CS 530
Bayesian Lab

Introduction

The data concerns data that might be acquired by sensors on a robot. The robot has to make decisions about what its environment contains, and that might include information about material types. In this case, there are three material types, which might be metal, wood and glass or some other set, so there are three classes. The independent variables are results from three sensors; a sonar device with real values between 0 and 1, a radio frequency device with real values between 0 and 1, and an infrared device with 0 or 1 values.

The data for this lab is based on actual data, but to avoid any analysis problems, has been recreated with workable parameters.

Creating the Data

The data was produced in octave using the following, where x_{ji} represents the data for class j variable i.

$$
\begin{aligned}
x_{11} &= normal\_rnd(0.20, 0.004, 100, 1); \\
x_{21} &= normal\_rnd(0.47, 0.003, 100, 1); \\
x_{31} &= normal\_rnd(0.80, 0.005, 100, 1); \\
x_{12} &= normal\_rnd(x11 - 0.015, 0.0002, 100, 1); \\
&\quad x_{12}(find(x_{12} < 0) = 0; \\
x_{22} &= normal\_rnd(x_{21}, 0.0004, 100, 1); \\
&\quad x_{22}(find(x_{22} < 0) = 0; \\
x_{32} &= normal\_rnd(x_{31}, 0.0005, 100, 1); \\
&\quad x_{32}(find(x_{32} < 0) = 0;
\end{aligned}
$$

The data for the third variable was generated by the following:

$$
\begin{aligned}
x_{13} &= (x_{11} > mean(x_{11}) - 1.5 * sqrt(var(x_{11}))). * binary\_rnd(0.80); \\
x_{23} &= (x_{11} > mean(x_{11}) - 0.0 * sqrt(var(x_{11}))). * binary\_rnd(0.30); \\
x_{33} &= (x_{11} > mean(x_{11}) + 1 * sqrt(var(x_{11}))). * binary\_rnd(0.60);
\end{aligned}
$$

Data Analysis

The first thing you want to do is analyze the data to see what your best choices are for classification.

Means

Class	x_1	x_2	x_3
1	0.196	0.058	0.93
2	0.470	0.119	0.47
3	0.800	0.199	0.12

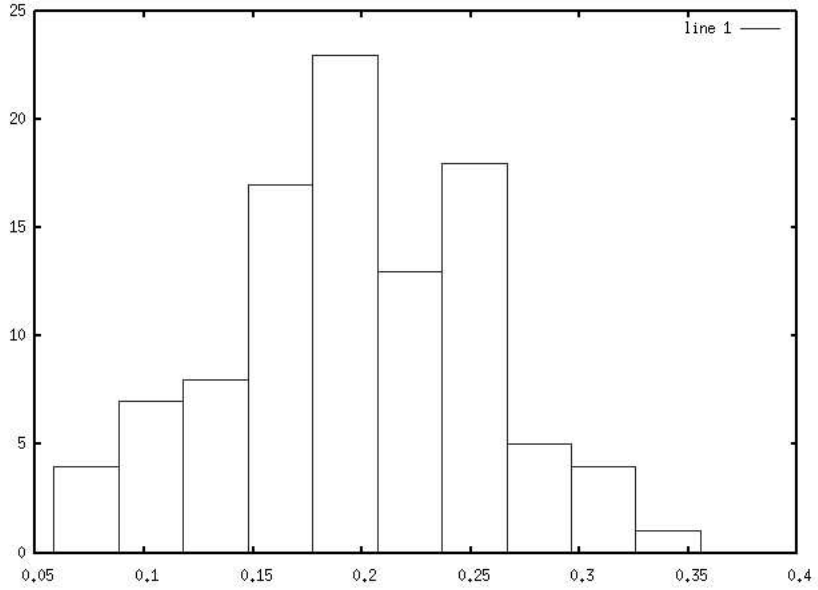
Variances

Class	x_1	x_2	x_3
1	0.004	0.003	0.066
2	0.002	0.003	0.252
3	0.004	0.005	0.107

Distributions

Next, you would be interested in the reasonableness of representing these distributions with standard distributions. In the case of variable 3, it is binary, so we will assume a binomial distribution with $\mu = np$, where p is the probability of a 1, and variance $\sigma^2 = np(1-p)$.

