Flow Cytometry and Machine Learning

Patrick Cheng

Applying Machine Learning to Medicine

1.0 Introduction

Medicine is an age old field, dating back to the days Aristotle theorized that the origin of disease was an imbalance of blood, phlegm, and bile. Computers have already rapidly changed the world. Can the relatively new field of machine learning change medicine?

1.1 White Blood Cells

The human body is made of trillions and trillions of cells. Groups of cells can work together to accomplish complicated things as organs, like making a beating heart or a thinking brain. The group of cells in our blood fulfill similarly specialized roles. Red blood cells supply oxygen to the body and carry away carbon dioxide, the byproduct of living. There is also a population of cells that make up our body's immune defense that are collectively called *"white blood cells"* or WBCs.

The vast majority of WBCs develop normally in bone marrow and go on to perform their vital jobs with no problem. Some cells, however, don't grow up as expected. In leukemia, a type of cancer, WBCs grow up mutated and begin to multiply more than usual. In these cases, not only do they not do their jobs properly but they also take resources from normal cells which in turn cannot perform their jobs properly. Doctors can treat leukemia with chemotherapy, but first they must identify the cancer, which is difficult because leukemia oftentimes looks and feels like other diseases. To make an accurate diagnosis, they run various laboratory tests.

One of the best laboratory tests is called *"flow cytometry."* Flow cytometry measures thousands of cells by measuring how much of certain proteins are in each cell. This can be thought of as the different traits of a person.

For example, if you were to measure a firefighter, a soldier, and a ballerina:

1. The firefighter and soldier's shoes weigh much more than a ballerina's
2. The firefighter and soldier's helmets weigh more than a ballerina's
3. The firefighter's carries a hose, whereas neither the soldier or ballerina does

It is easy to determine that among the three, a person with no helmet or hose but light shoes is a ballerina!

It turns out for human cells, the process is very similar. You can determine all sorts of a cell's characteristics by looking at the cell's protein levels. You can also detect the presence of cancer cells because they do not follow established patterns, like a person that wears no clothing but carries around 100 pounds of firefighting hose.

1.2 Flow Cytometry

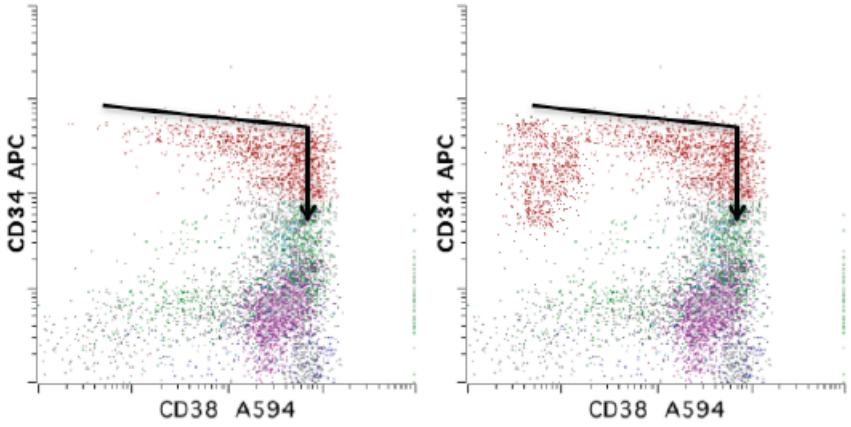
The proteins levels that flow cytometry measures are called names like "CD34" or "CD38", where CD stands for "conserved domain". This is exactly what they are- things about the cell that stay the same from cell to cell in the same category. If you look at the flow cytometry results for a normal growing white cell (figure 1), they may gradually gain in "CD38" fluorescence until it matures, at which point "CD34" protein levels drop off greatly. So CD38 is a marker for maturity, while CD34 is a childhood tag of sorts, that the baby white blood cell loses when they reach maturity.

It stands to reason that white blood cells usually have one or the other, or both markers. Cells with low levels of both CD38 and CD34 (figure 2) are highly suspicious and may signal a mutant population of cells that would warrant further investigation.

A typical flow cytometry readout may be time consuming for a pathologist to interpret for even the most straight forward cases. The main crux of the issue is that the doctors must examine only 2 protein markers at a time as this is the most digestible format for the human brain. For a panel of 30 markers, looking at every combination would mean examining several hundred different plots. Where humans have trouble comprehending more than 2 dimensions at a time, computers excel at it. Not only would the right model possibly save time, it might also be more accurate and consistent.

The cost savings standpoint is compelling; the cost of a pathologist is roughly $1 per minute, and the most straightforward flow cytometry may take 5-10 minutes to read, for the more than 150,000 new cases of leukemia per year, an automated flow reader would save around $750,000 to $1,500,000 per year.

FIGURE 1. Typical flow cytometry results for normal (left) and leukemic (right) patients



This does not count negative cases, periodic retesting for relapse, and retesting for treatment monitoring, which in reality account for a significant number of flow cytometries. Not to mention, flow cytometry is used for many other types of molecular classification and this method can easily be adapted across other uses as well.

