Classifying HIV Vaccination Status with Regularized Logistic Regression

Ariful Azad, Arif Khan, Bartek Rajwa, Alex Pothen

Purdue University

FlowCAP-III, NIH, November 29-30, 2012

This research was supported by grant 1R21EB015707 from the National Institute of Biomedical Imaging and Bioengineering and NSF grant CCF-1218916
Problem: Predict the vaccination status (pre- and post- vaccination) of samples from HIV patients. Half of the samples with known vaccination status are given as training set.

Method: We used the fraction of cells in different combination of Boolean gates, and Median Fluorescence Intensity (MFI) as features or explanatory variables. We then train a logistic regression model with Lasso regularization (RLR) with the training set and obtained a sparse model with four predictive features.

Results: The optimized RLR model performs good on training set with four (out of 37) misclassification. On the test set, the model classify 29 out of 37 samples with high confidence.
**Problem Description**

**Dataset**

- Application of a HIV vaccine on 74 subjects at two time points (before and after vaccination), 37 in training set and 37 subjects in test set.
- At each time point we have a POL-3 stimulated sample and two negative controls.
- Each sample has six markers. $CD3$, $CD4$, $CD8$ are for identifying T cell subpopulations. The remaining markers are cytokines $TNFa$, $IFNg$, and $IL2$.
Automated $CD4^+$ and $CD8^+$ T cell gatings

- We used norm1filter and norm2filter from flowCore to perform the automated gatings.
Automated Cytokine gating

- We applied **patient specific normalization** to all six samples from a particular subject and used norm2filter to identify $\text{TNFa}^+$, $\text{IFNg}^+$, and $\text{IL2}^+$ cells.
- Cytokine positive cells are extremely rare in $\text{CD8}^+$ cells, and we mainly used them when $\text{CD4}^+$ is unable to classify a pair of samples.
Feature Selection

- For each sample, we computed a Boolean (positive/negative) gating for each of the three cytokines.
- The Boolean gates can then be combined in $3^3 = 27$ ways by considering positive, neutral and negative levels of expression. We, however, kept only those combinations with at least one positive cytokine.
- We consider the fraction of cells within a Boolean gate combination as a feature.
- In addition we included median fluorescence intensity (MFI) of three cytokines as features in our model.
- Hence, we have about 21 features.
Model selection

- The dependent variable is the vaccination status of a sample (vaccinated or not-vaccinated)
- Therefore, this is a binary classification problem. We used Logistic Regression for this classification.
Logistic Regression

- Widely used for binary classification, e.g., Vaccinated and not-Vaccinated
- Explanatory variable $x_i$, such as fraction of cells in a combination of Boolean gate. e.g., $TNFa^+ IFNg^- IL2^+$
- Dependent variable $y_i$, Vaccinated, $y_i=1$ and not-Vaccinated, $y_i=0$
- Probability of $i^{th}$ sample being Vaccinated = $p_i$
- log odds for the event $y_i=1$, $logit(p_i) = log\left(\frac{p_i}{1-p_i}\right)$
Logistic Regression

- \( \text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_d x_{id} = \beta_0 + x^T \beta \)
- \( p_i = \frac{1}{1 + e^{-(\beta_0 + x^T \beta)}} \), logistic function
The dependent variable follows a binomial distribution, $y_i \sim \text{bin}(1, p_i)$

maximize the log likelihood:

$$\max_{(\beta_0, \beta) \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \{y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\}$$

which is equivalent to

$$\max_{(\beta_0, \beta) \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \{y_i (\beta_0 + x_i^T \beta) - \log(1 + (\beta_0 + x_i^T \beta))\}$$
Lasso Regularization

- Pick the **predictive features** by penalize models with too many parameters [Friedman et. al. 2009]
- maximize the log likelihood:

\[
\max_{(\beta_0, \beta) \in \mathbb{R}^{d+1}} \left[ \sum_{i=1}^{n} \{ y_i (\beta_0 + x_i^T \beta) - \log(1 + (\beta_0 + x_i^T \beta)) \} - \lambda \| \beta \|_1 \right]
\]

- Select a sparse solution with few non-zero values for $\beta_i$
- We used R package **glmnet** by Jerome Friedman, Trevor Hastie, and Rob Tibshirani.
Model Parameter Selection