For variable 1, a plot for variable 1 shows:



Which looks sort of like a Normal distribution. Using the χ^2 test, our null hypothesis is that the distribution is Normally distributed with $\mu = 0.196$ and $\sigma^2 = 0.004$.

Class	Count	Prob	Expected	$(f_i - e_i)^2/e_i$
0.05 - 0.09	4	0.04	4	0.0
0.09 - 0.13	11	0.10	10	0.10
0.13 - 0.17	16	0.19	19	0.47
0.17 - 0.21	28	0.25	25	0.36
0.21 - 0.25	22	0.22	22	0.0
0.25 - 0.29	12	0.13	13	0.08
0.29 - 0.33	4	0.05	5	0.20
0.33 - 0.37	3	0.01	1	4
			Sum	5.21

The χ^2 statistic for $\alpha = 0.5$ and $\nu = 7$ is 14.1, so the null hypothesis is accepted and we can assume a Normal distribution.

You could perform the same analysis on every variable if you had concerns about them being distributed differently. Certainly you should do one for x_2, but I won't show it here.

0.1 Correlations

In order to select the best variables to use a features you need to correlations between the variables and the dependent variable (the class). x_3 will have to be treated independently, but x_1 and x_2 can be handled together. The matrix of all observations can be created with:

$$x = [x_{11}, x_{12}, ones(100, 1); x_{21}, x_{22}, 2 * ones(100, 1)];$$

The correlation coefficients with the third variable can be found by using the Spearman correlation coefficient. Kendall's correlation can be used for the class correlation.

	x_1	x_2	x_3	Class
x_1	1.0	0.82	-0.275	0.89
x_2		1.0	-0.091	0.68
x_3				-0.330
Class				1.0

So it appears that the highest correlation with the class variable is x_1. However, there is a relatively high correlation between x_1 and x_2, so using both may not be particularly useful. x_3 has a relatively low correlation with the class variable, but it also has a relatively low correlation with x_1 and x_2.

Classifiers

Single Variable

The simplest classifiers are the single variable classifiers, so investigate those first. Since x_1 has the highest correlation, that will be our first choice. Since x_1 is a continuous variable and it can be assumed to be normally distributed, a Bayesian classifier will work well.

We need to find the $p(\omega_j|x_1)$ posterior distributions in order to choose a classifier.

$$p(\omega_1|x_1) = \frac{p(x_1|\omega_1)\ P(\omega_1)}{p(x)}$$
$$= \frac{N(x;0.196,0.004)(0.33)}{N(x;0.196,0.004)0.33\ +\ N(x;0.470,0.002)0.33\ +\ N(x;0.800,0.004)0.33}$$

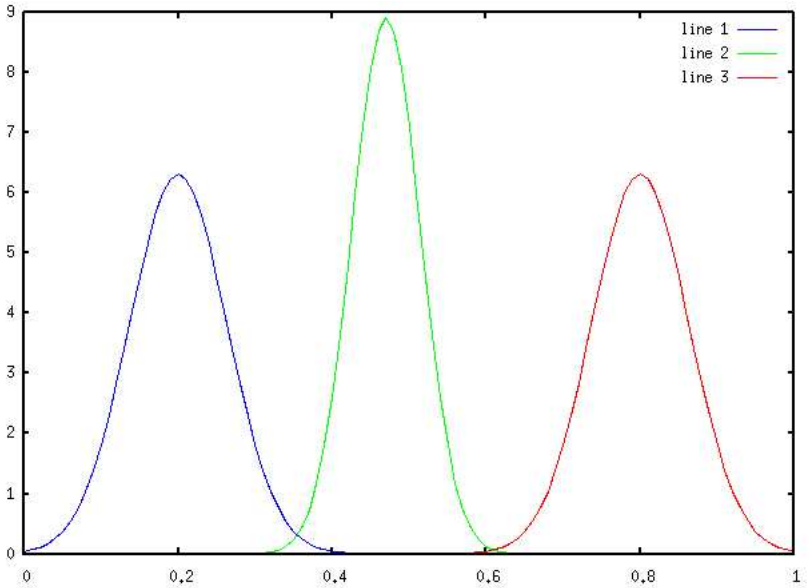
$$p(\omega_2|x_1) = \frac{p(x_1|\omega_2)\ P(\omega_2)}{p(x)}$$
$$= \frac{N(x;0.470,0.002)(0.33)}{N(x;0.196,0.004)0.33\ +\ N(x;0.470,0.002)0.33\ +\ N(x;0.800,0.004)0.33}$$