2.0 The Dataset

There is a publicly available flow cytometry repository at flowrepository.org. The dataset used in this study is a reference dataset for a commercial laboratory containing 359 people, including 43 patients suffering from acute myeloid leukemia (AML), a subtype of leukemia. The AML diagnoses were determined by the current manual interpretation methods.

The data is stored in an FCS format and can be converted into CSV for more ease of analysis in python. The module to do this was located at: https://genepattern.broadinstitute.org/gp/pages/login.jsf and the code for the module is located at: https://github.com/genepattern.

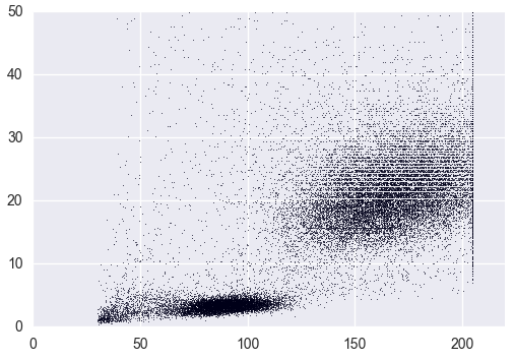
The dataset is packaged such that every patient has 7 corresponding tubes. In each tube, a different combination of 5 protein markers (the CD proteins discussed earlier) are measured for roughly 30,000 cells. The measure of the proteins is in units of fluorescence; the more of the protein is present, the brighter it lights up. (see appendix: how does flow cytometry work?) The fluorescence levels range from 0 (no protein present) to 1015 (lots of protein).

In summary, there are 359 patients, with 7 tubes of 30,000 cells each.

2.1 Exploring and Visualizing the Dataset

A scatter plot of the first patient's first two markers results in the scatterplot below.

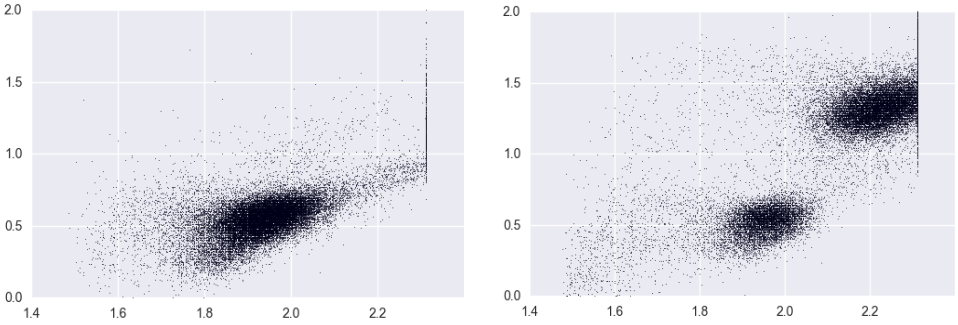
FIGURE 2. Scatterplot of patient 1, markers 1 and 2



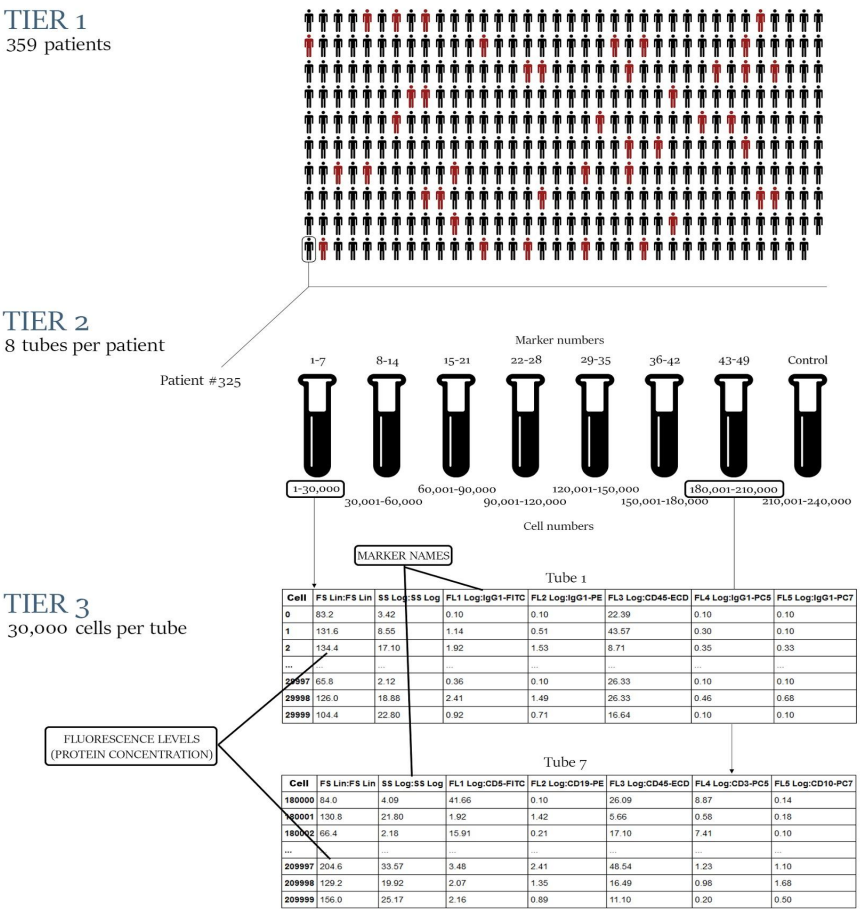
The dense blob of cell in the lower left seems like it has structure to it, but is vaguely diffi-

cult to define. If a log 10 transformation is applied there is better resolution. The same patient from above is shown on the left, compared with a leukemic patient on the right.

