BME 335

Spring 2016

Project I

Objective

The objective of this project is to apply concepts of probability and statistics to drive and justify decisions in a design problem. Specifically, you will design a method for classifying entities (in this class T cells) in groups according to measurable properties subject to random variability.

Introduction

T cells are an important component of the adaptive immune response. There are several varieties of T cells, such as helper T cells, regulatory T cells, etc. T cells from biological samples are normally characterized in terms of molecular markers they express, typically cell surface proteins. Different T cell subsets express markers in different proportions or different markers altogether (for additional background see supporting slides on Canvas).

Consider two hypothetical cell types, T1 and T2, and two markers, A and B. It has been determined the following empirical PDFs, f1 and f2, represent well the distribution of surface markers in T1 and T2 cells respectively.

$$f1_{AB}(a,b) = \alpha(0.05 + a^2)[(b-1)^4 + 0.025] \quad (\alpha=11.6, \beta=5.3)$$
$$f2_{AB}(a,b) = \beta(1-a^2)\left(0.05 + b^4 + \frac{a^2b^2}{2}\right)$$

In the expressions above, random variables a and b are normalized surface concentrations of markers A and B respectively. Therefore $0 \le a \le 1$ and $0 \le b \le 1$. The PDFs are zero outside these ranges.

Our goal is to design an algorithm that can classify cells in a mixture as T1 or T2 based on an automated measurement of A and B performed by a flow cytometer. The idea is that such algorithm can be incorporated in analytical instruments for generating a preliminary diagnostic of medical conditions that may manifest themselves as abnormal abundance of a particular immune cell type.

For simplicity we will assume the cell mixture contains only T1 and T2 cells.

The general strategy will be to define events in the (A,B) space with probabilities that depend on whether cell is T1 or T2. To do this we will divide the (A,B) plane into non-overlapping quadrants (the events), Qx (x=1, 2, 3, or 4), not necessarily equal, corresponding to (H,H), (H,L) ,(L,H), and (L,L), where H and L represent high or low marker concentration respectively. Because the quadrants represent a partition, measuring (a,b) for a given cell will result in the occurrence of one event. Based on the likelihood of the event for the different cell lines we can use a Bayesian approach to infer the cell type.

You are not expected to write a program that implements the algorithm, but it should be clear from your report how the classifier could be implemented. Critically, numerical parameters (e.g. quadrant boundaries) and detailed expressions should be provided.

Procedure

The following is a proposed approach for developing the classifier. You are welcome to perform additional steps or even follow your own path and remove steps that don't fit in your grand design plan (see "*Evaluation criteria"*).

In broad terms the suggested procedure is as follows:

1. Familiarize with the PDF's and CDF's to get some intuition of how probability of observing (a,b) is distributed in T1 and T2 cells.

2. Determine thresholds A_T and B_T, for markers A and B respectively, that define quadrants such that the probability density of (a,b) is concentrated in different quadrants for T1 and T2 cells.

3. Analyze the pros and cons of a non-Bayesian classifier that assigns the cell based on the quadrant corresponding to its (a,b) values.

4. Define and characterize a Bayesian classifier that uses posterior probabilities and takes into account prior information to make a cell type assignment.

5. Provide a design specification document with concise but detailed enough instructions for the programmer that must program this.

6. Optional: Test the classifier with simulated data (provided) or provide a specification document.

You don't have to follow this workflow exactly, especially if you have better ideas of how to design the classifier and/or evaluate its performance; but remember, what you decide to do has to show clear evidence that you understand the relevant concepts of probability and know how to apply them.

1. Preliminary study.

Get to know the PDF's. The primary goal of the preliminary study, is to assess the feasibility of the project and in particular get a sense of where (A,B) regions useful for cell type discrimination may lie (or if they even exist).

We suggest you use your favorite math software (Matlab, Mathematica, R, etc.) to plot the joint PDFs for A and B in T1 and T2 cells as surface or contour plots. These plots should give you a good idea of how the probability density for the two markers is distributed in the different cell types.

SUGGESTION: At this stage you may also find useful plotting the CDFs and marginal PDFs for A and B in T1 and T2 cells, although these may be a bit tricky to interpret.

