CAP 5638: Mid Term

Due on Monday, Nov. 23, 2015

XiuWen Liu 10:10am

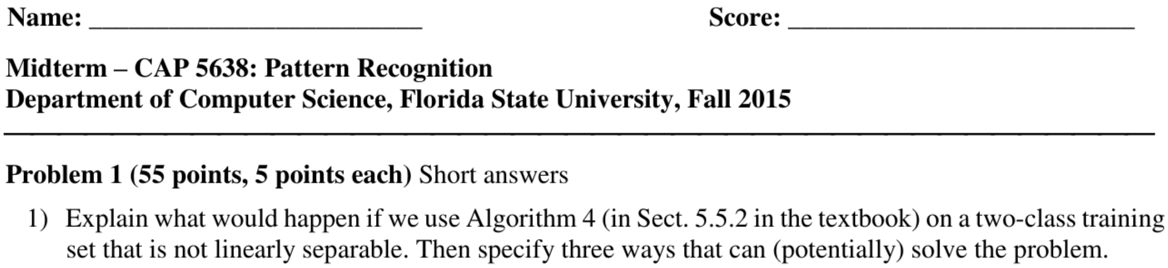
Jian Wang

Contents

Problem 1 **3**

Problem 2 **7**

Problem 1



Name: ________________________ **Score:** __________________________

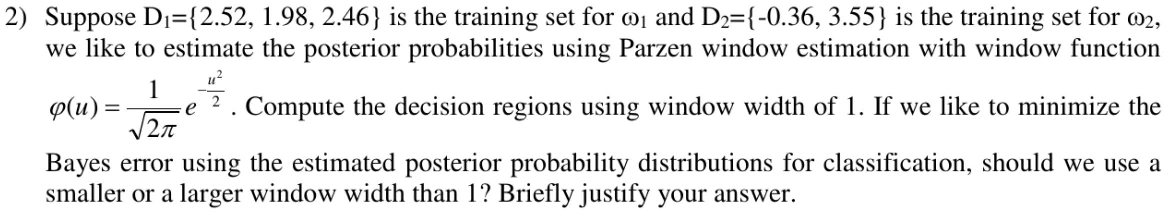
Midterm – CAP 5638: Pattern Recognition
Department of Computer Science, Florida State University, Fall 2015

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.

Answer:
If the two class training set is not linearly separable, the corrections in an error-correction procedure in algorithm 4 can never cease. So Algorithm 4 will never converge and can not stop.

The method to solve the problem:
1) We can use a nonlinear mapping to make the problem separable
2) We can nultiple linear discriminant functions
3) We can learn nonlinear discriminant functions directly.



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.

Answer:
1) For the parzen windows, we can use the following formula to calculate the posterior probability:

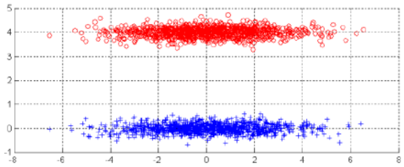
$$p_n(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi(\frac{x-x_i}{h_n})$$

where the h_n=1.
In this case, we can write the formula based on the dataset,given a test data x.

$$p_1(x)=\frac{1}{\sqrt{2\pi}}\frac{1}{n}(e^{-\frac{(x-2.52)^2}{2}}+e^{-\frac{(x-1.98)^2}{2}}+e^{-\frac{(x-2.46)^2}{2}})$$

$$p_2(x)=\frac{1}{\sqrt{2\pi}}\frac{1}{n}(e^{-\frac{(x+0.36)^2}{2}}+e^{-\frac{(x-3.55)^2}{2}})$$

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.

