イジングマシンを用いた正確ロジスティック回帰分析

大規模知識発見分野　津田研究室
佐藤 史歩

1. 背景

ヒトゲノム全体をほぼカバーする 1000 万カ所以上の一塩基多型（SNP）のうち、50 万～100 万か所の遺伝子型を決定する方法はゲノムワイド関連解析（Genome Wide Association Study；GWAS）と呼ばれ、主に SNP の頻度と、病気や量的形質との関連を統計的に調べるために用いられている。GWAS で用いられる代表的な統計モデルはロジスティック回帰であるが、尤度比検定など漸近論による統計解析を行う際、サンプルが小さい場合には不正確になってしまうという問題がある。それを回避する方法として正確ロジスティック回帰分析（Exact logistic regression）が提案されている。正確ロジスティック回帰分析は Cox によって 80 年代に提案された手法で、パラメータの反復最適化を行わず、サンプリングによって P 値計算を行う[1]。結果、サンプル数の大きさによって P 値の大きさが異なるという現象を回避することができる。しかしこの方法は計算が複雑であることが障壁となり、長年普及しなかった。

本研究では D-wave 量子アニーラの基底状態サンプリング[2]を用いて Exact logistic regression を実現することを目的とする。本稿では、理論的な背景を論じた後、実データを用いた実験を示し、今後の展望について述べる。

2. 方法

本章では Exact logistic regression を説明し、その後基底状態サンプリングの適用箇所について言及する。

例えば特徴量がm個$(1 \le j \le m)$, サンプル数がn個$(1 \le i \le n)$であるデータに対し、各特徴量における P 値を計算することを考える。Exact logistic regression では十分統計量として以下を使用する。

$$t_0 = \sum_{i=1}^{n} y_i \quad t_j = \sum_{i=1}^{n} x_{ji} y_i$$

k番目の特徴量についてk番目以外の特徴量を交絡因子として P 値を計算したい場合、まずはt_k以外に対して以下のように条件付けをし、条件を満たす Y を全列挙する。$\hat{t}_n$は元々のデータセットにおける Y に対して計算した十分統計量を表している。

$$Y = \{y | t_0 = \hat{t}_0, t_j = \hat{t}_j, j \neq k\} \qquad (1)$$

式(1)を満たすすべての Y サンプルに対しt_kを計算し、$\hat{t}_k$よりも小さいものの割合を P 値とする。以上が正確ロジスティック回帰の手順である。

本研究では、条件を満たす Y を、基底状態サンプリングによって生成する。イジングマシンにおいては、QUBO(Quadratic unconstrained binary optimization)と呼ばれる二次式で表されるエネルギー関数（ハミルトニアン）を最小化する基底状態（最適解）を発見するが、基底状態が複数あるときは、その中からサンプリングを行うことができる[2]。

Exact logistic regression をイジングマシンを用いて行うには、式 1 を満たす解が基底状態に

なるようなハミルトニアンを設計する必要がある。ここでは以下のように設定した。

$$H = (t_0 - \hat{t}_0)^2 + \sum_{j \neq k} (t_j - \hat{t}_j)^2$$

上記式を QUBO 形式にし、イジングマシンに入力した。

3. 実験

Exact logistic regression に関する文献[1]に掲載されていた骨肉腫に関するデータ（46 サンプル）を用いて、全列挙による P 値計算と、提案手法との比較を行った。このデータには、3 種類の説明変数と、1 種類の応答変数が存在する。今回の実験では、リンパ球浸潤に関する説明変数の P 値を、他の二変数を交絡因子として推定する。

提案手法の実装は、通常の計算機（2.7GHz クアッドコア Intel Core i7）におけるシミュレーテッドアニーリング(SA)と、D-Wave 実機における量子アニーリング(QA)の二種類を用意した。十分統計量を満たす解の全列挙は、4ti2[3]を用いて行った。SA、QA とも 1 万サンプルの生成を行った。両手法とも、全てのサンプルに関して必ずしも基底状態に到達できるわけではない。基底状態に到達したサンプルの数を有効サンプル数と呼ぶ。

スケーラビリティを評価するため、データ点を 20 点、25 点、30 点、35 点、40 点にダウンサンプリングしたものを使用した。各データ点数に関して、5 種類のデータセットを用意し、各手法に関して、計算時間、有効サンプル数、推定された P 値の誤差を測定した。30 点の結果を表 1 に示す。SA,QA とも、全列挙に比べて、大幅に高速であり、また、P 値の誤差は 0.06 程度に抑えられていることがわかる。

図 1 に SA と QA の計算時間の変化を示す。SA おいては、ビット数に比例して計算時間が増加しているが、QA においては変化がない。これは D-wave においては、アニーリングタイムが一定に設定されているためであり、そのため有効サンプルが得られる成功率も SA に比べて小さくなっている。問題の難しさに応じてアニーリングタイムを延長することで、成功率は改善できる可能性がある。SA と QA で計算量に大差はなかったが、量子計算機が発展すれば優劣が逆転する可能性は高い。

表 1：点数 30 のデータに対する実験結果。SA、QA とも 1 万サンプルを生成した。

	計算時間（秒）	有効サンプル数	P 値誤差
全列挙	54800	30200	0
SA	8.36	9210	0.006
QA	2.4	82	0.06

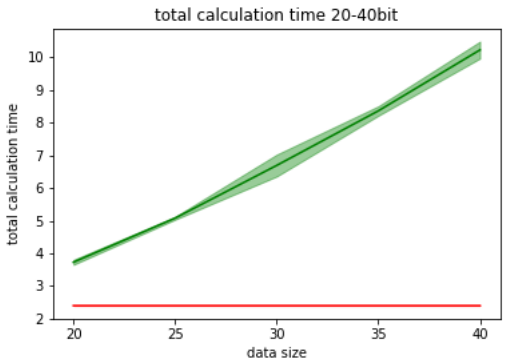

図 1：1 万サンプル生成時の SA（緑）と QA（赤）の計算時間

4. 参考文献

[1] Mehta, C. R., and N. R. Patel. Statistics in medicine 14 (1995): 2143-2160.

[2] Benedetti, M, et al. Physical Review A 94.2 (2016): 022308.

[3] https://4ti2.github.io/