Machine Learning in Untargeted Metabolomics: Final report for c150

Alex Tong

December 16, 2016

1 Background

I've been working on this project starting in February 2016 with Soha Hassoun. My contribution has been in the probabilistic model design and test. In the spring we started with a Bayesian Network designed in BayesNetToolkit. [?]

2 Introduction

3 Model Specification

Parameters:

λ_p: Probability that a pathway is active

μ_0: Probability that a feature is present given inactive pathway

μ_1: Probability that a feature is present given active pathway

Variables:

a_p: IRV indicating pathway p activity

$b_{p,f}$: IRV indicating feature f is associated with pathway p

$o_{p,f}$: IRV indicating whether feature f associated with pathway p is present in the sample due to pathway p

m_f: IRV indicating whether feature f is present in the sample

v_f: IRV (virtual evidence on feature f

Generative Model Prior:

P_p: $Bernoulli(\lambda_p)$ for $p = 1...P$

$o_{p,f}|P_p, \mu$: $Bernoulli(\mu_{P_p})$ for f in Features(p)

M_f $= (1 - \prod_p (1 - o_{p,f}))$ Equivalent to logical OR

$v_f \ = Bernoulli(\text{Measured P}(f))$

Observation:

$v_f \ = 1$

Posterior:

$$
\begin{aligned}
p(o|\lambda, \mu_0, \mu_1, b_{p,f}, a_p) &= \prod_p \prod_f (\mu_{a_p}^{o_{p,f}}(1-\mu_{a_p})^{(1-o_{p,f})})^{b_{p,f}} && (1)\\
p(m|o,b) &= \prod_f m_f = \prod_f (1 - \prod_p (1-o_{p,f})^{b_{p,f}}) && (2)\\
p(\lambda, \mu_0, \mu_1, a, o, m) &= p(a|\lambda)p(\lambda)p(\mu_0)p(\mu_1)p(o|\lambda, \mu_0, \mu_1, b_{p,f})p(m|o) && (3)\\
p(\lambda, \mu_0, \mu_1, a, o, m|v = \mathbf{1}) &= \frac{p(v|\lambda, \mu_0, \mu_1, a, o, y) * p(\lambda, \mu_0, \mu_1, a, o, m)}{p(v = \mathbf{1})} && (4)\\
&\propto p(v|m) * p(\lambda, \mu_0, \mu_1, a, o, m) && (5)
\end{aligned}
$$

Description: Equation 1 shows the likelihood of a given set of o variables. For example, if I wanted to calculate the probability of all $o_{p,f}$ variables being zero, I would need all given hyperparameters, λ, μ_0, μ_1, and the values of a_p. The likelihood as stated is a function of p variables, $a_{1...p}$. Note that with this likelihood function, it is simple to calculate the likelihood $P(o|\lambda, \mu_0, \mu_1, a)$, In fact, for a given $o_{p,f}$, we can calculate $p(o_{p,f}|a_p, b_{p,f}) = \mu_{a_p}^{o_{p,f}}(1-\mu_{a_p})^{(1-o_{p,f})}$ or

$$p(o_{p,f} = 1|a_p, b_{p,f}) = \mu_{a_p}$$

$$p(o_{p,f} = 0|a_p, b_{p,f}) = 1 - \mu_{a_p}$$

Equation 2 shows the likelihood of a set of metabolite observations given o. for example, the probability of getting $m_1 = 1, m_2 = 0, m_3 = 1$ given all of o, is a constant.

Equation 3 shows the likelihood over all hidden variables. This is derived from looking at our bayesian network, as each variable is independent.

Equation 4 shows the model likelihood given our observation of our virtual nodes. This is derived from bayes rule.

$$p(v_f|m_f) = P(metfrag) * P(\pi)$$

Reasonable Values:

- π should be nominally quite low, and may be lower for some metabolites than others possibly with the idea that larger molecules are harder to detect as there are more possible fragments.
- We will start with $\mu_0 = 0.001$ and $\mu_1 = 0.999$ as values very close to 0 and 1.
- We will start with $\lambda = 0.5$, but would like to move to a model where we can incorporate more reasonable priors separately on each pathway, something like a λ_p for $p = 1...P$.

4 Summary

5 Acknowledgments

Thanks to Nicholas Alden, Vlad Porokhin, Neda Hassanpour, Kyongbum Lee, Soha Hassoun who are also working on this project.

6 References