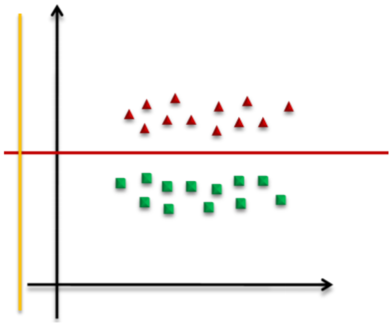


The decision boundary is the x_0 which make $p_1(x)$ equal to the $p_2(x)$, so the decision region is that, when $x > x_0$, x belongs to class 1 and when $x <= x_0$ x belongs to class 2

Answer:

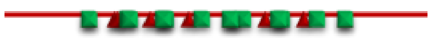
PCA: : Perform dimensionality reduction while preserving as much of the variance in the high dimensional space as possible.

FDA:Perform dimensionality reduction while preserving as much of the class discriminatory information as possible.

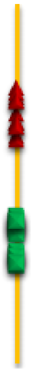


In this case: we can see that the data can be classified by a horizontal line, however the principal component will be the x_1.

So the PCA will have higher variance and bad discriminability:

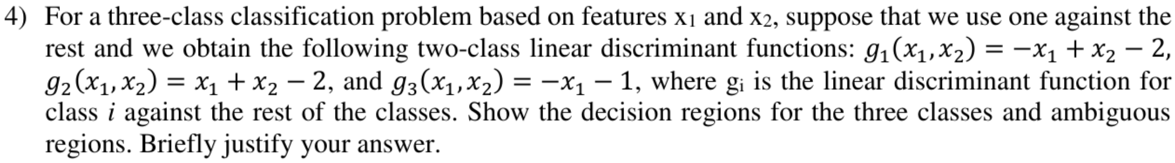


and the FDA will have smaller variance and good discriminability



For the area under the ROC curve, the PCA will have the region under the line $y = x$, which will be around 0.5.

and for the FDA the area of ROC curve will be around the unit square which will be around 1.



4) For a three-class classification problem based on features x1 and x2, 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 gi 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.

Answer:

Use the one against all method, so if the test data is belongs to class 1, the $g_1(x)$ will be bigger than the $g_2(x)$ and $g_3(x)$. Same situation for the class 2 and class 3. So when test data x belongs to class 1:
$g_1(x_1, x_2) > g_2(x_1, x_2) \Rightarrow -2x_1 > 0$ and $g_1(x_1, x_2) > g_3(x_1, x_2) \Rightarrow x_2 > 1$

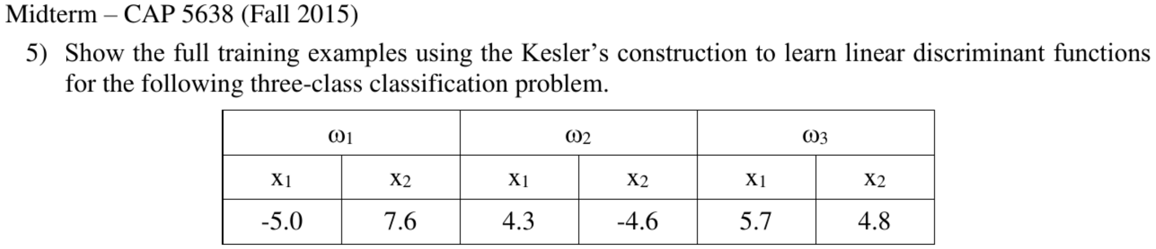
so: $x_1 < 0$ and $x_2 > 1$
When test data x belongs to class 1:
$g_2(x_1, x_2) > g_1(x_1, x_2) \Rightarrow -2x_1 > 0$ and $g_2(x_1, x_2) > g_3(x_1, x_2) \Rightarrow x_2 > 1$

When test data x belongs to class 2:
$g_1(x_1, x_2) > g_2(x_1, x_2) \Rightarrow 2x_1 > 0$ and $g_1(x_1, x_2) > g_2(x_1, x_2) \Rightarrow 2x_1 + x_2 - 1 > 0$

so : $x_1 > 0$ and $x_2 > -2x_1 + 1$

When test data x belongs to class 3:
$g_3(x_1, x_2) > g_1(x_1, x_2) \Rightarrow -x_2 + 1 > 0$ and $g_3(x_1, x_2) > g_2(x_1, x_2) \Rightarrow -2x_1 - x_2 + 1 > 0$

so : $x_1 > 0$ and $x_2 < -2x_2 + 1$
the decision regions for the three class is as the following chart.

