Linear Contextual Bandits for Online Warfarin Dosage Prediction

Ryan Silva
Department of Computer Science
Stanford University
rdsilva@stanford.edu

Jon Braatz
Department of Computer Science
Stanford University
jfbraatz@stanford.edu

Abstract

The problem of prescribing the correct dosage of the anticoagulant Warfarin to patients given their medical information can be formulated as a bandit problem when learning is done in an online setting as opposed to a batch setting. We explore the application of bandit algorithms to this problem using a data set comprised of medical information and the correct Warfarin dosage for 5000 patients as determined by trained physicians, and we compare their performance to baseline algorithms that don't alter their behavior as new patient outcome information becomes available. In particular, we investigate the LinUCB algorithm, the LinUCB hybrid algorithm, and the LASSO bandit algorithm and compare them to policies of recommending a fixed dosage and a dosage based on linear regression. We measure our implementations using (1) regret, which captures how well a given policy performs against an optimal policy, (2) the number of incorrect dosage decisions as a percentage of patients seen, and (3) performance on a test set. All bandit algorithms outperform the fixed-dose baseline and exhibit near sub linear regret behavior, and the LASSO Bandit is able to outperform the linear regression model with the aid of more patient features.

1 Introduction

Warfarin is a widely used blood anticoagulant that is commonly prescribed to prevent blood clotting. Warfarin can be challenging to prescribe in practice due to high variation of the correct dosage among patients. In addition, taking a substantially wrong dosage can have severe consequences. Typically, doctors prescribe a medium dose, monitor the patient's' response, then adjust the dosage [3]. Predicting the correct Warfarin dosage the first time would cut out this burn-in period and is of great interest since this would save time and cost for both the patient and doctor.

This paper addresses different approaches to solving the online decision making problem of prescribing doses of Warfarin to patients based on relevant health indicators. At each time step an agent observes a context about a patient, and must choose a level of Warfarin indicating high, medium, or low dosage, based on previous interactions with patients. We consider three contextual multi-armed bandit algorithms for this task: the LinUCB algorithm and LinUCB Hyrbid algorithms [4] as well as the LASSO Bandit algorithm [1].

Multi-armed bandits are well suited for this task since the setting naturally aligns itself to the bandit problem. Specifically, the agent has a choice between k = 3 actions, with the goal of maximizing expected total reward for a finite number of patients. The reward is stochastic and we choose a linear reward model with Gaussian noise. The payoff for an action is then 0 if correct, and -1 for an incorrect decision. This model assumes that the reward only depends on the features and that the reward is linear. We evaluate our implementations using cumulative expected regret and the average fraction of incorrect decisions, both on training data as well as a held out test set.

The online setting of this problem makes it challenging for an agent, which must trade off between exploitation to minimize regret and exploration in order to get better estimates of the parameters for each arm. The LinUCB algorithms use a form of Upper Confidence Bounds (UCB) to make this trade off in a principled manner, while the LASSO Bandit uses forced-sampling.

In Section 2, we review related work. In Section 3.3 we describe our approach for the LinUCB, LinUCB Hybrid and LASSO Bandit algorithms. In section 4 we provide our experimental results. In Section 5 we provide a discussion.

2 Background and Related Work

We use a publicly available patient dataset that was collected by staff at the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) for 5528 patients who were treated with Warfarin from 21 research groups spanning 9 countries and 4 continents. The dataset includes contextual information for each patient in the form of their gender, race, age, height, weight, and aspects of medical history such as the types of medications that they had been prescribed in the past. In addition, we are given the ground-truth optimal warfarin doses for each patient, which were found through a physician-guided process over the course of a few weeks.

We implement two baselines: the first is a fixed dose baseline that predicts a medium dosage (5mg/day) to all patients. This is the policy that physicians currently follow in practice[1]. The performance of this model over all patients in the dataset is 61.46%. Second, we evaluate the linear regression model detailed in [3], section s1f. Our experiments show the performance of this model to be 65.08% over patients where all metrics were available (2122 patients). This model can be seen as a sort of oracle, since it is allowed to first train over the entire dataset, and then is tested on the same dataset.

