), the Neymann-Pearson classifier and the Minmax classifier would all be the same.
2.
 - a. Suppose X has a density given by $f(x) = \frac{1}{(n-1)!} (\lambda x)^{n-1} \lambda e^{-\lambda x}$, $x > 0$, where $\lambda, n > 0$ are parameters. Assume we know n and want to estimate λ . Derive the ML estimate for λ based on N iid samples.
 - b. Suppose X is exponential with parameter λ . Derive the MAP estimate of λ based on n iid samples.
3.
 - a. Consider a 3-class problem and a linear classifier specified by three functions $g_j(X) = W_j^T X + w_{j0}$, $j = 1, 2, 3$. Recall that such a classifier classifies X into class- j if $g_j(X) \geq g_i(X), \forall i$. Show that the decision regions of such a classifier are convex. Now take the special case of $X \in \mathbb{R}^2$. Suppose $w_{j0} = 0, \forall j$. Take any three vectors of your choice as W_j , $j = 1, 2, 3$ and sketch the regions of the three classes under this classifier.
 - b. In the Perceptron algorithm (where we assume that all the patterns of negative class are multiplied by -1), we correct the weight vector as $W(k+1) = W(k) + X(k)$ if $W(k)^T X(k) \leq 0$. Suppose we make the correction whenever $W(k)^T X(k) \leq b$ where $b > 0$ is a user-chosen parameter. Will such an algorithm also converge (may be under some special conditions)? Explain. If it converges, what can you say about the hyperplane found.

✓
EI 214 Pattern Recognition and Neural Networks
Test 1

Time: 90 min

24 Feb 2010

Answer ALL questions

① → ①. Consider a 2-class PR problem with feature vectors in \mathbb{R}^2 . The class conditional density for class-0 is uniform over $[1, 3] \times [1, 3]$ and that for class-1 is uniform over $[2, 2+a] \times [2, 2+a]$, where $a > 1$ is some real number. Suppose the prior probabilities are equal. What is Bayes classifier (that minimizes risk under the 0-1 loss function) for different values of a ? What would be the probability of misclassification by Bayes Classifier? Consider a hyperplane given by $x + y = 5$ in \mathbb{R}^2 . Is this a Bayes classifier for some value of a ? Now suppose the prior probabilities of the two classes are not equal and are given by $p_0 = 0.6$ and $p_1 = 0.4$. What would be the Bayes classifier now?

② → ②. a. We have iid data, $X_i, i = 1, 2, \dots, N$, drawn from a Poisson distribution with parameter λ . We want to estimate λ from the data. Derive the maximum likelihood estimate for λ .

③ → ③. b. Let a class conditional density be a (one dimensional) normal distribution with mean μ_0 and variance 10000. Suppose we assume a density model as a normal density with mean μ & variance 1, and then estimate μ from the data. Suppose we have a large amount of training data. Will the estimated μ be close to μ_0 ? If we use the estimated density as the class conditional density, how good would be the performance of the classifier?

$\mu(\mu/\sqrt{10000})$
 $\sim N(\mu_0, 1)$
 $f(\mu) \sim N(\mu_0, 1)$

c. 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)

③ Given a set of vectors, $\{X_1, X_2, \dots, X_m\}$ in \mathbb{R}^n , their convex hull is defined as the set of all $X \in \mathbb{R}^n$ such that $X = \sum_{j=1}^m a_j X_j$ with $a_j \geq 0, \forall j$ and $\sum_{j=1}^m a_j = 1$. Show that if two sets of vectors in \mathbb{R}^n are linearly separable then their convex hulls do not intersect.

④ h) Suppose there is a scalar signal $x(n)$, $n = 1, 2, \dots$. You want to predict the value of $x(n)$ as a linear function of the past m values of the signal. Explain how you can learn this predictor using the LMS algorithm.

④ State whether the following statements are true or false and give a short explanation for your answer.

⑤ a) If the Bayes optimal classifier is linear then the classes are linearly separable. *not possible* *same as $\sigma_1 \neq \sigma_2$ for lin sep*

⑤ b) Consider a 2-class problem with one dimensional feature vector where the class conditional densities are normal with equal variances. The Bayes decision point and the Neyman-Pearson decision points will always be different.