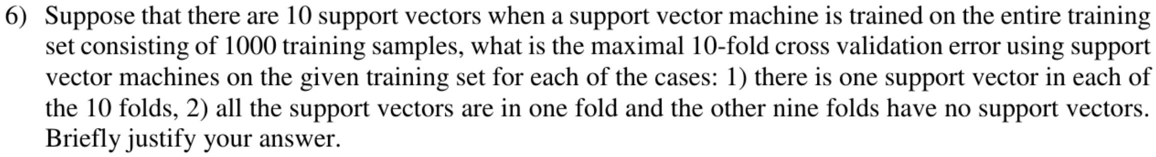


Midterm – CAP 5638 (Fall 2015)

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

ω1		ω2		ω3	
x1	x2	x1	x2	x1	x2
-5.0	7.6	4.3	-4.6	5.7	4.8

Answer:



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.

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.

Answer:

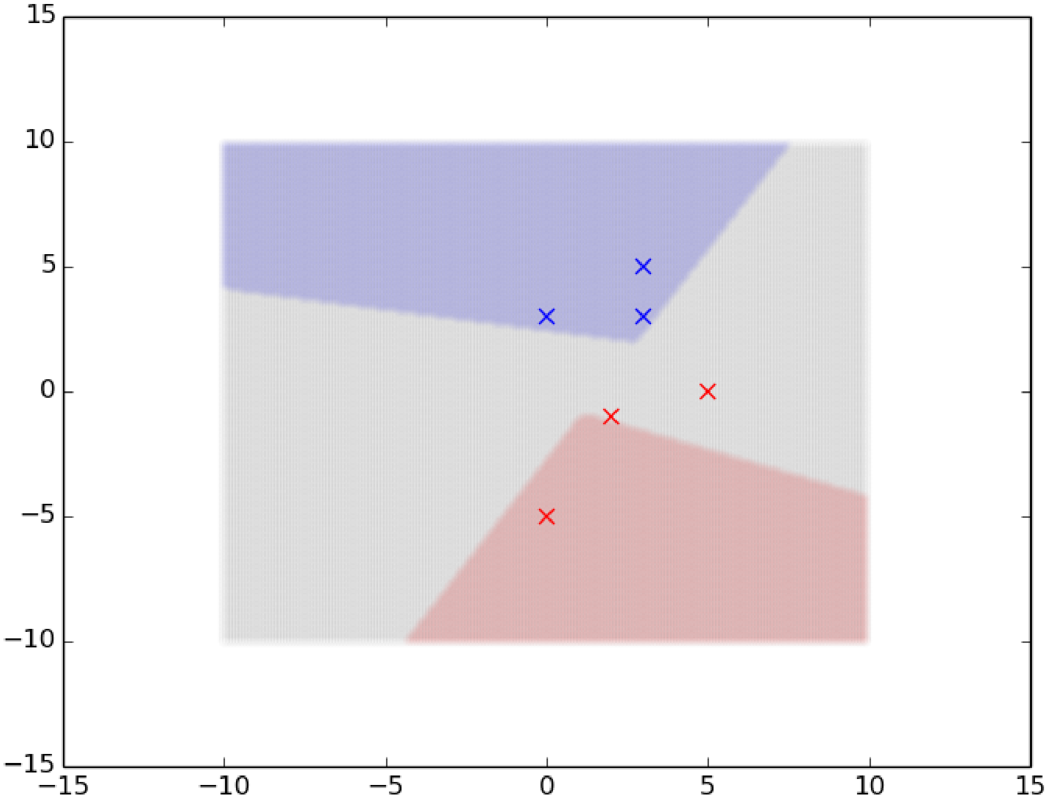
8) Suppose the feature from the previous question gives the lowest error among all the features, specify the weights for all the samples to be used for computing the second weak classifier. You need to justify your answers.