FIGURE 3. Scatterplot of patient 1 (left, healthy) and patient 7 (right, leukemic) with log transformation applied to protein levels



The log transformation spreads the data out much more evenly between a range of 0 to approximately 3, as the range of signal for these markers is 0 to 1015. There is also a very

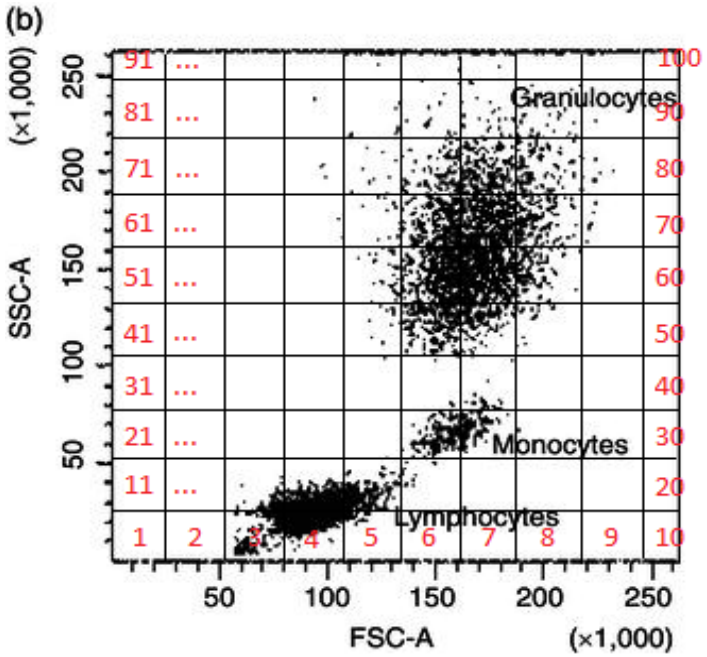


visible difference between the two patients. A machine learning algorithm would certainly be able to differentiate the two populations of cells.

2.2 Characterizing cell populations

Tracking the 30,000 cells individually in the scatterplot is very computationally onerous. One way to simplify the data is by histogramming the data by dividing the plot into boxes and counting the number of cells in each box (figure 6).

FIGURE 4. A labeled scatter plot divided into 100 boxes (numbered in red). This is a healthy patient with a fairly typical flow cytometry. A count of cells in healthy patients would show many cells of the lymphocyte type in boxes 4 and 14 just like this one.



A computer additionally has the ability to analyze in 3 dimensions and beyond. For a 3-dimensional histogram, there would be one thousand boxes, with each axis divided into 10. We can even divide 4-dimensional space into hypercubes, with ten thousand hypercubes! This can even be continued further, but is not optimal for analytical purposes, as the "curse of dimensionality" will begin to affect the analytical power of the machine learning model significantly (addressed in the machine learning chapter).

3.0 Machine Learning

Once the data has been scatter plotted and then converted into a histogram, the data can be processed and classified by a machine learning algorithm for each tube of cells. There will then be 7 different healthy or ill classifications for each patient, corresponding to each of the tubes of cells.

3.1 Choosing An Algorithm

There are a huge variety of machine algorithms that apply to many different situations. Choosing an incorrect machine algorithm is akin to using a flathead screwdriver to hammer the nails in a house. It might technically get the job done but the house will not be a safe structure.

There are many ways to choose an algorithm.

3.1.1 Function

There are *regression* algorithms that describe the data in an equation so that you can feed it data to calculate a predicted number. For an example, a weather algorithm that predicts the temperature based on predictors like humidity, pressure, and the previous day's high and low would be a regression algorithm.

For our data, however, we need a *classifier* algorithm; that is, an algorithm that sorts data into pre-specified labels - like healthy or sick!

3.1.2 Speed

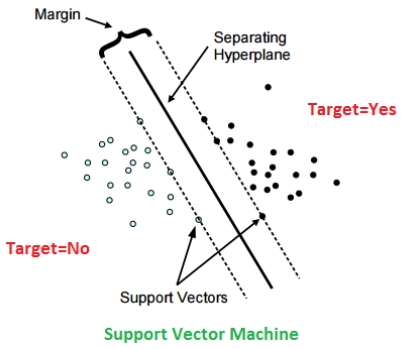
Algorithms can take a long time to make predictions. Other algorithms can predict quickly, but training takes a long time. For a leukemia data set, it wouldn't matter as much if it took days or even weeks to train as long as it could provide predictions quickly. After all, therapy is waits on timely diagnosis.

