DATA 1010
In-class exercises
Samuel S. Watson
07 November 2018

Problem 1

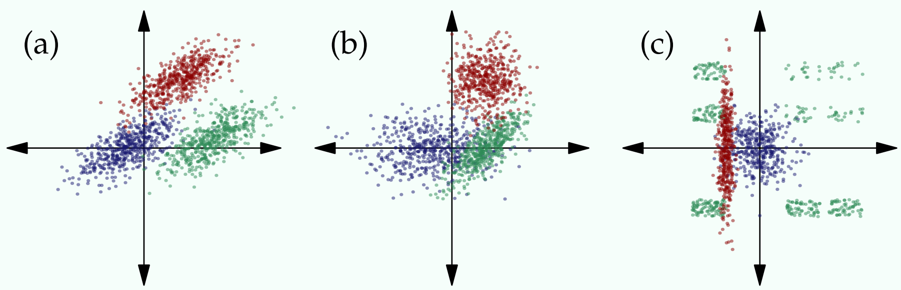
Suppose that f_1, f_2 are distinct normal densities and $p \in (0,1)$. Is $x \mapsto pf_1(x) + (1-p)f_2(x)$ a normal density?

Solution

No. One way to see this is not *generally* true is to consider the case where f_1 and f_2 have means which are very far apart. In that case, the graph of $pf_1(x) + (1-p)f_2(x)$ has two humps, which means that it cannot be Gaussian. In fact, $x \mapsto pf_1(x) + (1-p)f_2(x)$ is never Gaussian if f_1 and f_2 are not equal.

Problem 2

Match each data set with the best-suited model: Naive Bayes, LDA, QDA.



Solution

The correct order is CAB, since the third plot shows class conditional densities which factor as a product of marginals, the first plot shows Gaussian class conditional probabilities with the same covariance matrices, and the second plot shows Gaussian class conditional probabilities with distinct covariance matrices.

Problem 3

Consider the problem of classifying a US man or woman as a man or a woman based on their height. Suppose that the distribution of heights among women is normal with mean 63.7 inches and standard deviation 2.7 inches, and that the distribution of heights among men is normal with a mean of 69.1 inches and a standard deviation of 2.9 inches. The proportion of men or women in the US who are women is 50.8%, and the proportion of men is 49.2%.

(a) Based on the model specified above, is the distribution of heights among all men and women normal?

(b) Find the Bayes classifier for this example (the classifier with minimal error rate).

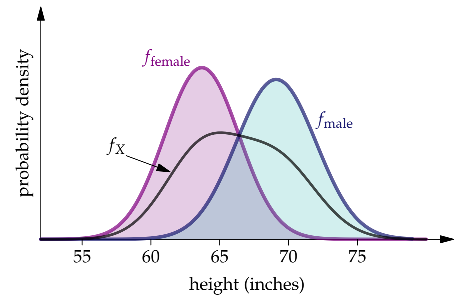
(c) What is the misclassification probability for the Bayes classifier?

Solution

(a) The distribution of heights has PDF

$$x \mapsto 0.508 f_{63.7,2.7}(x) + 0.492 f_{69.1,2.9}(x),$$

which is not a normal density (one way to see this is to plot it and look at the shape).



(b) The Bayes classifier maps a height x to `man` or `woman` according to whether $0.492 f_{69.1,2.9}(x)$ or $0.508 f_{63.7,2.7}(x)$ is larger. We can find the crossover point by setting these equal and solving:

```
using Roots, Distributions
woman_dist, man_dist = Normal(63.7,2.7), Normal(69.1,2.9)
woman_prop, man_prop = 0.508, 0.492
find_zero(x-> woman_prop*pdf(woman_dist,x) - man_prop*pdf(man_dist,x),(63.7,69.1))
```

We find that the Bayes classifier returns `male` for heights larger than 66.45 inches and `female` for lower heights.

(c) The probability of misclassification is

$$\mathbb{P}(\text{man, shorter than 66.45 inches}) + \mathbb{P}(\text{woman, taller than 66.45 inches}).$$

We can calculate this as `man_prop*cdf(man_dist,66.45) + woman_prop*(1-cdf(woman_dist,66.45))` = 16.7%.