Project Final Report - Estimating the Required Dosage of Warfarin

Fengjun Yang (fyang3), Chee Ching Ong (ccong)

December 2019

Abstract

We worked on predicting the therapeutic dosage of Warfarin, a blood anti-coagulant, based on patient's demographic and physiological data. Using the dataset released by the International Warfarin Pharmacogenetics Consortium, we explored the accuracy of various regression and classification techniques. We found that the highest accuracy among the tested models can be achieved using logistic regression with L1 regularization, which gives .769 accuracy, .713 sensitivity, and .812 specificity.

1 Introduction and Motivation

Warfarin is a prescription drug mainly used to treat blood clot-related symptoms such as deep vein thrombosis and to minimize the occurrence of stroke and heart attack in vulnerable patients. It works as an anticoagulant that inhibits blood clotting. Overdosing on Warfarin thus may lead to excessive bleeding. The goal of warfarin treatment is to prescribe the appropriate dosage to decrease the blood clotting tendency, while minimizing the risks of excessive bleeding. However, the amount of dosage is difficult to determine as it depends on a variety of demographic and physiological factors, and thus varies significantly from patient to patient.

We base our study on the data released by the International Warfarin Pharmacogenetics Consortium [1]. The input data contains patients' demographic and physiological data, and their clinically determined dosage. The aim of this project is to train a machine learning model that can classify the appropriate dosage for a patient from her demographic and physiological information into either a "small dosage" class or a "large dosage class". The two classes are divided by the threshold of 30mg per week, the mean therapeutic dose in the dataset. This threshold for categorizing dosage classes is consistent with previous works done in the field by [5, 7]. The reason for modeling the task as a classification problem stems from the fact that Warfarin doses are constantly monitored and adjusted by doctors and patients, and can vary from day to day because of dietary changes. The output of our model would only be used as a estimate of the initial dosage.

We applied both regression and classification algorithms for this task. In the case of regression algorithm, the predicted dosage class is given by binning the continuous output of the regression model. For baselines, we used linear regression and logistic regression, with L1 and L2 regularization. We also experimented with support vector machines (SVM), multi-layer perceptrons (MLP) and Gaussian processes (GP).

2 Related Work

Previous studies have explored both the classification formulation and regression formulation of the problem. For the regression formulation, Cosgun et al [2] and Hu et al [3] used techniques such as k-Nearest Neighbors (kNN), Support Vector Regression (SVR) and Random Forest Regression (RFR) to estimate the dosage required. For the classfication task, Sharabiani et al [5] [6][7] used Neural Network (NN) and Support Vector Machines (SVM) to predict the dosage class. Among these approaches, SVM, SVR and RFR in general performed better than kNN and NN.

A noticeable commonality among these previous studies regardless of their formulation is the employment of regularization techniques. This was necessary as models can easily over-fit given the large feature space. Thus, finding the optimal bias-variance trade-off is crucial in this task. In our experiments, we again confirmed the importance of regularizing the models for this task.

A potential flaw in the previous works lies in the metrics used to verify the performance of the model. The models aimed to achieve either RMSE, MAE, R^2 score, or classification accuracy. However, this fails

to take into account that the adverse effect of overdosing is more acute than that of under-dosing. We argue that a good evaluation metric should reflect the discrepancy in long-term and short-term risks, and a good prediction of initial dosage should more explicitly penalize the short-term risk of overdosing, especially considering the fact that long-term risks can be minimized by continuously monitoring and adjusting the dosage on a daily basis.

3 Dataset and Features

3.1 Dataset

The dataset was obtained from the International Warfarin Pharmacogenetics Consortium [1]. The dataset contains 5528 examples, and a rich set of 62 features, ranging from demographic and background information to genotypic features of the patient. Figure 1 describes some of the key features in the dataset. The physiological significance of the independent variables, in particular the key genotypes, were discussed in [4].

