CS 434a-541a
Nov. 8, 2004 Midterm

Instructions: Show all the work

Problem 1 (15%): Consider a one dimensional two category classification problem. Suppose the class conditional densities are given by $p(x|c_1)=N(2,1)$ and

$$p(x/c_2)=\begin{cases}\frac{1}{3} & \text{if } 0\le x\le 3\\ 0 & \text{otherwise}\end{cases}$$

(a) Design the ML classifier
(b) Suppose priors are $P(c_1)=2/3$ and $P(c_2)=1/3$. Design the MAP classifier
(c) Suppose priors are equal and the loss functions are as follows: loss for deciding c_1 when the true class is c_2 is 30, loss for deciding c_2 when the true class c_1 is 20. Loss for deciding the true class is 0. Design the Minimum Bayes risk classifier.

In each case above, give the decision boundaries and decision regions

Solution:

$$p(x/c_1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-2)^2}$$

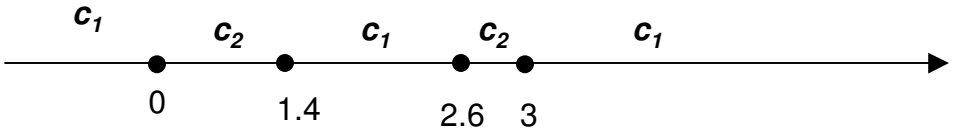
(a) According to ML classifier, $\frac{p(x/c_1)}{p(x/c_2)}\underset{<_{c_2}}{>_{c_1}}1$

For x <0 and x > 3, $p(c_2|x) = 0$, so decide class 1.

For 0<= x <= 3, $\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-2)^2}}{1/3}\underset{<_{c_2}}{>_{c_1}}1 \Leftrightarrow e^{-\frac{1}{2}(x-2)^2}\underset{<_{c_2}}{>_{c_1}}\frac{\sqrt{2\pi}}{3} \Leftrightarrow -\frac{1}{2}(x-2)^2\underset{<_{c_2}}{>_{c_1}}\ln\left(\frac{\sqrt{2\pi}}{3}\right) \Leftrightarrow (x-2)^2\underset{<_{c_1}}{>_{c_2}}0.36$

We need to solve $(x-2)^2=0.36$, which is easily solved by $x-2=\pm\sqrt{0.36} \Rightarrow x=2.60 \text{ and } 1.4$

Solution regions and boundaries:



(b) According to MAP classifier, $\frac{p(x/c_1)P(c_1)}{p(x/c_2)P(c_2)} \underset{c_2}{\overset{c_1}{\gtrless}} 1$

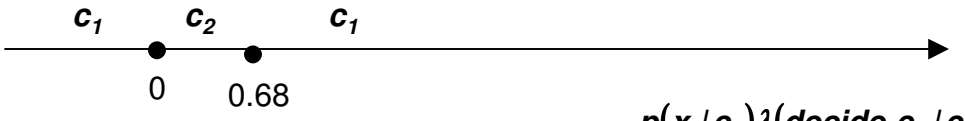
For x <0 and x > 3, $p(c_2|x) = 0$, so decide class 1.

For 0<= x <= 3,

$$\frac{\frac{2}{3}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-2)^2}}{\frac{1}{3}\cdot 1/3} \underset{c_2}{\overset{c_1}{\gtrless}} 1 \Leftrightarrow e^{-\frac{1}{2}(x-2)^2} \underset{c_2}{\overset{c_1}{\gtrless}} \frac{\sqrt{2\pi}}{6} \Leftrightarrow -\frac{1}{2}(x-2)^2 \underset{c_2}{\overset{c_1}{\gtrless}} \ln\left(\frac{\sqrt{2\pi}}{6}\right) \Leftrightarrow (x-2)^2 \underset{c_1}{\overset{c_2}{\gtrless}} 1.75$$

We need to solve $(x-2)^2 = 1.75$, which is easily solved by $x-2 = \pm\sqrt{1.75} \Rightarrow x = 3.32 \text{ and } 0.68$

3.32 >3 so only class 1 is possible. Therefore decision regions and boundaries are:



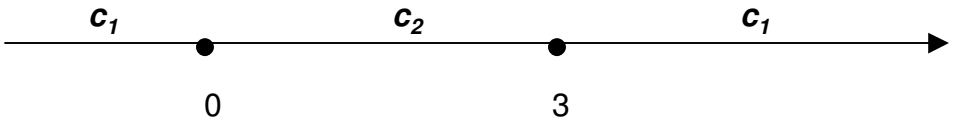
(c) According to Bayes classifier, since priors are equal, $\frac{p(x/c_1)\lambda(\text{decide } c_2/c_1)}{p(x/c_2)\lambda(\text{decide } c_1/c_2)} \underset{c_2}{\overset{c_1}{\gtrless}} 1$

For x <0 and x > 3, $p(c_2|x) = 0$, so decide class 1.