- The model parameters to be selected are $\beta_0, \beta_1...\beta_d$ and $\lambda$
- For fixed $\lambda$, $\beta_0, \beta_1...\beta_d$ are estimated by maximizing the log likelihood
- $\lambda$ is selected from n-fold cross validation (minimize $\sum o_i \log(\frac{o_i}{e_i})$)
Significance of the selected features

- A sparse solution with only four features being used

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient in the model</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFI $\text{TNFa}^+$</td>
<td>2.293</td>
</tr>
<tr>
<td>$\text{TNFa}^+ \text{IFNg}^+ \text{IL2}^+$</td>
<td>1.421</td>
</tr>
<tr>
<td>$\text{TNFa}^+ \text{IFNg}^- \text{IL2}^-$</td>
<td>0.397</td>
</tr>
<tr>
<td>$\text{TNFa}^- \text{IFNg}^- \text{IL2}^+$</td>
<td>-0.844</td>
</tr>
</tbody>
</table>

**Table**: Optimal Solution of the Regularized (Lasso) Logistic Regression
Model verification by incremental feature selection

- Build logistic regression model by **incrementally adding features**.
- Incrementally complex models from simpler models. Decrease the misclassification as we include features.

<table>
<thead>
<tr>
<th>Incremental Model features</th>
<th>p-value</th>
<th>AIC</th>
<th>Tr</th>
<th>Misclassification</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFI ( TNFa^+ )</td>
<td>2.46e-07</td>
<td>79.95</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>( TNFa^+ IFNg^+ IL2^+ )</td>
<td>2.20e-08</td>
<td>73.33</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>( TNFa^+ IFNg^- IL2^- )</td>
<td>3.15e-08</td>
<td>72.81</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>( TNFa^- IFNg^- IL2^+ )</td>
<td>4.69e-09</td>
<td>67.93</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
The RLR model predicts the probability of a sample being vaccinated. Low probability for non-vaccinated and high probability for vaccinated samples.

From a pair of samples (before and after vaccination) from a patient, the sample with high probability is predicted as vaccinated. Example:

Let \( p(s_1) \), and \( p(s_2) \) be the probabilities predicted by a trained RLR model for a pair of samples, \( s_1 \) and \( s_2 \) from a patient. If \( p(s_1) > p(s_2) \) then the model predicts \( s_1 \) as vaccinated and vice versa. \( |p(s_1) - p(s_2)| \) indicates the confidence on the prediction.
Prediction in the training set

- Four misclassification in the training set. Misclassified samples are marked with green circles.
Prediction in the test set

- Prediction in the test set. We have eight pairs of samples predicted with low confidence (green circles). Thus about 75% samples are classified with high confidence.
We used a logistic regression model with Lasso regularization (RLR) to classify samples to HIV vaccinated/not-vaccinated classes. The RLR model was able to automatically select the features predictive to the vaccination status.

Results: The optimized RLR model performs good on training set with four (out of 37) misclassification. On the test set, the model classifies 29 out of 37 samples with high confidence.
Thank You!
Preprocessing

- Each sample is stained against 6 markers aside from ViViD for live cell gating.
- The CD3, CD4, and CD8 markers are used to identify $CD4^+$ and $CD8^+$ subpopulations in T cells.
- The remaining markers are cytokines $TNFa$, $IFNg$, and $IL2$.
- Expression of the cytokines (quantity and quality) in T cells are known to be indicative of the HIV progression (Seder et. al. 2008, Betts et. al. 2006).
- Therefore, we aim to identify cells with at least one cytokine expressed, both in $CD4^+$ and $CD8^+$ T cells, and combine them to predict the disease (or vaccination status).
Model verification with single feature

- Consider a single feature at a time and build a logistic regression model
- Incrementally complex models from simpler models. Keeps almost same set of features.

<table>
<thead>
<tr>
<th>Model with feature</th>
<th>p-value</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFI $TNFa^+$</td>
<td>2.46e-07</td>
<td>79.95</td>
</tr>
<tr>
<td>$TNFa^+ IFNg^+ IL2^+$</td>
<td>7.47e-05</td>
<td>90.90</td>
</tr>
<tr>
<td>$TNFa^+ IFNg^- IL2^-$</td>
<td>1.68e-07</td>
<td>79.21</td>
</tr>
<tr>
<td>$TNFa^- IFNg^- IL2^+$</td>
<td>5.00e-01</td>
<td>106.13</td>
</tr>
</tbody>
</table>