Demographic	Gender, Race, Ethnicity, Age
Background Information	Height, Weight, Indication for Warfarin Treatment, Existing Conditions (e.g. Diabetes), Ongoing Medications (e.g. Aspirin, Antibiotics), International Normalized Ratio (INR)
Genotypic	Cyp2C9 genotypes, VKORC1 genotypes

Figure 1: Feature Examples

3.2 Feature Representations

We encode the features differently to stay consistent to their respective representations in the dataset. We used three different types of encoding. First, data represented as numerical values (eg. Height, Weight, etc.) are encoded into their numerical values. Numerical values that were missing from the dataset were imputed as the median of their respective fields. Secondly, data represented as binned classes (eg. age 10-19, 20-29, ...) are ordinally encoded. Missing values in these fields are encoded as 0, representing the field is missing. All the rest of the features are one-hot encoded, with an additional field denoting missing values. We dropped patient number, medication, and comorbidity as they are unique to each patient and do not generalize to new data points.

3.3 Experiment Setup

We split the data randomly into a training set (80%) and a test set (20%). 10-fold-cross validation was used on the training set to tune hyperparameters and to select the best model, while the test set was used exclusively to evaluate the performance of the chosen model.
The predicted dosage was represented as two classes (0, representing dose D $\leq$ 30 mg/week, and 1, representing dose D $>$ 30 mg/week).

4 Methodology

We implemented a variety of models to compare their performance over the dataset. First, we implemented our baselines using logistic regression. We then implemented four classification algorithms including k-nearest-neighbor (KNN), support vector machine (SVM), multilayer perceptron (MLP), and Gaussian processes classification. In addition, we also tested the performance of regression algorithms by binning their predictions with our specified threshold (i.e., given a regressor prediction, we define the output of our model to be the indicator variable of the prediction is greater than 30). For the regression models, we used linear regression as our baseline, tested the effectiveness of ridge and lasso regression, and also implemented SVMs and MLPs as a comparison. We detail the algorithms below.

Linear Regression
In linear regression, we use the following fitted model to predict the output.

$$h_\theta(x) = \theta^T x$$

θ is fitted by minimizing the residual sum of squares. With regularization, we minimize the following cost function, where λ is the regularization strength.

L1 Regularization

$$\sum_{i=1}^{n}(y^{(i)} - \theta^T x^{(i)}) + \lambda\|\theta\|_1,\ \lambda > 0$$

L2 Regularization

$$\sum_{i=1}^{n}(y^{(i)} - \theta^T x^{(i)}) + \lambda\|\theta\|_2^2,\ \lambda > 0$$

Logistic Regression
In logistic regression, we used the following fitted model to predict the output class probability

$$h_\theta(x) = g(\theta^T x) = \frac{1}{(1 + \exp(-\theta^T x))}$$

θ is fitted by maximizing the log likelihood of our data.

$$l(\theta) = \sum_{i=1}^{n} y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))$$

With regularization, we minimize the following cost function, where λ is the regularization strength.

L1 Regularization

$$J(\theta) = -\frac{1}{n} l(\theta) + \lambda \|\theta\|_1, \ \lambda > 0$$

L2 Regularization

$$J(\theta) = -\frac{1}{n} l(\theta) + \lambda \|\theta\|_2^2, \ \lambda > 0$$

In addition, we used a variable threshold δ for $0 \leq \delta \leq 1$. We predict a positive class if $h_\theta(x) > \delta$. Varying δ would allow us to control the relative performance in accuracy, sensitivity and specificity.

Support Vector Machines
In SVM, we solve the following optimization problem for finding the optimal margin classifier.

$$min_{w,b} \ \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$s.t. \ y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \ i = 1...m$$

$$\xi_i \geq 0$$

We used the Radial Basis Function (RBF) for the kernel function. The RBF is chosen for its ability to map the feature space to an infinite space.

$$K(x, z) = \exp\left(-\gamma \|x - x'\|^2\right), \ \gamma > 0$$

In our model, the hyper-parameters γ and C are varied. γ controls the influence of a single training example, while C controls the regularization strength.

kNN
In k-nearest-neighbors, the algorithm stores the data points in the training set. To classify a previously unseen point, the algorithm finds the k nearest training points around the queried point, and takes a simple majority vote of these k points. We tested $k \in 1, 2, 3, 4, 5$.