For 0<= x <= 3,

$$\frac{20\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-2)^2}}{30\cdot 1/3} \underset{c_2}{\overset{c_1}{\gtrless}} 1 \Leftrightarrow e^{-\frac{1}{2}(x-2)^2} \underset{c_2}{\overset{c_1}{\gtrless}} \frac{\sqrt{2\pi}}{2} \Leftrightarrow -\frac{1}{2}(x-2)^2 \underset{c_2}{\overset{c_1}{\gtrless}} \ln\left(\frac{\sqrt{2\pi}}{2}\right) \Leftrightarrow (x-2)^2 \underset{c_1}{\overset{c_2}{\gtrless}} -0.45$$

Since $(x-2)^2 > 0 \text{ for all } x$ we always decide class 2 for 0<= x <= 3



Problem 2 (8%): Suppose you have a three category 4 dimensional classification problem. Each class has a multivariate normal distribution. The covaraince matrices are equal for all classes and are

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 30 \end{bmatrix}$$

The means for each class are, respectively, $\mu_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$ $\mu_2 = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 3 \end{bmatrix}$ $\mu_3 = \begin{bmatrix} 3 \\ 4 \\ 1 \\ 2 \end{bmatrix}$

all classes have equal prior. Using the MAP classifier, classify sample $\boldsymbol{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \end{bmatrix}$

Solution: In case of equal priors and equal covariance matrices, we choose the class which has closest mean according to the Mahalanobis Distance

$$\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\Sigma^{-1}} = (\boldsymbol{x} - \boldsymbol{\mu})^t \sum^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

Inverse of the diagonal matrix is $\Sigma^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/10 & 0 & 0 \\ 0 & 0 & 1/20 & 0 \\ 0 & 0 & 0 & 1/30 \end{bmatrix}$, which is easy to compute since we know that $\Sigma\Sigma^{-1} = I$

$\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2_{\Sigma^{-1}} = \sum_{k=1}^{4} \sigma_k \left(\boldsymbol{x}^{(k)} - \boldsymbol{\mu}_i^{(k)}\right)^2$, where σ_k is the kth element on the diagonal

We can find the answer by observing that σ_1 is much bigger than σ_2 ,σ_3, and σ_4. Thus we can just look at the first feature for classification, and the closest mean is 1, which means $\boldsymbol{x}$ should be classified as class 1.

Alternatively, we could compute all 3 distances without making shortcuts.

$$\|\boldsymbol{x} - \boldsymbol{\mu}_1\|^2_{\Sigma^{-1}} = (0)^2 + \frac{1}{10}(1)^2 + \frac{1}{20}(2)^2 + \frac{1}{30}(2)^2 < 1$$

$$\|\boldsymbol{x} - \boldsymbol{\mu}_2\|^2_{\Sigma^{-1}} = (1)^2 + \frac{1}{10}(0)^2 + \frac{1}{20}(0)^2 + \frac{1}{30}(1)^2 > 1$$

$$\|\boldsymbol{x} - \boldsymbol{\mu}_3\|^2_{\Sigma^{-1}} = (2)^2 + \frac{1}{10}(3)^2 + \frac{1}{20}(0)^2 + \frac{1}{30}(0)^2 > 1$$

Problem 3 (12%): Consider a one dimensional two category classification problem. Suppose the class conditional densities are given by

$$p(x/c_1)=\begin{cases}0 & \text{if } x<0\\ \theta_1 e^{-\theta_1 x} & \text{if } x\geq 0\end{cases} \qquad p(x/c_2)=\begin{cases}0 & \text{if } x<0\\ \theta_2 e^{-\theta_2 x} & \text{if } x\geq 0\end{cases}$$

Suppose you have collected data for class 1 and stored it in array D_1= {1,5} and data for class 2 and stored it in array D_2={3,6,9}.