Answer:

Midterm – CAP 5638 (Fall 2015)

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.

Answer:



10) Given a training set D={x$_1$, …, x$_n$}, derive the maximum likelihood estimate for the following model:

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

Answer:

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 $f(n) = \begin{cases} 1 & \text{for } n \ge 0; \\ -1 & \textit{otherwise.} \end{cases}$.

Answer:

Problem 2

Midterm – CAP 5638 (Fall 2015)

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 $0 \le x \le \theta$ and 0 otherwise) for class 2.

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

Answer:

We denote variable x as 1 if x belongs to class 1, as 0 if x belongs to class 2.
p as the probability if $x = 1$.

so the likelihood for p given dataset$x_i$is:

$$\prod_{i=1}^{n} p^{x_i} p^{1-x_i}$$

so the maximum likelihood of p is:

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

In this case: the prior probability that data belongs to class 1 $p = 5/8$ and the priror probability that data belongs to class 2 is $1 - p = 3/8$

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.

Answer:

we know that the maximum likelihood for normal distribution is:
μ is the sample mean and the σ is the sample standard deviation.
in this case:

$$\mu = \frac{1}{5}(7.5 + 4.5 + 3.5 + 6.0 + 7.5) = 5.8$$

$$\sigma^2 = \frac{1}{5}((7.5 - 5.8)^2 + (4.5 - 55.8)^2 + (3.5 - 5.8)^2 + (6.0 - 5.8)^2 + (7.5 - 5.8)^2) = 2.56$$

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.

Answer:

$$p(\theta|D_0) = \begin{cases} \frac{1}{8}, & 2 \le \theta \le 10 \\ 0, & otherwise \end{cases}$$

$$p(\theta|D_1) = \begin{cases} \frac{1}{\theta}\frac{1}{(ln10 - ln2.5)}, & 2.5 \le \theta \le 10 \\ 0, & otherwise \end{cases}$$

$$p(\theta|D_2) = \begin{cases} \frac{1}{\theta^2}\frac{1}{(\frac{1}{3.5} - \frac{1}{10})}, & 3.5 \le \theta \le 10 \\ 0, & otherwise \end{cases}$$

$$p(\theta|D_3) = \begin{cases} \frac{1}{\theta^3}\frac{2}{(\frac{1}{4.0^2} - \frac{1}{10^2})} = \frac{1}{\theta^3}38.10, & 4.0 \le \theta \le 10 \\ 0, & otherwise \end{cases}$$

$$p(x|D_3) = \int p(x|\theta)p(\theta|D_3) = \begin{cases} \int_4^{10} \frac{1}{\theta}\frac{1}{\theta^3}\frac{1}{38.10} = \frac{1}{3}(\frac{1}{4^3} - \frac{1}{10^3})38.10 = 0.18573749999999997, & 0 \le x \le 4 \\ \int_x^{10} \frac{1}{\theta}\frac{1}{\theta^3}\frac{1}{38.10} = \frac{1}{3}(\frac{1}{x^3} - \frac{1}{10^3})38.10, & 4 \le x \le 10 \\ 0, & otherwise \end{cases}$$

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

Answer:

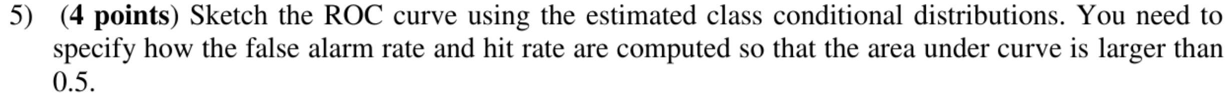
For the test set [-0.5,1.5,5.5,8.5], we define the discriminant function as:

$$g_i(x) = p(x|\omega_i)p(\omega_i)$$

when x= -0.5:
$g_1(x)$=6.6994574685438983e-05, $g_2(x)$=0, so x belongs class 1
$g_1(x)$=0.0042100787891683278, $g_2(x)$= 0.06965156249999999, so x belongs class 2
$g_1(x)$=0.15312144462991092, $g_2(x)$=0.023862593914350114, so x belongs class 1