$$p(\omega_3|x_1) = \frac{p(x_1|\omega_3))\ P(\omega_3)}{p(x)}$$
$$= \frac{N(x;0.800,0.004)(0.33)}{N(x;0.196,0.004)0.33\ +\ N(x;0.470,0.002)0.33\ +\ N(x;0.800,0.004)0.33}$$

Plots of the classes using:

$$plot\ \ (\ \ sx, normal_p df(sx, 0.2, 0.004), sx, normal_p df(sx, 0.47, .002),$$
$$sx, normal_p df(sx, 0.80, 0.004))$$

show:



So there are two decision thresholds, between classes 1 and 2 and between classes 2 and 3.

$$
\begin{aligned}
g_{12}(x) &= p(\omega_1|x_1) - p(\omega_2|x_1) \\
g_{23}(x) &= p(\omega_2|x_1) - p(\omega_3|x_1)
\end{aligned}
$$

You could do a g_{13}, but it would be redundant given the geometry.

We are interested in the points where the discriminant functions, g_{ij} are zero since these provide separating planes (points in this case). It is impossible to solve for these directly since there are no closed form solutions for the normal distribution. We can use tables or programs to generate the discriminant values and then interpolate:

Using *bayes_norm*,

$bayes\_norm(0.20, 0.004, 0.33, 0.47, 0.002, 0.33, 20)$

x_1	$p(\omega_1\|x_1$	$p(\omega_2\|x_1)$	$p(\omega_1\|x_1) - p(\omega_2\|x_1)$
⋮	⋮	⋮	⋮
0.25740	1.00000	0.00000	0.99999
0.29620	0.99958	0.00042	0.99915
0.33500	0.97510	0.02490	0.95021
0.37380	0.48094	0.51906	-0.03811
0.41260	0.03017	0.96983	-0.93966
0.45139	0.00147	0.99853	-0.99706
⋮	⋮	⋮	⋮

$g_{12} = 0$ at 0.37 and similarly, $g_23 = 0$ at 0.62.

Other solutions are to use an approximation to the Normal distribution, or to use a search function. We will be doing more of this later.

Classifier Error

The classifier error of any classifier is simply the probability the classifier make a mistake. This is always the sum of the probabilities of the possible errors, where an error is choosing a class when the the sample comes from another class.

$$
P_e = \sum_j \sum_{i \neq j} P(\omega_j | x \in \omega_i)
$$

where ω_j is the class chosen.

In this instance, this can be written as:

$$
\begin{aligned}
P_e = {} & p(x < 0.37|x \in \omega_2)P(\omega_2) + p(x < 0.37|x \in \omega_3)P(\omega_3 + p(x > 0.37|x \in \omega_1)P(\omega_1) + \\
& p(x < 0.62|x \in \omega_3)P(\omega_3) + + p(x > 0.62|x \in \omega_1)P(\omega_1) + p(x > 0.62|x \in \omega_2)P(\omega_2)
\end{aligned}
$$

$$\begin{aligned} P_e &= N_{cdf}(0.37, 0.47, 0.002)(0.33) + N_{cdf}(0.37, 0.800, 0.004)(0.33) + 1 - N_{cdf}(0.37, 0.20, 0.004) \\ & N_{cdf}(0.62, 0.800, 0.004) + 1 - N_{cdf}(0.62, 0.20, 0.004) + 1 - N_{cdf}(0.62, 0.470, 0.002) \\ &= 0.005 + 0.001 + 0.0007 + 0.0009 = 0.0082 \end{aligned}$$