(a) Find the maximum likelihood estimates for parameters θ_1 and θ_2.
(b) Using ML classifier, find decision regions and decision boundaries using the parameters estimated in (a)

Solution:

(a) The likelihood for the data for class 1 is

$$p(D/c_1)=\theta_1^2 e^{-\theta_1(x_1+x_2)}$$

Take the derivative and set it to 0:

$$\frac{d}{d\theta_1}p(D/c_1)=2\theta_1 e^{-\theta_1(x_1+x_2)}-(x_1+x_2)\theta_1^2 e^{-\theta_1(x_1+x_2)}=0\Rightarrow$$
$$\theta_1 e^{-\theta_1(x_1+x_2)}(2-(x_1+x_2)\theta_1)=0\Rightarrow \quad \theta_1=0 \text{ or } \theta_1=\frac{2}{x_1+x_2}=\frac{1}{3}$$

θ_1=0 is not a valid solution because p(x|c_1) = 0 is this case, which is not a density. Similarly, the likelihood for the data for class 2 is

$$p(D/c_2)=\theta_2^3 e^{-\theta_2(x_1+x_2+x_3)}$$

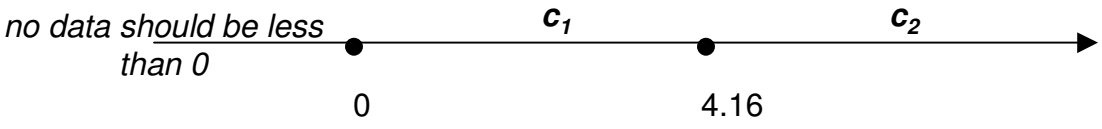
Take the derivative and set it to 0:

$$\frac{d}{d\theta_2}p(D/c_2)=3\theta_2^2 e^{-\theta_1(x_1+x_2+x_3)}-(x_1+x_2+x_3)\theta_2^3 e^{-\theta_2(x_1+x_2+x_3)}=0\Rightarrow$$

$$\theta_2^2 e^{-\theta_2(x_1+x_2+x_3)}(3-(x_1+x_2+x_3)\theta_2)=0\Rightarrow \quad \theta_2=0 \text{ or } \theta_2=\frac{3}{x_1+x_2+x_3}=\frac{1}{6}$$

(b)

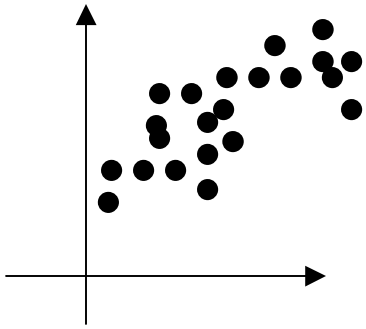
$$\frac{p(x/c_1)}{p(x/c_2)}\gtrless 1 \Rightarrow \frac{\frac{1}{3}e^{-\frac{1}{3}x}}{\frac{1}{6}e^{-\frac{1}{6}x}}\gtrless 1 \Rightarrow 2e^{-\frac{1}{6}x}\gtrless 1 \Rightarrow -\frac{1}{6}x\gtrless \ln\left(\frac{1}{2}\right) \Rightarrow x\gtrless 4.16$$



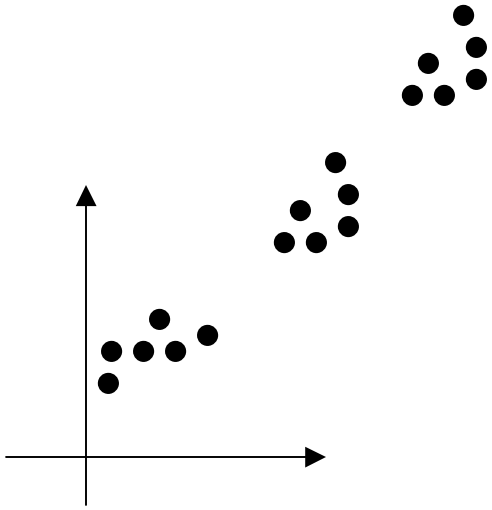
Problem 4 (6%): Sketch the best one dimensional representation of the following two dimensional set of samples