2. Defining quadrants.

In a typical experiment, a continuous stream of cells passes rapidly through a detector that can accurately measure the levels of a and b on the cellular surface. Our goal is to, based on these two numbers, determine the likelihood a cell is T1 or T2. For this to be feasible, we need to identify combinations of (a,b) values that appear with high probability in T1 cells but are rare in T2 ones and vice versa. During the preliminary studies you may have identified good candidate areas. To simplify the problem, we will consider only rectangular areas, arranged as quadrants, representing a partition of the (a,b) sample space.

At this stage you must make your first design decision: selecting the boundaries of the quadrants and justify your choice in terms of probabilistic arguments. Because quadrants are contiguous all you need are threshold values A_T and B_T for A and B respectively. To do this we suggest you look at the PDF's you plotted in the preliminary studies and ask which values of A_T and B_T result in quadrants such that a large portion of the probability density of (a,b) in T1 cells is concentrated in one quadrant and a large portion of the probability density of (a,b) in T2 cells is concentrated in another one. You must decide what "a large proportion" is. What are going to be the implications of this decision down the road?

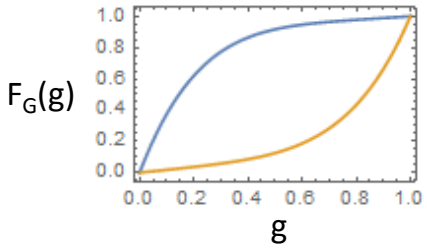

Figure 1: Marginal CDF for an hypothetical marker G in two hypothetical cell lines C1 (blue) or C2 (yellow) cells.

To justify your choice in quantitative terms it may be useful to look at the marginal PDF and both marginal and joint CDFs for A and B. For example, in the marginal CDFs in figure 1, we see that in –hypothetical- C1 cells (blue), values of b between 0 and 0.4 represent a cumulative probability of 0.8, that is 80% of C1 cells will have b<0.4 whereas almost 90% (see yellow line) of C2 cells will have b>0.4. You can use this kind of argument to justify your threshold choice, but we suggest you pay attention to the joint CDF as well to avoid surprises.

At the end of this process, you should have four quadrants Q_x (x=1, 2, 3, or 4). Indicate which quadrant(s) are associated with T1 cells and which one(s) are associated with T2's. Notice that some quadrants may be undecidable (are either highly probable or very unlikely for both cell types) and a cell type may be represented by more than one quadrant. Keep this in mind as you design your classifier.

3. Basic quadrant-based classifier.

Your values for A_T and B_T define four quadrants Q_x (x=1, 2, 3, or 4), which we can considered as four events (in the probabilistic sense) in terms of the levels of (a,b) present in a cell. The probability of observing each event depends on whether the cell is T1 or T2, a fact you used to associate quadrants with cell types above. Let Q_{T1} and Q_{T2} be the quadrants you associated with T1 and T2 cells respectively. If you associated more than one you can aggregate them.

Now, if we measure (a,b) in a cell and observe one of the Qx, we can assess what type the cell is (or is not) based on the relatively likelihood of observing that quadrant in either cell type. This is already a basic classifier. To assess the quality of this classifier, calculate the conditional probabilities P(Qx|T1) and P(Qx|T2) for x=1,2,3,4 (your quadrants). How do you interpret these probabilities?

How could you use these probabilities to perform a more rational threshold selection?

To get a better sense of the quality of the classification scheme, consider 100 T1 cells and calculate how many would fall on each quadrant. How many of these would be misclassified? Repeat for T2. How many T1 and T2 cells would remain unclassified (that is fall in a quadrant that you could not clearly associate with T1 or T2 status)?

How good/or bad would your classifier perform (how many errors will you observe) for samples with 50/50, 80/20 or 20/80 T1/T2 ratios?

SUGGESTION: The concepts of precision (cells correctly classified) and recall (cells your method fails to classify) may be useful to discuss the performance of your classifier.

4. The Bayesian classifier.

The classifier above, does not give you the probability of a cell being T1 or T2. It just decides for example that if Q_{T1} is observed, then the cell is T1. We don't have any indication of how much confidence we should place in the outcome. Furthermore, the number of classification errors depends on the relative abundance of T1 and T2 in the sample (why?).

