Homework 9: Lying, Cheating, and Mixture Models

36-462/662, Spring 2020

Due at 10 pm on Saturday, 28 March 2020

AGENDA: Practice working with mixture models, a.k.a. probabilistic clustering.

Mixture models where all the observed variables are categorical are often called **latent class models**, and the package `poLCA` provides functions for fitting them through the EM algorithm. The basic syntax is

```
poLCA(cbind(Var1,Var2,Var4)~1, data, nclass=k)
```

to fit a k-cluster model for the columns named `Var1`, `Var2` and `Var4`. Features are assumed to be independent within each cluster. The returned object contains a lot of information, including the distribution over each feature variable for each cluster, the probability or mixing weight for each cluster, and the posterior probability for each observation having come from each cluster.

The package `poLCA` also contains a data set, `cheating`, which is the result of a survey of about 300 undergraduates on cheating in college. Four variables record, for each student, whether they had ever lied[1] to get out of an exam; lied to avoid turning in a paper on time; bought a term paper or obtained a copy of an exam in advance (together, "fraud"); or copied answers on an exam from another student. The data set also records the students' GPAs, discretized into five categories. (See `help(cheating)` for details.)

1. (10) *Online questions* are online, due at the same time as the homework, and help orient you towards next week and its homework.
2. Load the data.
   (a) (5) What are the correlations between the four forms of cheating? (Don't list them all, use a table or graph.)
   (b) (3) What fraction of students have cheated at least once? What fraction of cheaters engage in multiple forms of cheating?

[1] More precisely, whether they *said* they had ever lied, etc., etc.

(c) (3) What fraction of students have lied to get out of an exam? What fraction of students have bought a term paper or snuck a look at an exam before taking it? What fraction of students who have lied to get out of an exam have committed fraud?

3. Fit a latent class model with two classes or clusters.

   (a) (3) For each class, what are the probabilities of each form of cheating? What is the probability of each class? (Again, use a table or graph.)

   (b) (4) For each cluster, use the estimated parameters of the model to find the probability that a member of the cluster has engaged in at least one form of cheating. *Hint:* It may be easier to first find the probability that they have not cheated in any way.

   (c) (5) For each cluster, find the conditional probability that a member of the cluster who has engaged in at least one sort of cheating has engaged in multiple forms of cheating. Again, use the estimated parameters of the model, not a new model. *Hint:* It may be easier to first find the probability that someone has cheated *exactly* once.

   (d) (8) Describe, in words, how the two classes differ from each other. Justify your description by referring to your numerical results.

4. *Conditioning*

   (a) (3) Find the probability (according to the model) that a random student has committed fraud.

   (b) (7) Suppose we know that a student has lied to get out of an exam, but not whether they have engaged in any other form of cheating. Find the conditional probabilities of the student being in each class. *Hint:* Bayes's rule.

   (c) (7) Find the probability (according to the model) that a student who has lied to get out of an exam has also committed fraud. *Hint:* Bayes's rule, and the law of total probability.

   (d) (7) Your answer for 4c should be several times larger than your answer for 4a. Explain how this is compatible with the fact that for both classes of student, lying to get out of exams is statistically independent of fraud.

5. *Cross-validation will continue until morale improves* We will use five-fold cross-validation of the log-likelihood to pick the number of clusters. *Hint:* §17.4.4 of *Advanced Data Analysis from an Elementary Point of View*, but remember that the features here are binary, so Gaussian distributions won't work, and some of the code given there is inapplicable (see below).

   (a) (5) Show, in math, how to calculate the probability that a latent class model assigns to an particular data point.

(b) (5) Provide comments for the following function, explaining the overall purpose of the function, what all the inputs are, what the output(s) are, how it relates to the math you just did in the previous part, and what each piece of the code does.

```
dmultbinarymix <- function(x, model, offset = 1, log = FALSE) {
    x <- x - offset
    prob.matrix <- sapply(model$probs, function(mat) {
        mat[, 2]
    })
    if (is.null(dim(prob.matrix))) {
        prob.matrix <- array(prob.matrix, dim = c(1, length(prob.matrix)))
    }
    class.probs <- model$P
    class.cond.prob <- function(x, c) {
        class.probs[c] * prod((prob.matrix[c, ]^x) * ((1 - prob.matrix[c, ])^(1 -
            x)))
    }
    one.point.prob <- function(x) {
        summands <- sapply(1:length(class.probs), class.cond.prob, x = x)
        return(sum(summands))
    }
    probs <- apply(x, 1, one.point.prob)
    if (log) {
        return(log(probs))
    } else {
        return(probs)
    }
}
```

(c) (5) Write a function which takes as its arguments an estimated latent class model and a data set, and returns the log-likelihood which the model assigns to the new data set. (This function should call `dmultbinarymix`, and should *not* call `poLCA`.) Check that this is working by seeing that it matches the log-likelihood returned by `poLCA` when run on the training data, and that it doesn't give exactly the same answer when you change the data.

(d) (5 Write a function to do v-fold cross-validation to pick the number of clusters. How many does it pick here, with $v = 5$?

(e) (5) Plot the cross-validated log-likelihood against the number of clusters.

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or

commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.

Extra credit (5 points): The data set also contains information on the GPA of the students in the survey. Suppose we pick $k = 2$ clusters. Describe, mathematically, how we could use the GPA information to estimate the probability of a student belonging to one class rather than another, *without* having to ask them about cheating. For full credit, implement your idea in code. How would you know whether it was working or not?