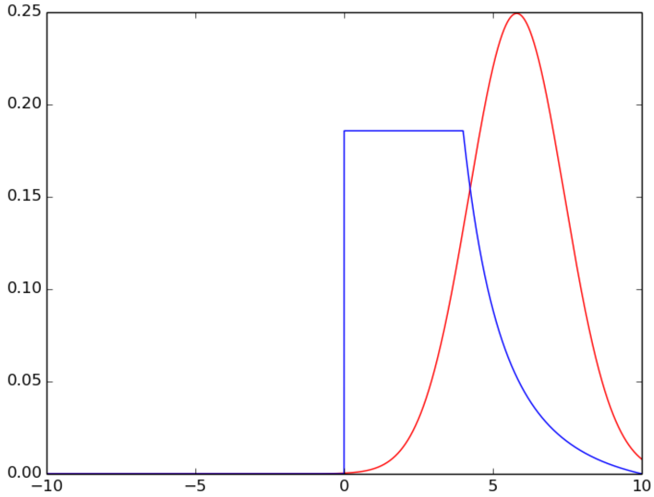
$g_1(x)$=0.037524023791059902, $g_2(x)$=0.0029924358843883576, so x belongs class 1

the final result are [1,2,1,1]

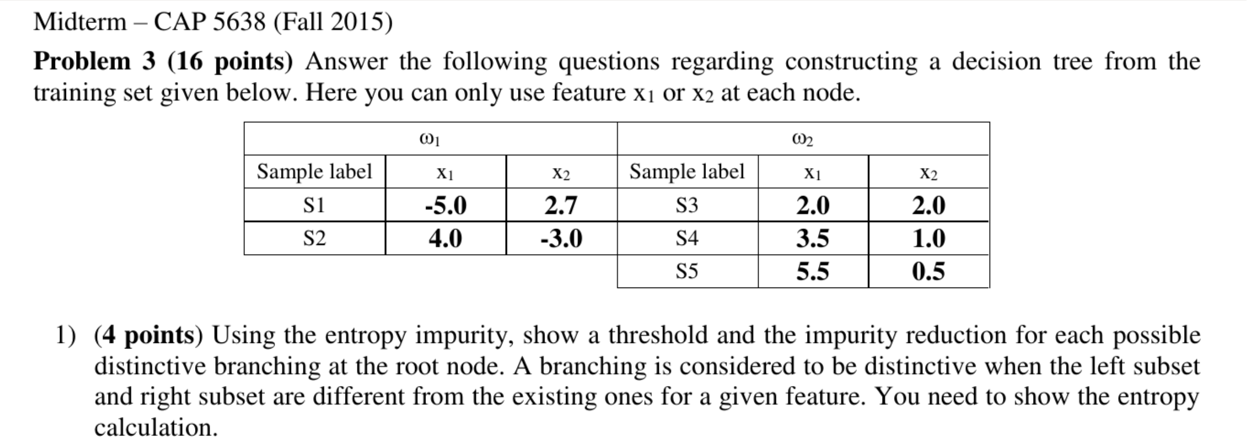


5) **(4 points)** Sketch the ROC curve using the estimated class conditional distributions. You need to specify how the false alarm rate and hit rate are computed so that the area under curve is larger than 0.5.