⑥ c) Suppose random variables Y and X are related by $Y = aX + b + \xi$ where ξ is a zero mean independent noise with variance σ^2 . If we have a very large training set and learn Y as a function of X using linear least squares regression method, then the final squared error on the training set may be far from zero.

⑦ d) In the Bayesian method of estimating parameters of a density function, if we choose the mode of the posterior density (of the parameter given the data) as our estimate then it will always be the same as the maximum likelihood estimate.

⑧ e) In a 2-class problem, if we want to estimate the prior probabilities of the two classes from the training set in a Bayesian framework, then the appropriate conjugate prior density in Bayesian estimation would be a uniform density.

Time: 90 min

25 Feb 2011

Answer ALL questions

1. Consider a 2-class PR problem with feature space \mathbb{R} . Let p_1 and p_2 be the prior probabilities and let the two class conditional densities be exponential with parameters λ_1 and λ_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.

2. Consider the following set of training patterns in \mathbb{R}^3 for a 2-class PR problem: $[0.8, 0.9, 0.1]$: Class I, and $[0.3, 0.2, 0.3]$: Class II. Show how the weight vector gets updated in the perceptron algorithm if you successively present each of the two patterns once. Take all components of the initial weight vector to be 1.

3. Consider Bayes classifier that achieves minimum risk. Let $L(i, j)$ denote the loss when classifier decision is i and the true class of the pattern is j . Suppose we have K classes. Consider a classifier that is allowed the option to 'reject' a pattern which will be done by the classifier assigning class $K + 1$ to the pattern. Define the loss function by

$$\begin{aligned} L(i, j) &= 0 \text{ if } i = j \text{ and } i, j = 1, \dots, K \\ &= \rho_m \text{ if } i = 1, \dots, K, \text{ and } i \neq j \\ &= \rho_r \text{ if } i = K + 1 \end{aligned}$$

Show that the optimal decision is: Decide on class i if (i). $q_i(X) \geq q_j(X), \forall j \neq i$, and (ii). $q_i(X) \geq 1 - (\rho_r / \rho_m)$; otherwise decide on class $K + 1$ (i.e., 'reject'). (Here, q_i is the posterior probability function for class i). Explain what this optimal decision means in the special cases of (a). $\rho_r = 0$, and (b). $\rho_r > \rho_m$.

3. Suppose we have n iid samples from a geometric distribution. Find the maximum likelihood estimator for the parameter p .

$$p(1-p)^{x_i-1} \quad L = p^n (1-p)^{\sum x_i - n}$$

$$\ln L = np + (\sum x_i - n) \ln(1-p)$$


$$\ln L = \frac{n}{p} + \frac{(\sum x_i - n)}{1-p}$$

$$\frac{1-p}{p} = \frac{\sum x_i}{n} - 1$$

$$\frac{1}{p} = \frac{\sum x_i}{n}$$

$$p = \sum x_i / n$$

✓ 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 g . Write an expression for the Bayes classifier (under 0-1 loss function) which uses these estimated densities.

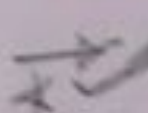
3)  Consider a two class problem where class conditional densities are normal with equal covariance matrices. Suppose we have a large amount of training data. Can you say something about the relationship between a classifier based on Fisher linear discriminant and the Bayes classifier that minimizes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

4. a. Briefly explain the following:

- (i). Least squares method for estimating a linear function.
- (ii). VC Dimension of a family of classifiers.

b. State whether or not each of the following statements are true

✓ In the Bayesian method of estimating parameters of a density function, if we choose the mode of the posterior density (of the parameter given the data) as our estimate then it will always be the same as the maximum likelihood estimate.

✗  In a 2-class problem, if we want to estimate the prior probabilities of the two classes, the maximum likelihood estimate is better than the Bayesian estimate, because the Bayesian estimate would be a uniform density.

100 ✓ Suppose we have n samples from a Gaussian density with mean μ_0 and variance 1000. We assume a density model that is Gaussian with mean μ and variance 1. We estimate μ from the data using maximum likelihood method. Then, no matter how much data we have, due to the mismatch between the true density and assumed density model, the estimated μ would not be close to μ_0 .

✓ 4. If the Bayes optimal classifier is linear then the classes are linearly separable.

✓ 5. If a set with m points is shattered by a family of classifiers then any set with $(m - 1)$ points would also be shattered.