To make a more objective cell-type assignment that takes into account the composition of typical samples, we will use the conditional probabilities P(T1|Qx) and P(T2|Qx), where Qx is the quadrant corresponding to the measured (a,b) values for the cell under consideration. To calculate these "inverse" probabilities we can use the Bayes formula. However, this approach requires a prior probability. We could use the relative abundances of T1 and T2 cells typical in a sample, we could let the user guess it, or we could use an uninformative prior that considers a cell is equally likely to be T1 or T2. Your next design decision is to pick one of these strategies.

To define the Bayesian classifier, write the expressions for the conditional probabilities P(T1|Qx) and P(T2|Qx) without assuming a specific T1/T2 ratio. Now use these probabilities to evaluate the performance of the classifier under different scenarios: e.g. samples with 50/50, 80/20 or 20/80 T1/T2 ratios.

How good is the Bayesian classifier? You can qualify this by calculating the probabilities of misclassification $P(T1|\ Q_{T2})$ and $P(T2|\ Q_{T1})$ or by reporting the fraction of T1/2 cells that are misclassified on each scenario. What would happen if you use the wrong prior (that is you assume p(T1)=0.8 but in your sample it is actually 0.2)?

SUGGESTION: You could test this by calculating how a classifier with prior p(T1)=0.8 would classify 20 T1's and 80 T2's.

For our application, errors in which a cell is classified as the wrong type are much worse that errors in which cells remain unclassified. In the immortal words of Lisa Simpson "*Remember, it is better to remain silent and be thought a fool than open your mouth and remove all doubt*". Let's look at the errors for T1 cells. Consolidate the quadrants you did not associate with a cell type (if you have more than one of those) into a Qu ("unclassified") and calculate P(Qu|T1). Repeat for T2. What are these probabilities telling you about the rate of misclassification (assigning the wrong type) and undecidability (failure to assign a type)?

5. Design implementation instructions

Someone will have to program this. Provide a concise set of instructions for the programmer that will implement your classifier in a flow cytometer device. We suggest you do this by providing step-by-step instructions of how would you, starting with a (a,b) value, assign a cell type. Assume the programmer does not care about the reasons behind your design choices; just needs to know how to implement the classificatory. Indicate, briefly, the type of errors you expect when testing the app. The document should not be longer than 1 page.

6. Optional: Test run

The attached file has simulated flow cytometry data for biomarkers A and B for 750 cells. Cells have been assigned an identification number from 1-750. Cells have also been classified in terms of shape into round (alpha) or elongated (beta) cells. The file contains the following columns: cell id number | cell morphology | A conc. | B conc.

We will use these data to test the classifier. We know cell morphology is a good predictor of cell type. We will use morphology data to test the output of the classifier.

To test the classifier we first need to clean the data.

Because of experimental error, some cells were assigned values of A or B outside of the 0-1 range. Using your favorite math software find and eliminate those cells (hint: check function *find* in Matlab).

Plot a scatter plot of A vs B. Use red dots for alpha and blue dots for beta cells. Make sure your plot axes include the whole physiological range (0-1) for both markers. Indicate was cell shape are likely T1's and which ones T2's. Justify.

Using the "quadrants" assign each cell to type 1 or type 2. Indicate in a table the total number of T1, T2, and unknown cells for the total sample. Assuming cell morphology accurately reflects cell type, quantify the errors made by the classification scheme.

Now we are going to test the Bayesian classifier. Use the morphology data to estimate the fraction of T1 and T2 cells. Use this ratio to estimate the prior probabilities.

For each cell, calculate P(T1|Q(a,b)) and P(T2|Q(a,b)), where Q(a,b) is the quadrant corresponding to the measured (a,b) values.

Use these probabilities to make an assignment. Compare with the assignments made based on the morphology data and those using the simple quadrant-based scheme.

Ideas for discussion and improvement

The questions below are there to give you some ideas for discussion. Do not answer them in the report.

- Why limiting the partition to 4 quadrants only?
- Could we get rid of quadrants and relay on P(T1|a,b) and P(T2|a,b) directly to make cell type assignments?
- A potential problem with the algorithm is bias; it may end up misclassifying more T1 cells as T2, than is mistaking T2 for T1, or if it missing more T1 cells as "unclassified" than T2's or vice versa.
- Should we stick to non-informative priors (assume we know nothing about the sample composition; that is each given cell has .5 probability of being T1 and .5 probability of being T2?
- What is the Bayesian classifier bringing to the table than the simpler approach in step 3 does not do?
- How would you use the conditional probabilities $P(Q_X|T_Y)$ to select optimal thresholds?