3.1.3 Robustness

Algorithms can excel in certain "spaces" where there are a limited number of predictors, but flounder in others with dozens, hundreds, or thousands of predictors, and vice versa. The way we have processed our data, the algorithm will have to process hundreds or thousands or predictors - the number of cells in each histogram "box".

3.2 Support Vector Machine

Given these conditions, an algorithm called Support Vector Machine (SVM) seems like a reasonable choice. SVM is an algorithm that is designed to fined a mathematical "boundary" separating sets of data as wide as possible.



For a set of data in two dimensions, this would be a line separating one label (healthy or sick) from another (figure).

Generally speaking, the boundary is found by choosing the small minority of data points that could be most meaningful in drawing up the boundary line. These few data points are the "support vectors" that make up the SVM (see appendix for more on SVM).

SVM offers many advantages that meet the criteria described previously. It is a classification algorithm that is very fast when classifying new samples, since it only requires knowledge of support vectors and not the entire dataset.

Furthermore it is a useful algorithm in high dimensional space due to the "kernel trick", which maps data in a way that helps to more clearly resolve data (figure). This is particularly useful for data with biological data.

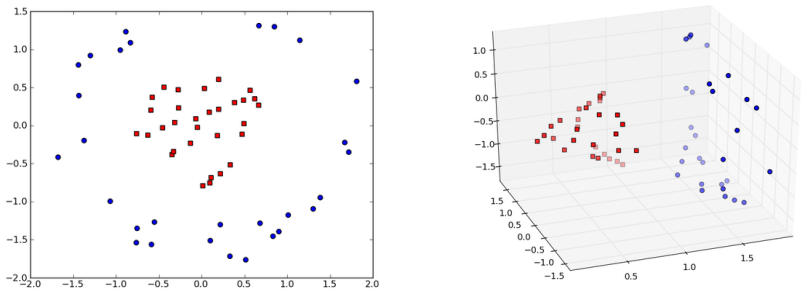


FIGURE 5. The Kernel Trick in SVM allows data to be separated in higher dimensions.

3.3 Overfitting

Overfitting is fitting an overly complex model to your data such that it produces excellent training scores, but the model is not useful because it does not generalize to predict new data well. In figure 9, the left figure is a regression line that is mathematically "correct", but clearly does not fit the data well. The overfitted figure on the right may technically fit the available training data exceedingly well, but is overly complex and likely does not predict new data well. In most cases, the middle figure is the balance to strive for.

For example, learning English by solely studying Shakespeare letter for letter, saying hello by saying "God ye good den, how now, Wench!" is a straight forward way of saying hello in a very specific time and place- Elizabethan era England. This is overfitting.

Underfitting, on the other hand, is learning English by only training on the most common words of the same Shakespeare text: "a, the, I, of, to, and." Correct for many more times and places, but also hardly enough to speak English fluently.

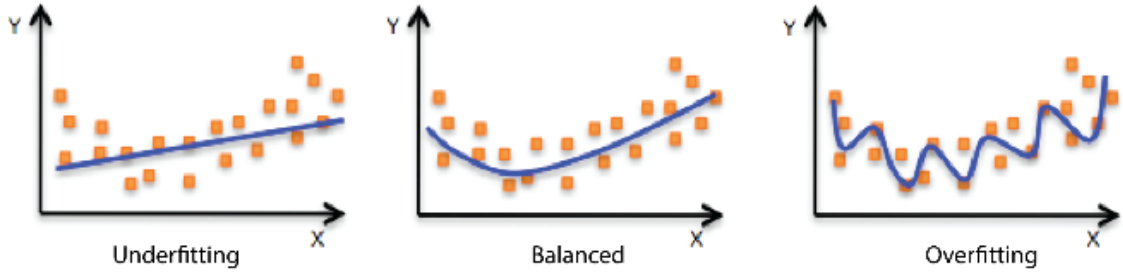


FIGURE 6. Examples of different algorithm fit types.

3.4 Regularization

In machine learning, the considerable processing power of the computer is harnessed, but can easily produce overfit. In order to avoid this, we can introduce "regularization", in which model complexity is mathematically "punished" by another term in the original SVM equation. This term is directly proportional to the misclassification errors produced, and it is also multiplied by a variable C that determines how severe the punishment is.

For a large C value, the punishment will be large, which means that the model will fit more exactly over the data, like the overfit model in figure 9. This will show high accuracy in training, but when the model is applied to other circumstances, it will fall apart.

Figure 10 shows an example of where introducing tolerance via adjusting C and allowing a "softer" margin may lead to a more robust model, even if it leads to some training misclassifications.