EL 214 Pattern Recognition and Neural Networks
Test I.

Time: 90 min

17 Feb 2012

2000

Answer ALL questions

1. a. Consider a 2-class PR problem with feature space \mathcal{R} . Let p_1 and p_2 be the prior probabilities. Let the class conditional density for Class-1 be exponential with parameter λ and that for Class-2 be normal with mean μ and variance σ^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.

Bayes NET

- b. Consider a 2-class PR problem with feature space \mathcal{R} . 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.

Does not give a solid answer for μ, σ same

2

- a. Consider the following set of training patterns in \mathbb{R}^3 for a 2-class PR problem: $[0.6, 0.5, 0.1]$: Class I, and $[0.7, 0.1, 0.3]$: Class II. Show how the weight vector gets updated in the perceptron algorithm if you successively present each of the two patterns once. Take all components of the initial weight vector to be -0.1 .

Bayesian estimation Computational complexity

- b. Consider a two class problem with one dimensional feature space. Suppose we have n_1 training samples from one class and n_2 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 σ . Write an expression for the Bayes classifier (under 0-1 loss function) which uses these estimated densities. Compare the computational complexity of this classifier with the Bayes classifier implemented with parametric density estimation assuming normal class conditional densities.

2nd order No eval of ex

3

- a. Suppose we have n iid samples from a geometric distribution. Find the maximum likelihood estimator for the parameter p .

Repeat 2011

- b. For the same problem as in (a.), suppose we want to use Bayesian estimation. What would be the conjugate prior? What is the MAP estimate for p in this case?



$$\begin{aligned}
 & (a-1)p^{a-2} \cdot (1-p)^{b-1} \\
 & - p^{a-1} \cdot (b-1)(1-p)^{b-2} = 0 \\
 & a-1 \cdot \frac{p^{a-1}}{p} = (b-1) \cdot \frac{(1-p)^{b-1}}{1-p} \\
 & \frac{a-1}{p} = \frac{b-1}{1-p} \\
 & \frac{1-p}{p} = \frac{a-b+1}{a-1} \Rightarrow \frac{1}{p} = \frac{a+b-1}{a-1} \\
 & \frac{1}{p} = \frac{a+b-1}{a-1}
 \end{aligned}$$

* Suppose there is a scalar signal $x(n)$, $n = 1, 2, \dots$. You want to predict the value of $x(n)$ as a linear function of the past m values of the signal. Briefly explain how you can learn this predictor using linear least squares method.

✓ Answer the following briefly. Each of your answers should not exceed five lines.

mode of $q(\theta|D)$
 $\hat{\theta} = \arg \max_{\theta} q(\theta|D)$

$$q(\theta|D) = \prod_{i=1}^N q(\theta|x_i)$$

$$q(\theta|D) = \frac{L(\theta|D)}{Z}$$

$$= \frac{L(\theta|D)}{\int L(\theta|D) d\theta}$$

mode is $\hat{\theta}$ that gives max. value for $q(\theta|D)$

$L(\theta|D)$ & $q(\theta|D)$ have same max. Depends on $p(\theta)$

i. In the Bayesian method of estimating parameters of a density function, if we choose the mode of the posterior density (of the parameter given the data) as our estimate, will it always be the same as the maximum likelihood estimate?

Consider a 2-class problem with both class conditional densities being normal with means 0 & 3 and variances 1 & 100. Suppose we assume that both classes have normal density with variance 1 and estimate the means from the data using maximum likelihood method. Would the estimate of the means be accurate? Would the Bayes classifier implemented with these estimated densities be accurate?
 No - not at all so full marks

iii. If, in a 2-class problem, the Bayes optimal classifier is linear then would the classes be linearly separable?

iv. In Bayesian Estimation, what is a conjugate prior?

v. Explain what is a sufficient statistic.

$$x(\theta|D) = \frac{L(\theta|D)p(\theta)}{q(\theta|D)}$$

$$= \frac{L(\theta|D)p(\theta)}{\int L(\theta|D)p(\theta) d\theta}$$

→ in estimate or mean $\hat{\theta} = \frac{1}{N} \sum x_i$
 stage 1: $\hat{\theta} = \frac{1}{N} \sum x_i$
 stage 2: $\hat{\theta} = \frac{1}{N} \sum x_i$
 prior density which has v.l.s in some form of posterior

