E1 213 Pattern Recognition and Neural Networks

Practice Problems: Set 1

1. Consider a 2-class problem with one dimensional feature space. Let the class conditional densities be: $f_0(x) = e^{-x},\ x > 0$, and $f_1(x) = 1/2a,\ x \in [-a,\ a],\ a > 0$. The prior probabilities are equal. Assume we are using 0–1 loss. Find the Bayes classifier. For the case when $a = 0.25$, find Bayes error.

Answer: Since we are using 0–1 loss and since prior probabilities are equal, the Bayes classifier is: $h_B(x) = 0$ if $f_0(x) > f_1(x)$ and $h_B(x) = 1$ otherwise.

Since $f_0(x) = 0$ for $x < 0$, we have $h_B(x) = 1$ for $x < 0$. (Actaully, we need not worry about the region $x < -a$ because a pattern coming from that region has probability zero. But since we want to think of h_B as a function on $\Re$, we can assign class-1 in that region).

Now consider the region $x \geq 0$. If $a \leq 0.5$ then $(1/2a) \geq 1$ and hence $f_1(x) \geq f_0(x)$ for $0 \leq x \leq a$ and there after $f_1(x) = 0$. Hence till a we classify into class-1 and beyond that we classify into class-0.

If $a > 0.5$ then $f_0(x) > f_1(x)$ if $e^{-x} > (1/2a)$. This is true if $x < \ln(2a)$.

Putting all this together, the Bayes classifier for this problem is the following:

- If $a \leq 0.5$ then
$$h_B(x) = \begin{cases} 1 & \text{if } x \leq a \\ 0 & \text{if } x > a \end{cases}$$
- If $a > 0.5$ then
$$h_B(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ 0 & \text{if } 0 < x < \ln(2a) \\ 1 & \text{if } \ln(2a) \leq x \leq a \\ 0 & \text{if } x > a \end{cases}$$

For $a = 0.25$, we classify into class-1 till 0.25 and into class-0 after that. Hence, we make an error if we get a class-0 pattern with $x \leq 0.25$. Hence, bayes error is

$$0.5 \int_0^{0.25} e^{-x}\, dx = 0.5(1 - e^{-0.25}) \approx 0.11$$

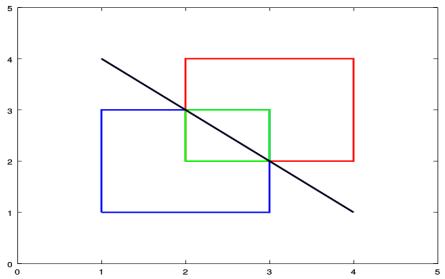
(Please plot f_1 and f_0 for two different values of a so that you would see all this more clearly).

2. 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 and we are using 0–1 loss. Consider line given by $x + y = 5$ in $\Re^2$. Is this a Bayes Classifier for this problem? Is Bayes Classifier unique for this problem? If not, can you specify two different Bayes classifiers? Suppose the class conditional densities are changed so that the density for class-I is still uniform over $[1,3] \times [1,3]$ but that for class-II is uniform over $[2,5] \times [2,5]$. Is the line $x + y = 5$ a Bayes classifier now? If not, specify a Bayes classifier now. Is the Bayes classifier unique now? For this case of class conditional densities, suppose that wrongly classifying a pattern into class-I is 10 times more expensive than wrongly classifying a pattern into class-II. Now, what would be a Bayes classifier?

Answer: Consider the first case. With equal priors and 0–1 loss, the decision of Bayes classifier is simply based on which class conditional density has a higher value.

It is easy to see that in $([1,3] \times [1,3]) - ([2,3] \times [2,3])$ Bayes classifier would assign Class-I because the other class conditional density is zero. Similarly in $([2,4] \times [2,4]) - ([2,3] \times [2,3])$ the decison is Class-II.

The situation is as shown in the figure below.



The only thing remaining to be decided is what to do in the overlapping region shown as a green rectangle in the figure. In this regions, both class conditional densities have the same value and hence it does not matter which class you assign. (This is like having an arbitrary rule to 'break ties' while deriving Bayes rule).

Thus, the Bayes classifier here is not unique. For example, we can assign all points in the green rectangle in the figure to Class-I or Class-II thus giving two different Bayes classifiers.

The line $x + y = 5$ shown in the figure is also a Bayes classifier here. It assigns half the points in the green rectangle to one class and the other half to the other class.