Solutions

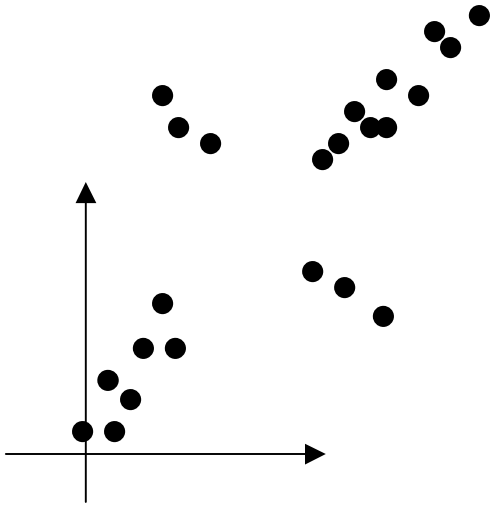
(a)



(b)



(c)



Problem 5 (8%): Recall that the idea behind the Fisher linear discriminant is to project samples to a line so that samples from different classes project to clusters on a line which overlap as little as possible. Let the projected means be μ_1, μ_2 and the projected variances be σ^2_1, σ^2_2. Suppose we want to modify the Fischer linear discriminant approach by using a different objective function $J(\mathbf{v})$ for finding the best line direction $\mathbf{v}$ to project samples to. We plan on maximizing $J(\mathbf{v})$. Which of the following would be a good objective functions to use, which are not? Give 1 sentence explanations.

Solutions

(a) $J(v) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2}$

This is not a good objective function because it does not enforce small projected variance for class 2

(b) $J(v) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 - \sigma_2^2}$

This is not a good objective function because it enforces differences between variances to be small, but we need the individual varinces for each class to be small

(c) $J(v) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2} + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}$

Good objective function because it enforces differences between the means to be large, and the individual variances for each class to be small

(d) $J(v) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 / \sigma_2^2}$

Not good objective function because it enforces variance for the second class to be large, abut we need it to be small

Problem 6 (15%): Consider a one dimensional two class classification problem, where we have collected the following data for each class: $D_1 = \{-1,-2,3,3,6,7\}$ and $D_2 = \{-3,-2,3,5,8\}$. Suppose we decided to use Parzen windows with window width $h = 2$ and $\varphi(x)$ defined as

$$\varphi(u) = \begin{cases} -\frac{3}{4}u^2 + \frac{3}{4} & \text{if } -1 \le u \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Classify sample $x = 4$ using the ML classifier

(b) Show that the density estimates using Parzen windows with $\varphi(x)$ given as above are true density functions

Solution:

(a) $$p_1(4) = \frac{1}{6}\sum_{i=1}^{6}\frac{1}{2}\varphi\left(\frac{4-x_i}{2}\right) = \frac{1}{12}\left(\varphi\left(\frac{1}{2}\right)+\varphi\left(\frac{1}{2}\right)\right) = \frac{\varphi\left(\frac{1}{2}\right)}{6}$$

$$p_2(4) = \frac{1}{5}\sum_{i=1}^{5}\frac{1}{2}\varphi\left(\frac{4-x_i}{2}\right) = \frac{1}{10}\left(\varphi\left(\frac{1}{2}\right)+\varphi\left(-\frac{1}{2}\right)\right) = \frac{\varphi\left(\frac{1}{2}\right)}{5}$$

where the last step holds because $\varphi\left(\frac{1}{2}\right) = \varphi\left(-\frac{1}{2}\right)$

Thus $x = 4$ should be classified as class 2

(b) To show that density estimate is a true density, all we need to show that the window function $\varphi(u)$ is nonnegative and it's integral is 1.

Obviously $\varphi(u)$ >=0, since $-\frac{3}{4}u^2 + \frac{3}{4}$ is nonnegative between –1 and 1

$$\int_{-\infty}^{\infty}\varphi(u)du = \int_{-1}^{1}\left(-\frac{3}{4}u^2+\frac{3}{4}\right)du = -\frac{3}{12}u^3 + \frac{3}{4}u\Big|_{-1}^{1} = \frac{6}{12}+\frac{6}{12} = 1$$

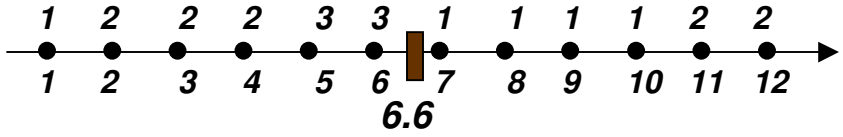
Problem 7 (10%): Suppose we have a collected the following one dimensional data from three classes: D_1={1,7,8,9,10}, D_2={2,3,4,11,12}, D_3={5,6}.