El 213 Pattern Recognition and Neural Networks
Test 1

Time: 90 min

27 Feb 2015

Answer ALL questions

- ✓ a. Consider a 2-class problem with feature space \mathbb{R} 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

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

- b. In the perceptron algorithm, when the current hyperplane misclassifies the next pattern, we use that pattern to update the hyperplane. However, it is not guaranteed that the updated hyperplane correctly classifies the current pattern. Modify the perceptron algorithm by introducing a step size (which can be a function of the current pattern) so that we can ensure that the updated hyperplane would correctly classify this pattern. What can you say about the convergence of this modified perceptron algorithm?

2. a. Define an unbiased estimator. Consider the density model $f(z|\theta)$ which is uniform over $[0, \theta]$. Is the maximum likelihood estimate for θ unbiased?
- b. Let X be geometric with parameter p . (That is, the mass function is $f_X(x) = (1-p)^{x-1}p, x=1, 2, \dots$). Suppose we want Bayesian estimate for p . What would be the conjugate prior? Calculate the posterior density for p given n iid data. What is the MAP estimate? What will this reduce to if the prior density is uniform?

dec bound

$$(x-a_1) = \pm (x-a_2)$$

$$x-a_1 = -x+a_2$$

$$x = \frac{a_1+a_2}{2}$$

$$\theta = \max x_i$$

$$Y = \max(x_i)$$

$$F_Y(y) = P(Y \leq y) = P(x_1 \leq y, \dots, x_n \leq y) = P(x_1 \leq y)^n = \left(\frac{y}{A}\right)^n$$

$$f_Y(y) = F_Y'(y) = \frac{n y^{n-1}}{A^n}$$

$$E(\max(x_i)) = E(Y) = \int_0^A y \cdot \frac{n}{A^n} y^{n-1} dy = \frac{n}{A^n} \frac{y^{n+1}}{n+1} \Big|_0^A = \frac{n}{n+1} A < A$$

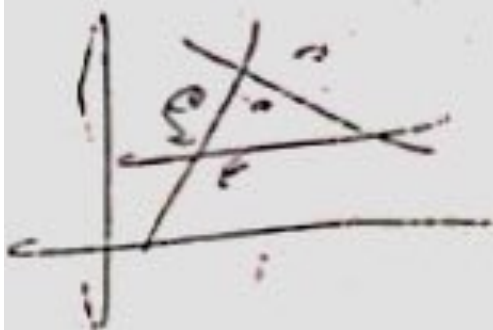
not unbiased.

3. a. Explain how one can view the linear least squares solution as a maximum likelihood estimate under a reasonable probability model.

b. Suppose we have a data set of $2n$ points with X_1, \dots, X_n being from class C_0 and X_{n+1}, \dots, X_{2n} from class C_1 . Suppose we use kernel density estimate with Gaussian kernel to estimate the class conditional densities. Write the expression for the Bayes classifier using these estimated densities. (Assume equal priors and 0-1 loss). Instead of this, suppose we assume the two class conditional densities to be Gaussian and estimate them using Maximum likelihood method. Compare the amount of computation needed to classify a new feature vector in these two methods.

4. a. Consider a problem with M classes. We say a given training set is linearly separable if there exist M linear discriminant functions, $g_j, j = 1, \dots, M$, such that $g_j(x) > g_i(x), \forall i \neq j$, whenever, $x \in C_j$. We say the training set is totally linearly separable if patterns of any one class are linearly separable from patterns of all other classes. Show that total linear separability implies linear separability but the converse is not true.

- b. Consider a 2-class problem with feature space being \mathbb{R} . Suppose the two class conditional densities are normal with same variance, σ^2 , and means μ_1 and μ_2 . Given any threshold, τ , let $a = P[X > \tau | X \in C_1]$ and $b = P[X > \tau | X \in C_2]$. Express the discriminability, $d = |\mu_1 - \mu_2|/\sigma$, in terms of a and b . Consider two cases: Case-1: $a = 0.8, b = 0.3$, Case-2: $a = 0.3, b = 0.7$. Which case has higher discriminability?