Answer:
The following chart showed the probability density function of the two class,



Roc curve is:



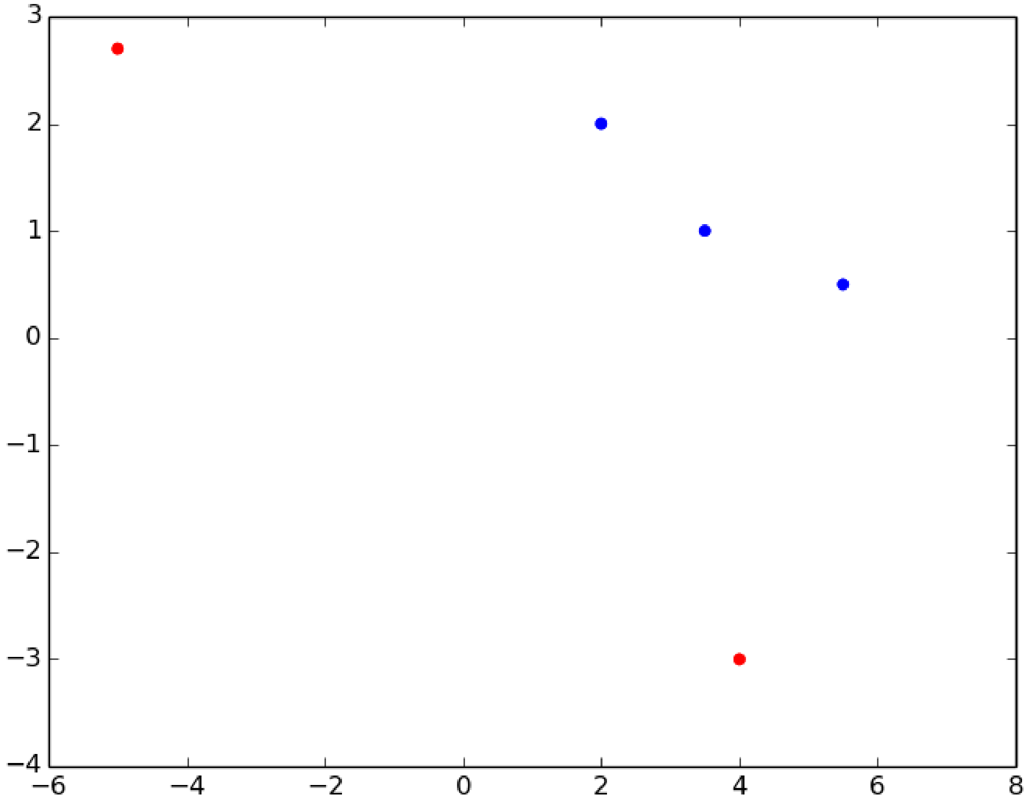
Midterm – CAP 5638 (Fall 2015)

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.

Answer:
The data was shown on the following chart:



first, we can calculate the original entropy:

$$Entropy_0 = -\frac{2}{5} * log_2\frac{2}{5} - \frac{3}{5} * log_2\frac{3}{5} = 0.9710$$

next need found a root which has the biggest information gain.
the sorted x_1 value is [-5,2,3.5,4,5.5]
the sorted x_2 value is [-3,0.5,1,2,2.7]

calculate the mid point for each interval of x_1 and x_2,show the process in the following:

when $x_1 < -1.5$, on the left one node is class 2, entropy is 0, on the right, three node is class 1, one node is class 2, so the entropy will be:

$$Entropy = -\frac{1}{4} * log_2\frac{1}{1} - \frac{3}{4} * log_2\frac{3}{4} = 0.8113$$

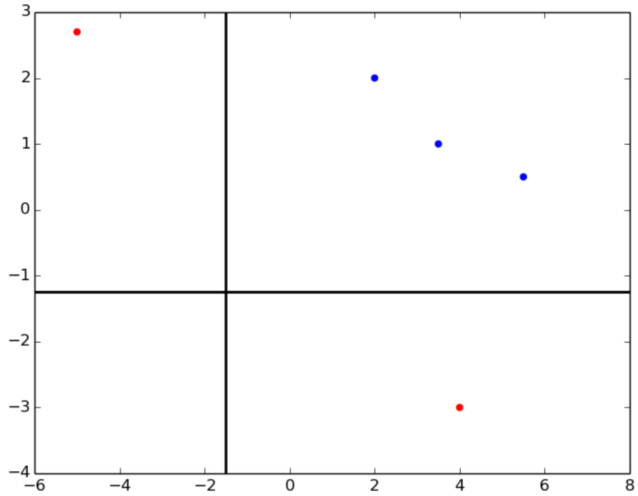
so the information gain is:

$$\Delta i(N) = i(N) - \sum_{1}^{2} p_k i(N_k) = 0.9710 - \frac{4}{5} * 0.8113 = 0.32196$$

for $x_1 < 2.75$: $\Delta i(N) = 0.025$
for $x_1 < 3.75$: $\Delta i(N) = 0.025$
for $x_1 < 4.75$: $\Delta i(N) = 0.1710$
for $x_2 < -1.25$: $\Delta i(N) = 0.32196$
for $x_2 < 0.75$: $\Delta i(N) = 0.025$
for $x_2 < 1.5$: $\Delta i(N) = 0.025$
for $x_2 < 2.35$: $\Delta i(N) = 0.32196$

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.

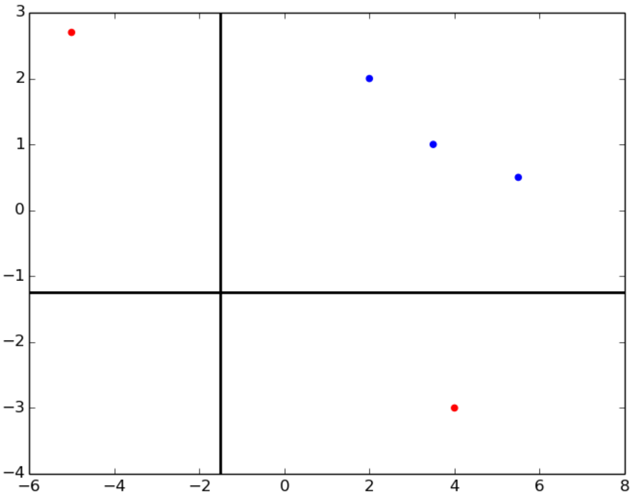
Answer:
we can find a decision tree by: $x_1 < -1.5$, data belongs to class 2, and when $x_1 > -1.5$ and $x_2 < -1.25$ data belongs to class 2, otherwise data belongs to class 1.



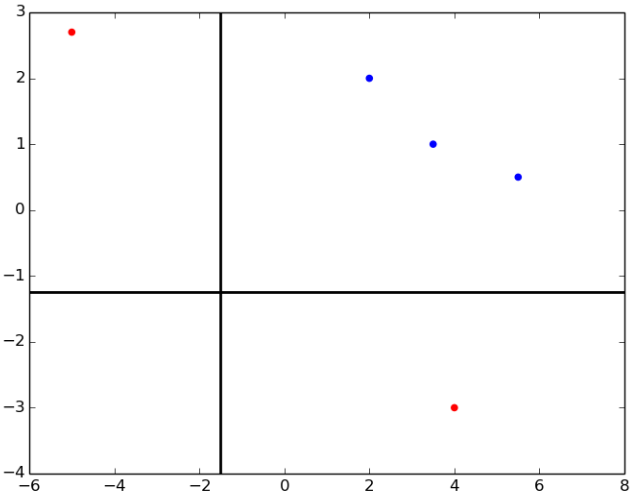
3) (**4 points**) Specify the classification rules for ω₁ and ω₂ respectively based on the decision tree you obtained for part 2).

Answer:

similar answer for the problem 2, $x_1 < -1.5$, data belongs to class 2, and when $x_1 > -1.5$ and $x_2 < -1.25$ data belongs to class 2, otherwise data belongs to class 1.



$x_1 < -1.5$, data belongs to class 2, and when $x_1 > -1.5$ and $x_2 < -1.25$ data belongs to class 2, otherwise data belongs to class 1.



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.

Answer:
for formula for the Gain ratio is:

$$\frac{i(N) - \sum_{k=1}^{B} P_k i(N_k)}{-\sum_{k=1}^{B} P_k log_2 P_k}$$

For this problem, it seems no influence of for the classification of the results.