Report

The report should include the following sections:

Cover page with Honor pledge

The first page of the report should include your name, the names of your team mates, and the title of your project. The cover page should also include the honor pledge below and your signature indicating you accept to abide by it.

"As a student of The University of Texas at Austin, I shall abide by the core values of the University and uphold academic integrity."

Introduction to the design problem and strategy

Explain the problem you are trying to solve and the general strategy you used to solve it (~2 paragraphs).

Classifier design process

Describe you design process and justify each design decision (even those made for you, for example using 4 quadrants) using concepts from probability such as conditional and marginal probabilities, cumulative probability, etc. Includes figures as necessary to support your points. Figures must be labeled, numbered, and have a short caption indicating what are we looking at. Figures must be referenced in the text (indicate what you are learning for the figures you select to include). Do not include superfluous figures. All probability calculations should be shown explicitly together with the corresponding expressions. All claims (e.g. "we find 80% of T1 cells have b concentrations lower than xyz". You must show how you calculate that and/or the figure you are using to get that number from. Do not "estimate", calculate! At each stage, indicate the performance metrics you are using and quantify the errors the classifier is likely to make.

This section should not be longer than 4 pages + figures.

OPTIONAL: Simulated data results.

Discussion

Discuss the performance of the classifier in its various incarnations, including possible pitfalls and improvements (~ 2 or 3 paragraphs)

Conclusion

Short paragraph summarizing the outcome of your project.

Design implementation document

As described in (5).

IMPORTANT:

Students must write their own reports. Similar figures are expected for members of a group, but wording should be your own. Generating "customized" group reports with slightly tweaked wording will likely be flagged as plagiarism by Canvas. Plagiarism or any other violations of academic integrity will be reported to SJS.

Evaluation criteria

The overarching evaluation criteria are: demonstrates understanding of the class material and proficiency applying it to a specific biomedical engineering scenario. The grade for this project will be based on your report. It is therefore in your best interest the report is clearly written, well organized, complete, and coherent. It has be clear how and why you went from point 1 to point 2, 3, etc. We will use the rubric below:

Organization of the report (0-5)

5- Well organized/presented. The report follows a clear logic path. Explanations are sufficient and clear. Included figures are sufficient to support the work. Figures are numbered and labeled, and are referenced appropriately in the text. All required sections are present.

0- Poorly organized and/or incomplete. Logic path is unclear and/or there are significant logic gaps. Explanations are insufficient to understand the work. Key figures are missing, mislabeled, or not numbered. Sections are missing or incomplete.

Understanding of the topics (0-5)

5- Reflects good understanding of the class material, in particular PDF's, conditional probability, marginal probability, expectation and variance, and Bayesian approaches. This is evidenced in the report by correct calculations, results that are consistent, use of error checks, and adequate assessment of intermediate results. An integral understanding of the material is clearly present in the report.

0- No evidence of understanding of the class material, in particular PDF's, conditional probability, marginal probability, expectation and variance, and Bayesian approaches. Incorrect calculations (e.g. using wrong expressions), results inconsistent with data or concepts from probability, failure to check for errors, and/or gross misinterpretation of intermediate results.

Application Area (0-5)

5- Demonstrates ability to transfer concepts from probability to the solution of a real-life problem. The student clearly connects elements of the engineering problem with concepts from probability and uses these concepts to formulate a solution. The connections are explicitly indicated in the report. Assumptions are explicitly indicated and are adequate. Overall results are interpreted in the context of the original problem and caveats or limitations discussed. Improved or alternative approaches are proposed and justified.

0- Does not demonstrate ability to transfer concepts from probability to the solution of a real-life problem. Student fail to frame the problem in terms of probability concepts or does it incorrectly. Assumptions are made but are not explicitly discussed. Incorrect or inadequate approach to the solution of the engineering problem. Lack of interpretation of the overall results in the context of the original problem and caveats or limitations not discussed. Improved or alternative approaches are not proposed or proposed but not justified.

We encourage original thinking and creative solutions, but keep in mind your work still need to demonstrate knowledge of the topics and ability to apply it.