Now consider the case where class-II is uniform over $[2, 5] \times [2, 5]$. Now at all these points the value of class conditional density is $1/9$. Hence, in the common region now, the density of class-I has higher value (namely, $(1/4)$). Hence, all points in the common region have to be assigned to class-I. Thus, the line is no longer a Bayes classifier. Also, the Bayes classifier is unique here (assuming the domain, $\mathcal{X}$, to be $([1, 3] \times [1, 3]) \cup ([2, 5] \times [2, 5])$).

Now consider the case where we are not using 0–1 loss and wrogly classifying into class-I is 10 times costlier than wrongly classifying into

class-II. Since priors are equal, this means we will put something into class-I only if the value (at that point) of $f_I(x)$ is 10 times that of $f_{II}(x)$. In the common region, the ratio is (9/4). Hence, now the Bayes classifier would put all points in the common region in Class-II.

3. Consider a general K-class problem with a general loss function. Let $h(X)$ denote the output of the classifier on X. Let $R(\alpha_i|X)$ denote the expected loss when classifier says α_i and conditioned on X. That is, $R(\alpha_i|X) = E[L(h(X), y(X))|h(X) = \alpha_i, X]$, where, as usual, $y(X)$ denotes the 'true class'. We had only considered deterministic classifiers where h is a function that assigns a unique class label for any given X. Suppose we use a stochastic classifier, h, which, given X, outputs α_i with probability $p_h(\alpha_i|X)$. (Note that we would have $p_h(\alpha_i|X) \geq 0$ and $\sum_i p_h(\alpha_i|X) = 1$). For this classifier, show that the risk is given by

$$R(h) = \int \left[\sum_{i=1}^{K} R(\alpha_i|X) p_h(\alpha_i|X) \right] \, f(X) \, dX$$

where $f(X)$ is the density of X. Using the above expression, find the best choice of values for all the $p_h(\alpha_i|X)$ and hence conclude that we do not gain anything by making the classifier stochastic.

Answer

Using the same notation as in class, we have

$$\begin{aligned}
R(h) &= E[E[L(h(X), y(X)) \mid X]] \\
&= E[\sum_{i=1}^{K} (E[L(h(X), y(X)) \mid h(X) = \alpha_i, X] \Pr[h(X) = \alpha_i \mid X]] \\
&= E[\sum_{i=1}^{K} R(\alpha_i|X) p_h(\alpha_i|X)] \\
&= \int \left[\sum_{i=1}^{K} R(\alpha_i \mid X) p_h(\alpha_i|X) \right] \, f(X) \, dX
\end{aligned}$$

Suppose for some X, we have $R(\alpha_i \mid X) \leq R(\alpha_j \mid X)$, $\forall j$. Since p_h is a probability mass function, for any classifier h, we have

$$\min_j R(\alpha_j \mid X) \leq \sum_{i=1}^{K} R(\alpha_i \mid X) p_h(\alpha_i|X)$$

Let $h2$ be any stochastic classifier with $0 < p_{h2}(\alpha_i \mid X) < 1$. Suppose $h1$ be a dterministic classifier with $h1(X) = \alpha_i$. Then we have

$$R(h1(X) \mid X) = R(\alpha_i \mid X) < \sum_{i=1}^{K} R(\alpha_i \mid X) p_{h2}(\alpha_i | X)$$

Thus, at every X a deterministic classifier has lower risk than a stochastic classifier, proving that Bayes classifier would be a deterministic classifier.

4. Let $x_1, \cdots, x_n$ be *iid* data drawn according to exponential density with parameter λ. Derive the ML estimate for λ. (The exponential density is given by $f(x) = \lambda e^{-\lambda x},\ x > 0$).

Answer: The density model is

$$f(x \mid \lambda) = \lambda \exp(-\lambda x),\ x \geq 0$$

The likelihood function is

$$L(\lambda \mid \mathcal{D}) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i)$$

The log likelihood function is

$$l(\lambda \mid \mathcal{D}) = \sum_{i=1}^{n} (\ln(\lambda) - \lambda x_i)$$

Differentiating w.r.t. λ and equating to zero, we get

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

This gives us the final ML estimate as

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

5. Suppose X is uniformly distributed over $[0,\ \theta]$, with $\theta > 0$ being the unknown parameter. (The uniform density is given by $f(x) = 1/\theta$, if $0 \leq x \leq \theta$ and $f(x) = 0$ otherwise). Suppose we have three *iid* samples,