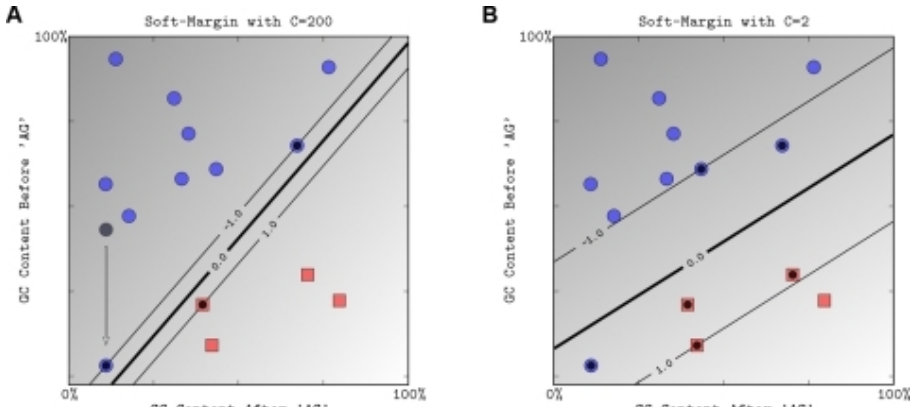


FIGURE 7. A strict boundary (left) may lead to overfitting, which is highly accurate in testing but performs poorly on new data. A soft margin (right) could apply to new data better.

3.5 Cross Validation

Cross validation is another way of reducing overfitting. As overfitting arises from over-reliance on a small amount of training data, we can chop our data up in several ways to present more. This helps lend more power to our model by making it more generalizable.

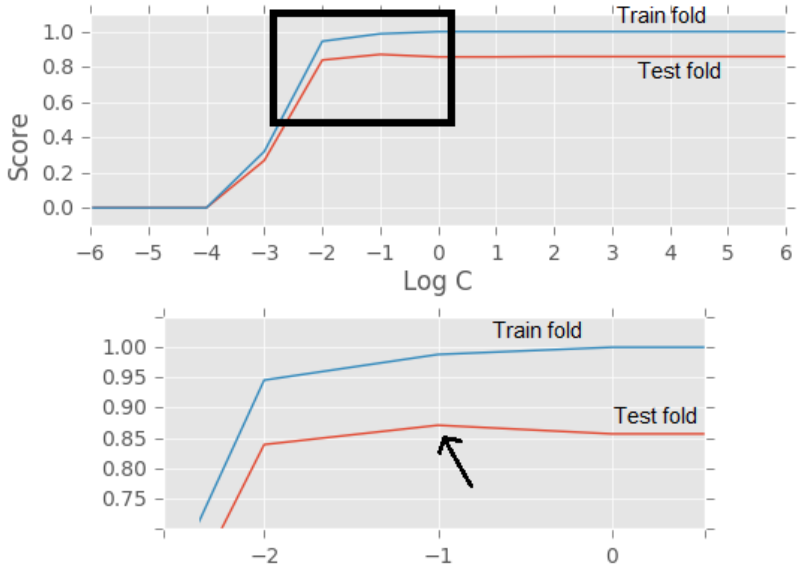
While one fifth of the data is held out for the test section (aka "test fold"), the remaining four fifths of data is trained to find an optimal C value. The chosen regularization C term is then used to make predictions on the test fold.

In figure 11 a sample grid search can be seen. At this stage we have generated a large dataset for a single tube of 359 patients by all possible marker combinations and are selecting a C to use for the 72 patients held out as a test fold. The grid search essentially repeats training and thentesting over a range of C between 10^-6 to 10^6.

10-fold cross validation was used during this process. One fifth of the total set is held out as previously mentioned, then the remaining o ? and ? for the purpose of grid searching C. For each C in the search, 10 models are trained and the results averaged in an effort to reduce variance. The reason for the nested split is to conserve the all-important separation of training influence from the final test fold until the predictions are ready to actually be made. Using test data as part of the training fold is a direct way to cause overfit.

It is important to select a C value based on test score rather than validation score. In figure 12 you can see both test and training score plotted for each value of C. While it is tempting to choose values of C for which the training score was a perfect 1.0, that would be the definition of overfit. Instead, as you can see in the exploded section, there is a peak for the test score while the training score is still rising, and this is the optimal C that should be chosen. In this example C would be 10^-1, which would then be used for original test fold prediction. It should be noted that this particular shape of score plot is conserved through every SVM grid search.

FIGURE 8. Training and test fold accuracy



3.6 Measuring Success

Once predictions are made for every tube, there will be 7 predictions per patient as each patient corresponds to 7 samples of cells. The eventual purpose of this algorithm is likely to screen patients for suspicious cases. The cost of missing a patient and having them go untreated is much more severe than having a pathologist check flow cytometry results unnecessarily, especially since they are already checking every patient every time anyways. Therefore, it is reasonable to simply assume that a patient is positive from any one positive result from any of their seven tubes, as this is maximally cautious.

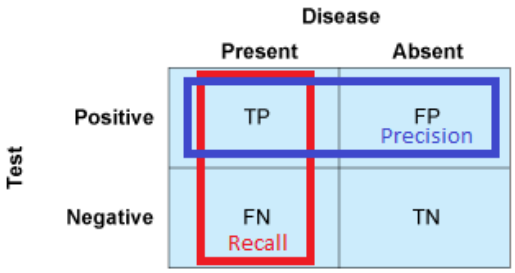
3.6.1 Misleading Statistics

In order to measure how good the overall algorithm is, a definition of "accuracy" needs to be made. In the literal sense, accuracy means the number of correct answers out of the total, which is straight forward, but not the most meaningful in this scenario. The reason why is that this definition of accuracy is misleading in datasets where the number of actual positives is very low, as in the following scenario:

1% of a population has a disease, similar to the common cold. If someone invented a "fake" test for this disease which said simply that literally everybody was healthy, they would still be 99% accurate! Of course, this is completely meaningless.