Consortium (2009) [3] provides background for how the Warfarin dataset was gathered. They provide figures on which features are most statistically significant and strategies for dealing with missing data: imputation of missing values, omission of data points (patients) with missing values, and treatment of "missing" as an additional discrete value. They also provide a linear regression model for numerical real-valued prediction of Warfarin dosage over the most significant features. We use this model as a baseline in our project as mentioned above, and we provide comparisons in Figure 3.

Li et. al. [4] describes a contextual bandit approach for recommending articles to users in an online fashion based on information about the user. They similarly assume a linear reward model and discrete binary reward. They propose an algorithm named LinUCB, as well as a variant named LinUCB Hybrid, both of which we implement as a first approach in our project.

Bastani and Bayati [1] present the LASSO bandit algorithm, and provide results on the same Warfarin data set we use. The LASSO bandit is an efficient algorithm for working with high dimensional covariates in the contextual bandit setting. Their results show that the algorithm is able to outperform physicians in selecting a correct dose for a majority of patients (the fixed-dose baseline above). We implement this algorithm and provide experimental results of our own.

3 Approach

3.1 Problem Model

The Warfarin dataset provides medical data and the correct dosages for 5528 patients We remove this information from the training data and use it to determine rewards for the environment. We discretize the action space into k = 3 actions; low: less than 21 mg/week, medium: 21-49 mg/week and high: more than 49 mg/week, as defined in Consortium (2009).

The linear reward structure for each arm is $r_t(X_t, a_i) = X_t^\top \beta_i + \epsilon_{i,t}$, where $\beta_i \in \mathbb{R}^d$ is an unknown parameter the bandit algorithm is attempting to learn and $\epsilon_{i,t}$ is centered gaussian noise with constant variance. The bandit objective is to minimize total expected regret,

$$R_T = \sum_{t=1}^{\top} \mathbb{E}[\max_j [X_t^\top \beta_j] - X_t^\top \beta_i]$$

. The actual payoff in the environment is -1 for incorrect and 0 for correct decisions. Since the maximum reward is always 0, the total regret is equal to -1 times the number of incorrect decisions.

3.2 Dataset and Preprocessing

We use a one hot encoding for categorical features: that is, for a feature that can take on n different values, we create n features which are 0 except for the corresponding new feature, which is 1. We also normalize each feature vector and add a constant feature to each vector to allow the model to train a bias term. After dropping variables with under 50% support among the 5528 patients, our final data vectors contain 93 features.

In addition, in order to more accurately compare to the linear regression baseline, we have a separate preprocessing step which we use when comparing with this model. This preprocessing step removes patients with missing data, and only uses the subset of features specified in [3], leaving a total of 2122 patients, each with 12 features.

3.3 Algorithms

3.3.1 LinUCB

This algorithm estimates the expected reward of the current patient using ridge regression, then selects an arm based on the principle of UCB. The data involved for each arms' coefficient's is the subset of patients for which the agent selected that arm. This encapsulates the exploration/exploitation trade-off: selecting a promising arm will likely lead to better short term rewards, at the cost of starving other the arms of data they can use to better estimate their parameters for future patients. The ridge regression estimate is

$$\beta_a = (D_a^\top D_a + I_d)^{-1} D_a^\top c_a$$

and the estimated reward is $x_{t,a}^\top \beta_a$. The estimate of how far the estimated reward is from the expected reward for the patient can be bounded with probability at least $1 - \delta$. Specifically:

$$x_{t,a}^\top \beta_a - \mathbb{E}[r_{t,a}|x_{t,a}] \leq \alpha \sqrt{x_{t,a}^\top (D_a^\top D_a + I_d)^{-1} x_{t,a}}$$

, where $\alpha = 1 + \sqrt{\log(2/\delta)/2}$. The upper confidence bound for arm a is then

$$a_t = \arg\max_{a \in A_t} x_{t,a}^\top \beta_a + \alpha \sqrt{x_{t,a}^\top (D_a^\top D_a + I_d)^{-1} x_{t,a}}$$

. We use an incremental approach to efficiently calculate β_a. The asymptotic upper regret bound of this algorithm is $\tilde{O}(\sqrt{KdT})$ [4].

