イジングマシンを用いた正確ロジスティック回帰

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、現存の統計的仮説検定を行う際いくつかの問題が生じる。例えば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ったり、適合度検定にて作成した分割表が疎となる場合、最尤推定を実施することができない場合がある。それを回避する方法として Exact logistic regression（正確ロジスティック回帰）が提案されている。Exact logistic regression は Cox によって 70 年代に提案された手法で [1]、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 [2]。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、長年普及されてこなかった。本研究では D-wave 量子アニーラの基底状態サンプリング [3] を用いて Exact logistic regression を実現することを目的とした。骨肉腫データに適用した結果を示した後、現状の課題や今後の展望について述べる。

2　方法

2.1　量子アニーリング

量子アニーリングとは量子効果を用いて組合せ最適化問題を解く手法のことである。組合せ最適化問題は,、エネルギーが

$$H(\boldsymbol{\sigma}) = \sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{N} h_i\sigma_i \tag{1}$$

$$\sigma_i \in \{-1, +1\}$$

で与えられるイジング模型の基底状態 (最低エネルギー状態) を求める問題とみなすことができる。ここで $\boldsymbol{\sigma}$ はイジングスピンと呼ばれ, J_{ij} はスピン間の相互作用, h_i は局所磁場を調整するパラメータである. 量子アニーリングでは、はじめに各スピンの状態を不確定,、つまり ±1 を重ね合わせた状態に設定する。そして,、量子効果を徐々に小さくして各スピンの状態が ±1 のどちらかに確定した状態が自律的に選ばれ, その状態が基底状態となる。カナダの D-Wave Systems Inc. は量子アニーリングの動作原理が採用されたハードウェアを世界で初めて開発し、現在ではクラウド経由で利用できるようになっている [4]。

2.2　Exact logistic regression

本章ではまず 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。Exact logistic regression は共変量と応答変数ベクトルとの内積値を用いた条件付き分布を用いることで、検定したい説明変数に対する正確な P 値を推定することが可能である [1]。例えば共変量が m 個 $(1 \leq j \leq m)$、独立した標本データが n 個 $(1 \leq i \leq n)$ あり、各共変量における 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。

$$t_0 = \sum_{i=1}^{n} y_i\, t_j = \sum_{i=1}^{n} x_{ij} y_i \tag{2}$$

k 番目の共変量について、k 番目以外の共変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。

$$Y = \{y | t_0 = \hat{t}_0, t_j = \hat{t}_j, j \neq k\} \tag{3}$$

式 (3) を満たすすべての Y サンプルに対し t_k を計算し、その割合を P 値とする。

$$P(t_k = \hat{t}_k | t_0 = \hat{t}_0, t_j = \hat{t}_j) = \frac{c(t_k = \hat{t}_k, t_j = \hat{t}_j)}{\sum_u c(t_k = u, t_j = \hat{t}_j)} \quad j \neq k \tag{4}$$

$c(\alpha = \hat{\alpha}, \beta = \hat{\beta})$ は $\alpha = \hat{\alpha}, \beta = \hat{\beta}$ を満たす Y の個数を表している。以上が Exact logistic regression の手順である。

本研究では, 条件を満たす Y の全列挙過程を、基底状態サンプリングに変更した. 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。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。Exact logistic regression をイジングマシンを用いて行うには、式 (3) 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。

$$H = 2\sum_{p<q} y_p y_q + (1-2\hat{t}_0)\sum y_p + \hat{t}_0^2 + \sum_{j\neq k}(2\sum_{p<q} x_{pj}x_{qj}y_p y_q + (1-2\hat{t}_j)\sum y_p x_{pj} + \hat{t}_j^2) \quad (5)$$

3　結果

実験用のデータには 46 サンプルをもつ単変量解析結果データである骨肉腫データを使用した [5]. 骨肉腫データのうち特徴量は 3 つ使用し,20 サンプル, 25 サンプル,30 サンプル, 35 サンプル, 40 サンプルを 46 サンプルからランダムに選択したデータをそれぞれ 5 つ作成した. 説明変数は多変量解析にて有意性が認められた特徴量であり，特徴量に 1 つ含まれている. 応答変数は 3 年再発せず生存したか否かの二値変数である. 実験用のプログラムは Python3.7.0 を用いて実装した.。量子アニーラとして D-Wave Systems Inc. の D-Wave2000Q をクラウドサービスである Leap を介して利用した。まず計算時間を計測した結果を図 1に示す。

40bit に至るまでサンプル獲得と時間計測が可能であった SA と QA を比較すると、QA のほうが計算時間が短いことがわかる。さらに表 1からより詳細に結果を観察すると、QA はデータサイズが大きくなるほど SA よりも短時間で計算ができることがわかる。

表 1: SA と QA におけるサンプリング時間

	データサイズ [sec]				
手法	20bit	25bit	30bit	35bit	40bit
SA	3.73	5.08	6.70	8.36	1.02×10^1
QA	2.40	2.40	2.40	2.40	2.40
SA / QA	1.55	2.12	2.53	3.48	4.26

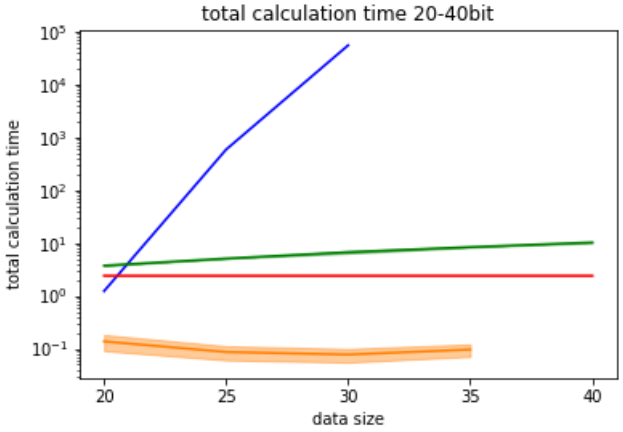

図 1: 計算時間の比較

4　考察と今後の課題

表??より様々なサンプリング手法にて実験した結果,、QA を用いて行なった結果が,、もっとも最短で解を検出できたことが分かった。古典アニーリングとして知られている SA と比較しても、80 倍-1189 倍速く計算できている。今後の課題として、 選択的推論への展開、他のイジングマシンへの展開などが挙げられる。

参考文献

[1] Cox, D. R. ”Analysis of Binary Data” *Methuen*(1970).

[2] Mehta, Cyrus R., and Nitin R. Patel. ”Exact logistic regression: theory and examples.” *Statistics in medicine* 14.19 (1995): 2143-2160.

[3] Benedetti, Marcello, et al. ”Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applications in deep learning.” *Physical Review A* 94.2 (2016): 022308.

[4] Johnson, Mark W., et al. ”Quantum annealing with manufactured spins.” *Nature* 473.7346 (2011): 194.

[5] Jaffe, Norman, et al. ”Osteosarcoma: Intra‐arterial treatment of the primary tumor with cis‐diammine‐dichloroplatinum II (CDP): Angiographic, pathologic, and pharmacologic studies.” *Cancer* 51.3 (1983): 402-407.