In the leukemia dataset this is a very relevant concern because there are only 43 positive samples to begin with, and in the greater population that this algorithm might be used in, the overall percentage of true leukemic patients is even lower.

3.6.2 F-Score



A common alternative is the F-score, which is defined as twice the harmonic mean of Precision (True Positives over Total Test Positives) and Recall (True Positives over Total Actual Positives). Note these formulas avoid using the true negatives in their calculations (figure 13). So for cases like the common cold (and leukemia!), the 99% that are truly negative do not skew the score, and give a misleading measure of test power.

3.7 Summary

To summarize, from 359 patients, combinations of markers are plotted by scatterplot. The scatterplot is converted into a histogram by dividing the plot into equal boxes. The numbers of cells falling into each box is recorded in a string of numbers. The SVM trains itself on these numbers, finding a model that is precise but avoids overfitting as much as possible. The end result is some boundary that separates the positive cases from the negative cases. This boundary space can then be used to predict new cases. F-score is the method used to measure the predictions at the end.

4.0 Results

4.1 Exploratory Analysis

Running the model with a very basic scenario of the first two markers in the first tube produces interesting results: an F-score of 0.76 and a recall of 0.67. This means 67% of actual sick patients were detected. This seems like a poor score but this is to be expected. Diagnosing based on only two markers out of 35 is like tying shoelaces with one hand.

If the test is expanded to include all markers, and even all combinations between them, the model would intuitively become more accurate to become more accurate. Indeed, the F-score increases to 0.85 and recall increases to 0.95. There are still 12 false positives and 2 false negatives though, and while it seems like the test is already more powerful, for these 14 patients (especially the false negatives) it is simply not good enough. Fortunately there are still ways to make the model more accurate.

4.2 Feeding the SVM More Dimensions

As mentioned earlier, the pathologists that review flow cytometry results view them as a long series of two dimensional scatterplots. Theoretically, one could construct a three dimensional representation of the data, but this would be hard to manipulate on a computer and even harder to comprehend by a human. And of course beyond three dimensions, the data becomes simply incomprehensible.

However, to a computer, the extra dimension is simply more numbers to crunch. Therefore, it can look at data in more than two dimensions. It's even possible to analyze all seven markers at once.

If the SVM is trained and tested on these plots with extra dimensions, the F-scores seem to continue to improve to a point (figure 14).

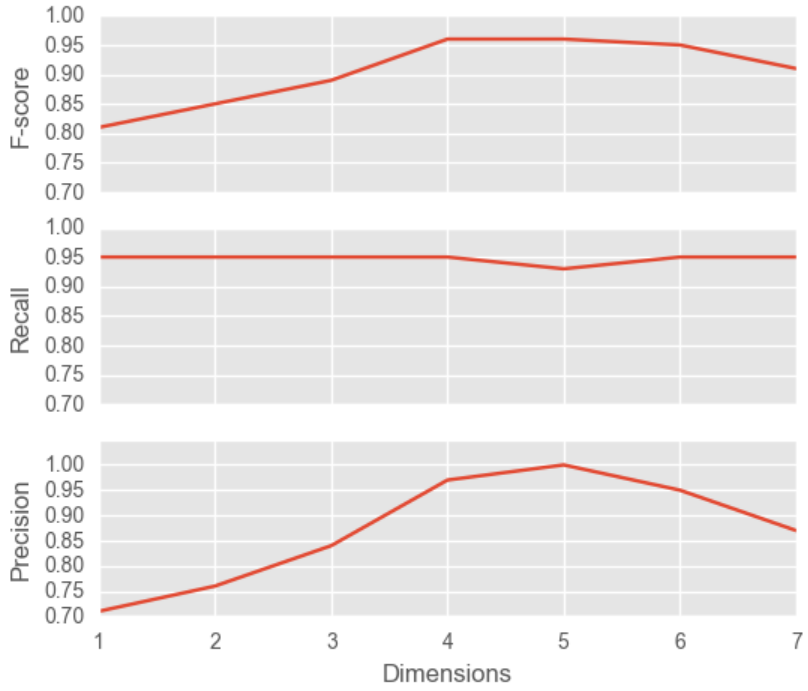
TABLE 1. Measurement of Accuracy and Regularization Terms for Higher Dimensional Space

Dimensions	F-score	Recall	Precision	TP	FP	FN	Features	Median C	S.d. of C	Log s.d. Of C
1	0.81	0.95	0.71	41	17	2	72	1	16.82	1.23
2	0.85	0.95	0.76	41	13	2	1295	10	21.18	1.33
3	0.89	0.95	0.84	41	8	2	9188	10	0	0
4	0.96	0.95	0.97	41	1	2	25519	10	0	0
5	0.96	0.93	1	40	0	3	32384	10	0	0
6	0.95	0.95	0.95	41	2	2	19158	10	0	0
7	0.91	0.95	0.87	41	6	2	4294	10	25.23	1.41

It seems that positive cases are detected at 95% (recall), except for 93% at 5 marker. Otherwise, the model becomes more discerning of negative patients (precision) until the 5 marker combination level. Afterwards, precision begins to drop again. F-score, as a combination of recall and precision, tops out with precision at 4 and 5 concurrent markers.