3.3.2 LinUCB Hybrid

This is a modification to the LinUCB algorithm, which allows for an additional set of parameters to be shared across all arms, as well as a context z that includes information about the arm. The new expected reward under this model is

$$\mathbb{E}[r_{t,a}|x_{t,a}] = z_{t,a}^\top \beta^* + x_{t,a}^\top \theta_a^*$$

, where β^* is shared by all arms, and is the result of ridge regression over all previous patients. In our algorithm, we choose to set $z_{t,a} = x_{t,a}$.

3.3.3 LASSO Bandit

Instead of using an UCB, the LASSO Bandit algorithm uses a forced-sample approach to exploration. At predetermined times, the bandit chooses a certain arm, and at all other times acts greedily. This method produces i.i.d. data from the forced-samples, which can be used as an unbiased estimation of β_i. This creates two sets of data for each arm: the forced sample data T_i, and the all-sample data S_i which includes all patients where that arm was chosen. On each dataset, LASSO regression is used to estimate the parameters for each arm. From the T_i data parameters, a subset of arms is chosen for which the estimate is within $h/2$ of the maximum estimate where h is a hyper-parameter for the algorithm. Then, from this subset of arms, new parameters are estimated using the all-sample data in S_i. From these new parameters, the action is greedily chosen:

$$\pi_t \leftarrow \arg\max_{i \in K} X_t^\top \hat{\beta}(S_{i,t-1}, \lambda_{2,t-1})$$

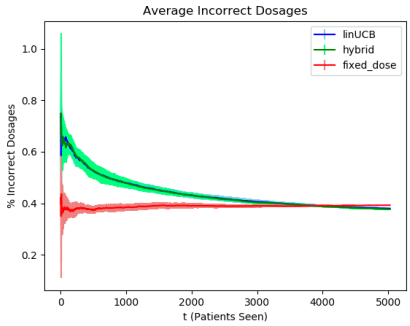
. Under i.i.d. data, LASSO regression is guaranteed to converge. Combining the forced-sampling estimator for the pre-processing step then guarantees convergence of the all-sample estimator. The asymptotic upper regret bound of this algorithm is $O(K[\log T + \log d]^2)$ [1].

3.3.4 XGBandit

We also noted that the LASSO regression algorithm in LASSO Bandit could be replaced with any other regression algorithm of our choice. Since we knew that the XGBoost regression algorithm [2] often outperforms algorithms based on linear regression like LASSO, we decided to see if replacing the LASSO regression step in LASSO Bandit with XGBoost yielded performance improvements.

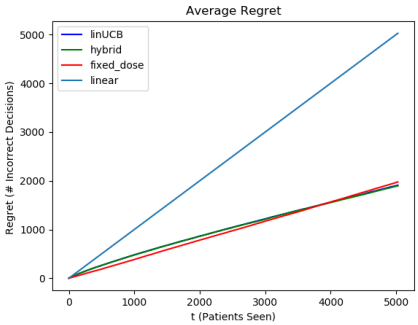
4 Experimental Results

4.1 LinUCB Results



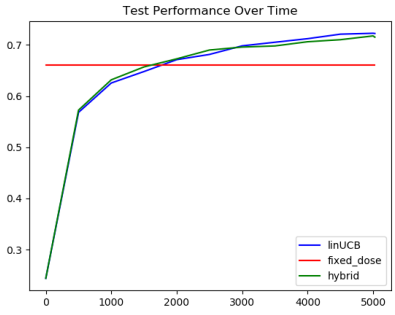
(a)

The number of incorrect decisions as a fraction of the number of patients seen. The shaded regions represent error confidence bounds. The bandit algorithms continue to improve dosage predictions as more patients are observed. Both start to be comparable to the fixed dose baseline around t=300, and are outperforming the baseline by 4000 patients.



(b)

Regret averaged over 10 training data permutations, as a function of patients seen in the data. LinUCB and LinUCB Hybrid exhibit nearly identical regret functions. Both algorithms begin to show sub-linear behavior, as they start to dip under the linear fixed-dose policy around t = 4000 patients.