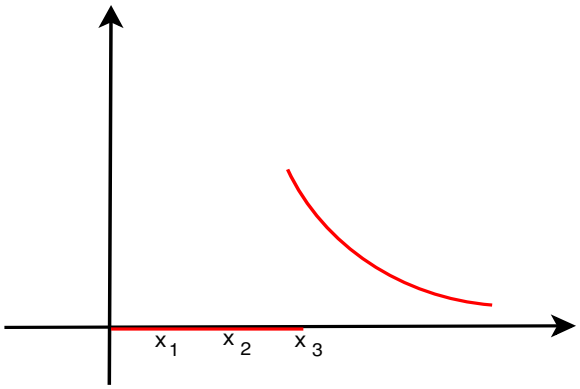


Figure 1: Schematic plot of likelihood function value, $L(\theta)$, on y-axis versus θ on x-axis (taking $\theta > 0$).

$1.75, 0.5, 2.2$. What is the value of the likelihood function $L(\theta|\mathcal{D})$ for (i). $\theta = 10$, (ii). $\theta = 1.9$? Now consider the general case where we represent the three *iid* samples as x_1, x_2, x_3. Plot the likelihood function (that is, plot $L(\theta|\mathcal{D})$ versus θ). Now, consider the case where we have n *iid* samples, what is the ML estimate for θ.

Answer: Given the three data points, the likelihood function is given by

$$L(\theta \mid \mathcal{D}) = f_\theta(1.75) f_\theta(0.5) f_\theta(2.2)$$

This is a function of θ. Note that the density model, f_θ is given by $f_\theta(x) = (1/\theta)$ if $0 \leq x \leq \theta$ and $f_\theta(x) = 0$ if $x > \theta$.

So, $L(10 \mid \mathcal{D})$ is $1/1000$, because each factor is $1/10$ because $\theta = 10$.

For $\theta = 1.9$, note that $f_\theta(2.2) = 0$. Hence, $L(1.9 \mid \mathcal{D})$ is zero.

Now consider plotting $f_\theta(1.75) f_\theta(0.5) f_\theta(2.2)$ versus θ. Till θ reaches 2.2, at least one of the three terms in the product is zero and hence the product is zero. For $\theta \geq 2.2$, the product would have value $1/\theta^3$.

From the above we can generalize as follows. Given three data points, x_1, x_2, x_3, $L(\theta \mid \mathcal{D}) = 0$ if $\theta < \max(x_1, x_2, x_3)$. For $\theta \geq \max(x_1, x_2, x_3)$, $L(\theta \mid \mathcal{D}) = (1/\theta^3)$. So, the likelihood function attains its maximum value at $\max(x_1, x_2, x_3)$. This is schematically illustrated in figure 1.

Generalizing this we have the following. Given n samples, $x_1, \cdots, x_n$, the ML estimate for θ is $\max_i(x_i)$.

6. Suppose you have n samples from a normal density with mean μ and variance 1. You estimated the mean using the sample mean. Then you discover that your friend had m samples from the same density and has estimated the mean using sample mean. How should you combine your estimates to get a better estimate.

Answer: Let us write $\hat{\mu}_n$ and $\hat{\mu}_m$ for the two estimates with sample sizes n and m respectively.

Intuitively taking average of averages does not make sense. If we take sum of all the $m+n$ samples and divide by $m+n$, it may be better. Thus a good way to combine might be

$$\hat{\mu} = \frac{n\hat{\mu}_n + m\hat{\mu}_m}{n+m}$$

We can argue it like this. Consider a combination

$$\hat{\mu} = \alpha\hat{\mu}_n + (1-\alpha)\hat{\mu}_m$$

Now we can ask what value of α would minimize the variance of the estimate $\hat{\mu}$. (Since it is an unbiased estimate, variance is the mean square error).

Since we can assume that the samples of the two people are independent, $\hat{\mu}_n$ and $\hat{\mu}_m$ are independent. Also, since variance of the underlying distribution is given as 1, the variance of these two estimates are $1/n$ and $1/m$ respectively. Let variance of $\hat{\mu}$ be η^2. Then

$$\eta^2 = \alpha^2\frac{1}{n} + (1-\alpha)^2\frac{1}{m}$$

By differentiating this with respect to α and equating to zero, we get

$$\frac{\alpha}{n} = \frac{1-\alpha}{m} \quad \Rightarrow \quad \alpha = \frac{n}{m+n}$$

That shows that our intuitive idea is correct.