In addition, our regularization term has stayed relatively consistent despite the complexity of the data ballooning into seven dimensional space.

FIGURE 9. Measurement of Accuracy and Regularization Terms for Higher Dimensional Space



5.0 Discussion

Overall, the original question of machine learning prediction for flow cytometry seems to have been answered. The SVM algorithm performs at a consistent rate of detecting 95% cases of leukemia.

5.1 Analyzing More Than Two Markers

The algorithm's precision is higher as it reaches 5 marker analysis, but then seems to fall off again. An explanation for the dropoff may be that the "curse of dimensionality" comes into play (figure 16). That is, with every dimension added to your histogram, the "space" within which your data exists becomes exponentially larger. Perhaps the information gained from more markers "wins out" to begin with but then succumbs to dimensionality, giving rise to the peak at 4 and 5 concurrent markers.

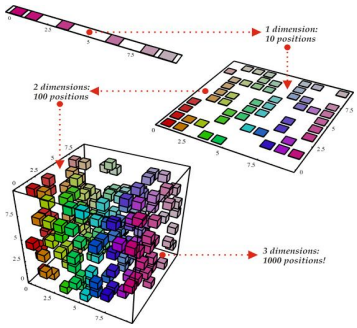
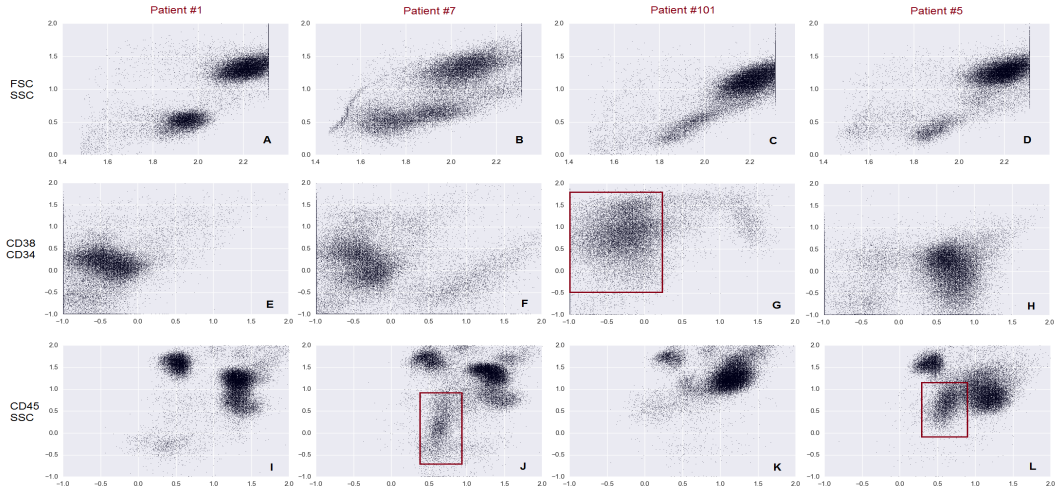


FIGURE 10. The Curse of Dimensionality

In any case the better precision in higher dimensional space is very interesting and suggests that there is value after all in analyzing more than 2 markers at once.

5.2 Patients #7 and #101

FIGURE 11. Outlier Results: Patient #1 (normal), Patients #7, #101, and #5 (leukemic)



Every model missed the same two positive cases: patients #7 and #101. Since the two cases are consistent against all 7 models, they may both have something in common that our model is not able to detect.

It is also possible that these two cases represent difficult to detect "edge" cases. Preliminary investigation of these two cases suggests that there may be some separability as suggested by the plots in figure 18. The rows are matching plots that are typically one of the first inspected by pathologists during routine evaluation. The red boxes indicate some of the suspicious populations that would certainly increase a human's degree of suspicion. This may be a consequence of the SVM method. A different classifier algorithm, such as "K-Means" might yield more sensitivity for these patients.

Another avenue for improvement is to change the relative weighting of marker combinations in diagnosis. Some markers have more predictive value than others; certainly the ones plotted in the figure are generally significant. It is possible that the signal from these combinations is being diluted by opposing or neutral signals from other markers. This is an area that a tree boosting algorithm such as XG Boost may certainly help.

6.0 Summary

Machine learning algorithms are a promising way to accurately read flow cytometry. Exploring a single algorithm and a single way to manipulate the data shows an F-score of 96% in an environment computing data of 4 and 5 dimensions at once. This implies that our current human-read flow cytometry may be improved upon. If optimized for further accuracy, flow cytometry may be a way to save significant time, energy, and money, while providing accurate diagnoses for any patient needing flow cytometry.