(c)

Performance on a test set of 500 patients. At intervals of 100 patients, the current policy was tested on this held out set of patients. On the test set, the bandits begin to outperform the baseline much earlier, with fewer than 2000 patients needed.

Figure 1: These plots summarize our experiments with the LinUCB algorithms. A feature vector of size n = 93 was used, and δ was set to 0.1 for all runs and algorithms. A test set of 500 patients was held out, making the training set size 5028. For all graphs, the policies were run over 10 random permutations of the training data, and the runs were averaged per each algorithm.

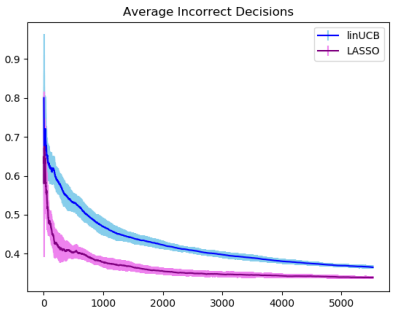
Final Average Metrics: 93 features, 5058 training points, 500 test points			
Algorithm	Average Final Regret	Average Final Incorrect Decision %	Average Final Test Performance
Fixed Dose Baseline	1976.0	0.3929	0.66
LinUCB	1909.3	0.3797	0.7224
LinUCB Hybrid	1908.1	0.3794	0.7152

Figure 2: The averaged final metrics for the graphs above for each policy.

Final Average Metrics: 12 features, 2122 training points		
Algorithm	Average Final Regret	Average Final Incorrect Decision %
Fixed Dose Baseline	771.0	0.3633
S1f Baseline	741	0.3491
LinUCB	787.6	0.3711
LinUCB Hybrid	781.8	0.3684

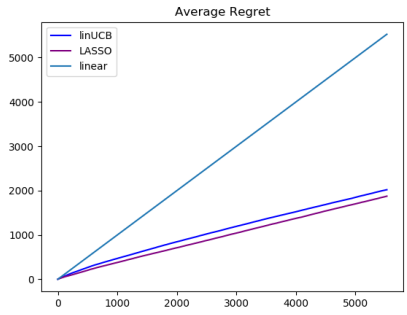
Figure 3: As stated in section 3.2, a different subset of the data was used for comparison to the linear regression model. This table represents the final metrics for each algorithm on this data set.

4.2 LASSO Results



(a)

Comparison of LASSO and LinUCB by number of incorrect decisions as a fraction of the number of patients seen. The shaded region represent error confidence bounds. The LASSO algorithm out performs LinUCB for all time steps, and additionally starts converging much more quickly, indicating it is more data efficient.



(b)

Average regret for LASSO and LinUCB as a function of patients seen in the data. LASSO has lower regret for all time steps, but appears to have roughly the same asymptotic behavior as LinUCB.

Figure 4: These plots summarize our experiments with the LASSO algorithm. A feature vector of size n = 93 was used, and the same parameters as described in [1] were used. Our results are comparable to those described by the authors. For all graphs, the policies were run over 10 random permutations of the training data, and the runs were averaged per each algorithm.