Using octave, you could find this sum with:

$$\begin{aligned} P_e &= normal_cdf(0.37, 0.47, 0.002) * 0.33 + (1 - normal_cdf(0.37, 0.20, 0.004)) * 0.33 + \\ & normal_cdf(0.37, 0.80, 0.004) + (1 - normal_cdf(0.62, 0.47, 0.002)) * 0.33 + (1 - normal_c \\ &= 0.0055 \end{aligned}$$

x_2

The same analysis can be carried out for x_2, which has thresholds at $x_2 = 0.093$ and $x_2 = 0.165$. The error for this classifier is $p_e = 55555$

x_3

Since x_3 is a binary variable, we can't use a Normal distribution approximation. If you are familiar with statistics, you have probably heard of the Normal approximation for the binomial distribution, but that it the distribution of the number of successes in n trials. This doesn't apply when you are simply attempting to classify based on the sample producing a one or a zero. Bayes Law doesn't apply well to this situation because the class conditional distributions are not Normal and don't even have strong central tendancy. It really isn't necessary though, since we are really interested in:

$$g(x_3) = \max_{j=1}^{3} p(\omega_j|x)$$

x_3 can only be 0 or 1, and for each class, there is a different probability. We have the following:

Class	P(0)	P(1)
ω_1	0.07	0.93
ω_2	0.53	0.47
ω_3	0.88	0.12

So if $x_3 = 0$, the highest conditional probability is for ω_3, and if $x_3 = 1$, the highest conditional probability is for ω_1, so are classifier is evident.

The error for this classifier is, as usual, the probability of making a mistake. If $x_3 = 0$, the probability that the sample came from class 1 or 2 is:

$$\begin{aligned} p(error|x_3 = 0) &= p(x_3 = 0|\omega_1)P(\omega_1) + p(x_3 = 0|\omega_2)P(\omega_2) \\ &= 0.07 \cdot 0.33 + 0.53\ cdot0.33 \\ &= 0.20 \end{aligned}$$

$$\begin{aligned} p(error|x_3 = 1) &= p(x_3 = 1|\omega_2)P(\omega_2) \ + \ p(x_3 = 1|\omega_3)P(\omega_3) \\ &= 0.47 \cdot 0.33 \ + \ 0.12 \ cdot0.33 \\ &= 0.19 \end{aligned}$$

So the classifier error is $0.20 + 0.19 \ = \ 0.49$.

So the best single variable classifier is x_1 with an expected error of 0.0082.

Bivariate Classifiers

For bivariate classifiers, there are three choices:

$$\begin{vmatrix} x_1 \\ x_2 \end{vmatrix},$$

$$\begin{vmatrix} x_1 \\ x_3 \end{vmatrix}$$

$$\begin{vmatrix} x_2 \\ x_3 \end{vmatrix}$$

For the first case, we could use the bivariate normal distribution, but for the other two cases that wouldn't work. We could look for distributions that represent a mixture of a Normal and a binomially distributed variable, but that is an ugly problem. So we will resort to a standard form of classifier, a linear classifier. For x_1 and x_2, we can use one of the forms discussed in the book for linear classifiers for Normally distributed variables.

There are three cases: all variables have the same variance for all classes and there are no cross-correlations, all classes have the same covariance matrix and the classes have different covariance matrices. From before, the variances are:

	Variances		
Class	x_1	x_2	x_3
1	0.004	0.003	0.066
2	0.002	0.003	0.252
3	0.004	0.005	0.107

It is not too much of a stretch to assume that the variances for all the variables are the same for all classes. Typically, if you want to make such an assumption, you use the average variance over all classes. In this case, let $\sigma^2 \ = \ 0.0036$ which is the average variance over all classes and both variables.