Appendix A: Support Vector Machine

For a set of two binarily labeled sets of data, there exists a "hyperplane" in between the sets that separates the two. For simplicity, we can consider two dimensional data first, for which the hyperplane is a line. Surrounding the line on either side are two halves of the boundary zone that contains no data. There are many possible lines to choose (in most cases an infinite number), but the goal of the support vector machine is to find the line for which this boundary is as wide as possible.

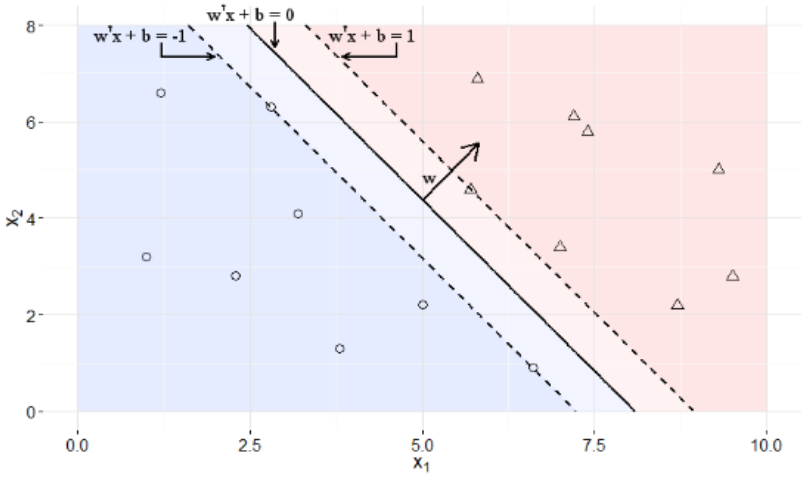


FIGURE 12. Support Vector Plotting

The general equation for the line or hyperplane is: wTx + b = 0 where w is called the weight vector perpendicular to the margin and b is some constant to move the line away from the origin. We may label positive examples +1 and negative examples -1 such that positive samples are all wTx + b ≥ 1 and wTx + b ≤ -1.

Consider two points as close as possible to each other along these boundary lines wTx+ + b = 1 and wTx- + b = -1. The magnitude of the vector between these two points is equivalent to the width of the boundary, and furthermore, the line is parallel to the weight vector w. The difference between these equations gives wT(x+-x-) = 2. Since w and the vector between the points are parallel, this is also equivalent to ||wT|| * ||(x+-x-)|| = 2. Divide by w to get the equation for the width of the margin: ||(x+-x-)|| = 2 / ||wT||. We then see that we maximize the boundary by minimizing w.

You can convert this into a quadratic programming equation: subject to the constraints of: 0 ≤ αi and y ∑ αi yi = 0. The terms α can best be thought of as weights for each sample, and it turns out that once solved, most α are equal to zero. The points with non-zero α are those that lie close to the boundary that contribute the most to its definition. These points are what are called the support vectors.

Appendix B: Flow Cytometry

Flow cytometry is a cell analysis technique that was first used in the 1950s to measure the volume of cells in a rapidly flowing fluid stream as they passed in front of a viewing aperture. Since that time, innovations from many engineers and researchers have culminated in the modern flow cytometer, which is able to make measurements of cells in solution as they pass by the instrument's laser at rates of 10,000 cells per second (or more). Today's instruments offer an increased number of detectable fluorescent parameters (from 1 or 2 up to ~30 or more), all measured at the same time on the same cell. Because of its speed and ability to scrutinize at the single-cell level, flow cytometry offers the cell biologist the statistical power to rapidly analyze and characterize millions of cells, albeit at the expense of the morphological characteristics and subcellular localization that microscopy can provide.

References and further reading:

- Wood BL. Myeloid malignancies: myelodysplastic syndromes, myeloproliferative disorders, and acute myeloid leukemia. Clin Lab Med. Sep 2007; 27(3):551-575, vii.
- Cherian, Sindhu, M.D. Flow Cytometry for Acute Myeloid Leukemia. https://www.cytometry.org/public/educational_presentations/Cherian.pdf.
- Aghaeepour, N., Finak, G. Critical assessment of automated flow cytometry data analysis techniques. FlowCAP Consortium. January 10, 2013.

Figures:

- *Fluorescent flow antibody* http://www.signalsblog.ca/inside-a-cancer-stem-cell-researchers-toolbox-csc-markers-flow-cytometry/
- *2d SVM* http://dni-institute.in/blogs/building-predictive-model-using-svm-and-r/
- *SVM kernel trick* https://www.quora.com/Support-Vector-Machines-How-does-going-to-higher-dimension-help-data-get-linearly-separable-which-was-non-linearly-separable-in-actual-dimension
- *Overfitting* http://docs.aws.amazon.com/machine-learning/latest/dg/images/mlconcepts_image5.png
- *Regularization* https://openi.nlm.nih.gov/detailedresult.php?img=PMC2547983_pcbi.1000173.g003&req=4
- *Curse of dimensionality* https://haifengl.files.wordpress.com/2016/01/cursedimensionality.jpg