4.3 XGBandit Results

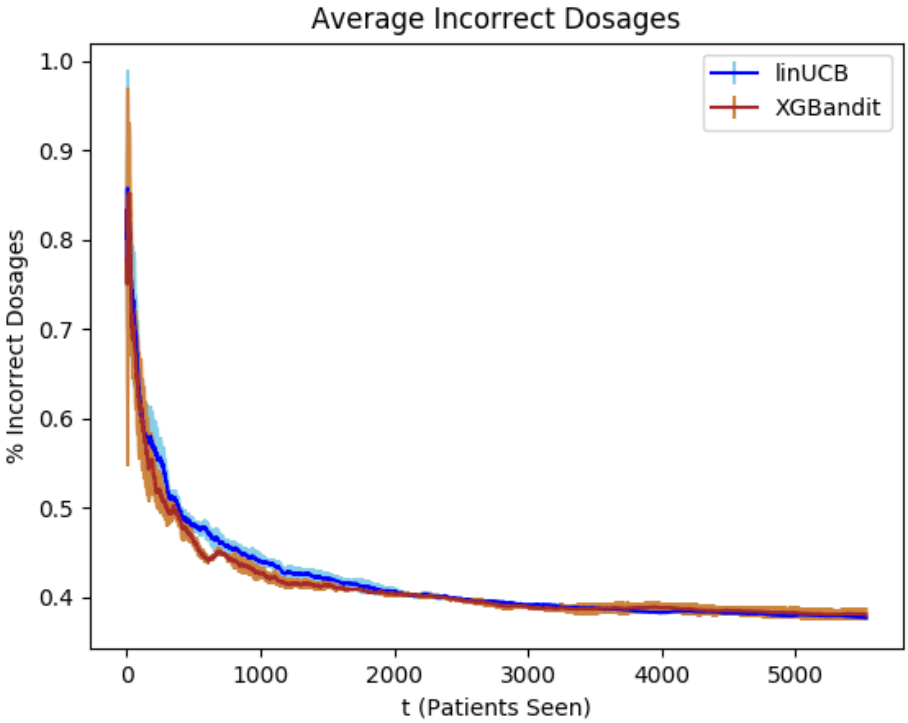


Figure 5
Comparison of XGBandit and LinUCB by number of incorrect decisions as a fraction of the number of patients seen with the same features used in the s1f baseline, average over 3 runs. The shaded region represent error confidence bounds. The XGBandit algorithm trains slightly faster than LinUCB, but LinUCB eventually slightly exceeds the performance of XGBandit.

Final Average Metrics		
Algorithm	Average Final Regret	Average Final Incorrect Decision %
Fixed Dose Baseline	1976.0	0.3929
LinUCB	1909.3	0.3797
LinUCB Hybrid	1908.1	0.3794
LASSO	1874.2	0.3390
XGBandit	1927.7	0.3810

Figure 6

5 Conclusion

From our experiments, we can conclude that Linear Contextual Bandits perform well on the learning problem we have posed using the Warfarin data set, even with the assumptions (reward linearity, Gaussian noise, i.i.d. data) we have made. All bandit algorithms we implemented outperformed the fixed dose baseline as shown in figure 1a and 4a, and seem to exhibit near sub-linear regret in 1b and 4b. Additionally, on a held out test set, the LinUCB algorithms exhibit particularly good performance after observing around 5000 patients (71-72% correct decisions).

In our comparison of LinUCB and LASSO, we determine that the LASSO Bandit performs best under this scenario, outperforming the semi-oracle s1f baseline with additional features, as seen by

the average final incorrect decisions of s1f (.34) and LASSO (.33) in figures 3 and 5 respectively. However, parameter estimation in LinUCB can be computed in closed form and updated efficiently. Additionally, LASSO performs two regressions over subsets of the data, allowing it to converge more quickly as seen in figure 4a, but slightly more computationally expensive. It may be possible to combine the ridge regression in LinUCB with the algorithm in the LASSO Bandit in practice to combine the advantages of both, with some loss of convergence guarantees.

We assumed that the reward is linear, which allows us to explore different properties of linear bandit algorithms on the Warfarin data set and allows for convergence guarantees, however it may be interesting to relax this assumption and use other (non)linear models in future work. Still, linear contextual bandits are a promising direction for medical diagnosis and may be able to be used for many other applications in medicine.

6 Contributions

Ryan authored the fixed baseline, co-authored the data preprocessing script, and LinUCB, LinUCB Hybrid, and LASSO implementations; co-authored the environment and testing script; produced the graphs and tables; co-wrote all sections of the paper.

Jon authored the linear regression baseline, co-authored the data preprocessing script, and LinUCB, LinUCB Hybrid, LASSO, and XGBandit implementations; co-authored the environment and testing script; co-wrote all sections of the paper.

References

[1] Hamsa Bastani and Mohsen Bayati. "Online decision-making with high-dimensional covariates". In: (2015).

[2] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

[3] International Warfarin Pharmacogenetics Consortium. "Estimation of the warfarin dose with clinical and pharmacogenetic data". In: *New England Journal of Medicine* 360.8 (2009), pp. 753–764.

[4] Lihong Li et al. "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 661–670.