(a) classify sample ***6.6*** using ***knn*** algorithm with ***k*=*6***

(b) Apply the ***editing*** algorithm to the data and show the reduced set of samples it finds

Solution:

(a)



Nearest 6 neighbors of ***6.6*** are 7,8,9 from class 1 and 5,6 from class 3 and 4 from class 2. Thus the largest number of neighbors is from class 1 and ***6.6*** should be classified as class 1

(b) Samples 3, 8, 9, 12 have all the neighbors in the Voronoi diagram of the same class, and thus they can be removed. The resulting reduced set of samples: D_1={1,7,10}, D_2={2,4,11,}, D_3={5,6}.

Problem 8 (15%):

(a) Write pseudo code for minimizing with gradient decent the objective function

$$J(v) = \frac{v^t A v + v^t C v}{v^t B v}$$

where v is a vector of unknowns of size n, and A, B, C are completely known matrices of size n by n.

(b) Suppose you do not limit the number of iterations performed by your algorithm. How should you set the learning rate to ensure the convergence of your algorithm?

Solution:

(a) We can write $J(v) = \frac{v^t (A + C) v}{v^t B v}$

$$\nabla J(v) = \frac{(2(A+C)v)(v^t B v) - (v^t (A+C)v)(2Bv)}{(v^t B v)^2}$$

set $k = 1$ **and** $v^{(1)}$ =to some initial guess for the weight vector
choose $\eta^{(k)} = 1/k$

while $\eta^{(k)} \left| \nabla J(v^{(k)}) \right| > \varepsilon$

$v^{(k+1)} = v^{(k)} - \eta^{(k)} \nabla J(v)$

$k = k + 1$

(b) Learning rate as above, 1/k will work because the steps eventually will be too small to pass the condition of the while loop

Problem 9 (11%): *Give a one sentence answer to the questions below:*

(a) Which classifier should you use for minimum error rate classification?
Map Classifier

(b) What does it mean to say that 2 features are negatively correlated?
When one increases, the other tends to decrease, and when one decreases, the other tends to increase

(c) Suppose we have an estimator $\boldsymbol{\theta}^*$ for a parameter $\boldsymbol{\theta}$. State 2 conditions for $\boldsymbol{\theta}^*$ to be a good estimator of $\boldsymbol{\theta}$.
Estimator $\boldsymbol{\theta}^*$ should be unbiased and have low variance

(d) Name one difference between ML and Bayesian parameter estimation
(1)*In Bayesian estimation, unknown parameter is modeled as a random variable, while in ML estimation it is a unknown but fixed;(2) BPE envolves integration, MLE envolves differentiation*

(e) Suppose we are not happy with our classifier and decide to increase the number of features hoping to improve the performance. Suppose classification error goes up, instead of expected down. Name one reason why this may have happened.
(1) o*verfitting; (2) samples are not dense enough*

(f) Name one drawback for using $\boldsymbol{k}$ nearest neighbors rule for density estimation
(1) The resulting density is usually not even a density, (2) Has discontinuities

(g) Name one advantage for using weights $\boldsymbol{b}=[1,1,\ldots 1]^{\boldsymbol{t}}$ in the MSE procedure
(1) The resulting MSE solution is basically identical to Fischer's linear discriminant solution; (2) MSE solution approaches the Bayes discriminant function as the number of samples goes to infinity

(h) Suppose you are using gradient descent and printing out the value of the objective function at each iteration. This value has been going down, as it is supposed to, for 100 iterations, but then at iteration 101 it went up. What happened?
We overshot the a local minimum

(i) For which functions gradient decent should not be even attempted?
Discontinuous functions

(j) Suppose we have 5 samples and we know that the first sample should lie closer to the separating hyperplane and the last sample should lie far from the separating hyperplane, compared to the other samples. Suggest an appropriate $\boldsymbol{b}$ for the MSE procedure.
b = [1 10 10 10 100]

(k) What is an optimal data distribution for a linear discriminant function?
Gaussian distribution with equal covariances