We will need three discriminant functions, one for each pair of classes. The form of the linear discriminant is $\mathbf{w}_i^t\mathbf{x} \ + \ w_{i0}$, where:

$$\begin{aligned} \mathbf{w}_i &= \frac{1}{\sigma^2}\mu_\mathbf{i} \\ w_{i0} &= -\frac{1}{2\sigma^2}\mu^t\mu \ + \ \ln P(\omega_i) \end{aligned}$$

Since the classes are equiprobable, you can ignore the $p(\omega_i)$ term. So,

$$\begin{aligned} g_1(\mathbf{x}) &= \frac{1}{0.0036}\begin{vmatrix} 0.199 \\ 0.058 \end{vmatrix} + -\frac{1}{2 \cdot 0.0036}\begin{vmatrix} 0.199 \\ 0.058 \end{vmatrix}^t + \begin{vmatrix} 0.199 \\ 0.058 \end{vmatrix} \\ g_2(\mathbf{x}) &= \frac{1}{0.0036}\begin{vmatrix} 0.470 \\ 0.119 \end{vmatrix} + -\frac{1}{2 \cdot 0.0036}\begin{vmatrix} 0.470 \\ 0.119 \end{vmatrix}^t + \begin{vmatrix} 0.470 \\ 0.119 \end{vmatrix} \\ g_3(\mathbf{x}) &= \frac{1}{0.0036}\begin{vmatrix} 0.800 \\ 0.199 \end{vmatrix} + -\frac{1}{2 \cdot 0.0036}\begin{vmatrix} 0.800 \\ 0.199 \end{vmatrix}^t + \begin{vmatrix} 0.800 \\ 0.199 \end{vmatrix} \end{aligned}$$

Using the form:

$$\begin{aligned} g(x) &= \mathbf{w}^t(\mathbf{x} \ - \ x_0) \\ \mathbf{w} &= \mu_i \ - \ \mu_j \\ w_0 &= \frac{1}{2}(\mu_i \ + \ \mu_j) \end{aligned}$$

$$\begin{aligned} g_{12}(\mathbf{x}) &= \begin{vmatrix} -0.274 \\ -0.060 \end{vmatrix}^t (\mathbf{x} \ - \ \frac{1}{2}\begin{vmatrix} 0.665 \\ 0.177 \end{vmatrix}) \\ g_{13}(\mathbf{x}) &= \begin{vmatrix} -0.604 \\ -0.141 \end{vmatrix}^t (\mathbf{x} \ - \ \frac{1}{2}\begin{vmatrix} 0.996 \\ 0.257 \end{vmatrix}) \\ g_{23}(\mathbf{x}) &= \begin{vmatrix} -0.330 \\ -0.080 \end{vmatrix}^t (\mathbf{x} \ - \ \frac{1}{2}\begin{vmatrix} 1.270 \\ 0.318 \end{vmatrix}) \end{aligned}$$

Where these functions are zero define lines which separate the classes. For example,

$$\begin{aligned} g_{12}(\mathbf{x}) &= -0.274\, x_1 \ - \ 0.060\, x_2 \ + \ 0.193 \ = \ 0 \\ &\rightarrow \ x_1 \ = \ -0.22 x_2 \ + \ 0.347 \end{aligned}$$

Similarly for g_{13} and g_{23}

Classifier Error

The error for this classifier can't be calculated using the class conditional probabilities since we don't have them (ostensibly). What you would do is process your training sample to see how many time you get the wrong sign from your discriminant function. The fraction of incorrect values is the probability of an error.

Using this command to count the number of values less than zero:

$$length(find(docalc(w12, [x11, x12], x0) < 0)$$

We find that for $(x_1, x_2|\omega_1)$ there were two negative values out of 100, and for $(x_1, x_2|\omega_2)$ there was 1 positive value out of 100, so the error rate is $3/200 = 0.015$. This is actually higher than the single variate classifier ($P_e = 0.0082$), but given the difference in the approach and the error calculation, close enough to be considered the same. However, the same error indicates that using a single variate classifier is better.