Machine Learning

Second Midterm

Answer all questions

Answer all questions Time: 90 minutes

19th Mar, 2014

1. Let $D = \{(x_i, y_i) | i \in [N]\}$ be a dataset containing N observation, label pairs with observations denoted by $x_i \in \mathbb{R}^d$ and labels denoted by $y_i \in \{-1, 1\}$.
 - (a) State the ν -SVM formulation for learning a linear classifier on D 1 marks
 - (b) Derive the dual 3 marks
 - (c) Define margin-errors. Derive a relationship between margin-errors, number of Support vectors, and ν . 4 marks
 - (d) Can we use $\nu = \frac{1}{2}$? Give reasons 2 marks
2. Consider a binary classification task with a reject option, in which a classifier can choose to reject instances about which it is unsure, with a corresponding cost of $c \in (0, \frac{1}{2})$; misclassifications incur a cost of 1 as usual. This can be formulated as a learning task with instance space \mathcal{X} , label space $\mathcal{Y} = \{\pm 1\}$, prediction space $\hat{\mathcal{Y}} = \{-1, +1, \text{reject}\}$, and loss $\ell_{\text{reject}(c)} : \{\pm 1\} \times \{-1, +1, \text{reject}\} \rightarrow [0, 1, c]$ defined as

$$\ell_{\text{reject}(c)}(y, \hat{y}) = \begin{cases} c & \text{if } \hat{y} = \text{reject} \\ 1 & \text{if } \hat{y} \in \{\pm 1\} \text{ and } \hat{y} \neq y \\ 0 & \text{if } \hat{y} = y. \end{cases}$$

Let D be a joint probability distribution on $\mathcal{X} \times \{\pm 1\}$, with marginal distribution μ on \mathcal{X} and conditional label probabilities given by $\eta(x) = P(y = 1|x)$, and for any classifier $h : \mathcal{X} \rightarrow \{-1, +1, \text{reject}\}$, define

$$\text{er}_D^{\text{reject}(c)}[h] = \mathbb{E}_{(x,y) \sim D} [\ell_{\text{reject}(c)}(y, h(x))].$$

Derive a Bayes optimal classifier in this setting, i.e. a classifier $h^* : \mathcal{X} \rightarrow \{-1, +1, \text{reject}\}$ with

$$\text{er}_D^{\text{reject}(c)}[h^*] = \inf_{h : \mathcal{X} \rightarrow \{-1, +1, \text{reject}\}} \text{er}_D^{\text{reject}(c)}[h].$$

What happens if $c \geq 0.5$.

10 marks

3. Let $\mathcal{X} \subset \mathbb{R}^d$ be a vector space.

(a) Define a kernel function

1 mark

- (b) Using the definition of kernel function: deduce that for any $x, z \in \mathcal{X}$, $K(x, z) = e^{-\|x\|^2 - \|z\|^2}$ is a kernel function. 2marks
- (c) Using the above prove that $e^{-t\|x-z\|^2}$ is a kernel function for any $t > 0$. 2 marks
- (d) For any normalized kernel function, $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ prove that $|k(x, z)| \leq 1$ for any $x, z \in \mathcal{X}$. 5 marks
4. Let $f(x) = w^T x + b$ be a model described by the parameters $w \in \mathbb{R}^d, b \in \mathbb{R}$. Let us fit the model to a dataset $D = \{(x_i, y_i) | i \in [N]\}$ by least squares where $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$. Show that

$$SS_{tot} = SS_{reg} + SS_{res}$$

$$SS_{tot} = \sum_{i=1}^N (y_i - \bar{y})^2, SS_{exp} = \sum_{i=1}^N (\bar{y} - f(x_i))^2, SS_{res} = \sum_{i=1}^N (f(x_i) - y_i)^2$$

where $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$.

10 marks

Note: The implication of this is total variation in the response variable y , measured by SS_{tot} , is the sum total of variance explained by the regression, SS_{reg} , and residual variance, SS_{res} . This motivates an often used measure of fit to data popularly known as $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$.

5. Let $f(x), D$ be defined as in Question 4. Let $y = f(x) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$. Let $w \sim N(0, \Lambda)$ and $b \sim N(0, 1)$. Let

$$S = \frac{1}{N} \sum_{i=1}^N x_i x_i^T, c = \sum_{i=1}^N y_i x_i$$

Derive the MAP estimate of w, b in terms of σ^2, S, c, Λ .

10 marks