Multi-Layer Perceptron
A multi-layer perceptron is a fully-connected feed-forward neural network. In our model we used the ReLU activation function $a(z) = \max(0, z)$. We tested several architectures ranging from one to five hidden layers, with up to 2000 neuron in each layer. We also employed early-stopping (with a 10% of the training data used for validation purposes) as a regularization technique.

5 Experiments and Results

5.1 Evaluation Metrics

The relevant metrics include Accuracy, Sensitivity and Specificity.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

We are more concerned with the scenario of over-dosage, as this would lead to increased vulnerability of a patient to excessive bleeding. In the case of under-dosage, since the condition of a patient would be regularly monitored, the dosage can be increased gradually if necessary. Hence, Specificity is deemed more important than Sensitivity. To choose our best model, we define the overall performance metric P:

$$P = 0.8(Specificity) + 0.2(Sensitivity)$$

5.2 Baseline Models

In the logistic regression model, the threshold value δ is varied from 0 and 1 inclusive and the Receiver-Operating Curve (ROC) is generated. As we increase the threshold, specificity increases, but at the expense of sensitivity, i.e. there is a trade off between these two metrics. The AUC is 0.857, which indicates that the baseline model is a reasonably good model. While AUC gives an overall indication of model performance, it is also important to measure specificity and sensitivity so that we would be able to select particular threshold values to optimize the overall performance metric P.

We also tested linear regression model as a baseline, where the regression output is discretized into two classes (high and low dosage). The confusion matrix in Table 1 summarizes the predictions. The values are normalized by the total number of examples in the validation set. For this baseline, we achieved accuracy, sensitivity and specificity of 0.766, 0.823 and 0.715 respectively.

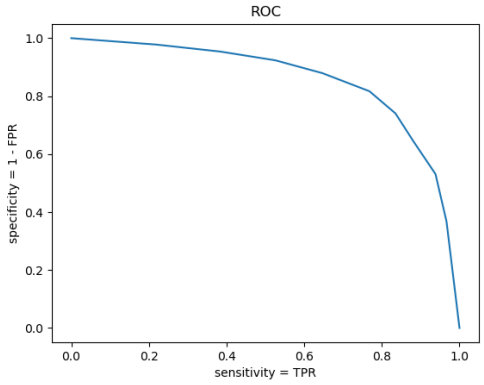

Figure 2: ROC for logistic regression baseline model

Actual / Predicted	Positive	Negative
Positive	0.377	0.155
Negative	0.079	0.389

Table 1: Confusion Matrix for Linear Reg Baseline

5.3 Ridge Regression with Regularization

To cope with the large feature space, regularization was used. For the ridge regression model, as we increased the penalty, the training score decreased. This is expected as the training fit would be the highest for the zero penalty case, which corresponds to ordinary least squares regression. There exists an optimal penalty ($\lambda = 1$) which yielded an optimum R^2 score of about 0.46 on the validation set.
Following the same discretization procedure, the best ridge regression model achieved accuracy, sensitivity and specificity of 0.772, 0.827 and 0.725 respectively.

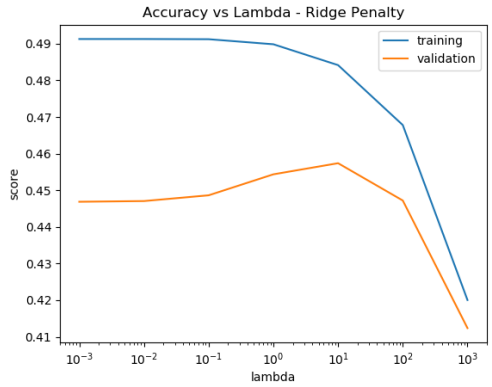

Figure 3: Ridge Regression Results

5.4 LASSO Regularization

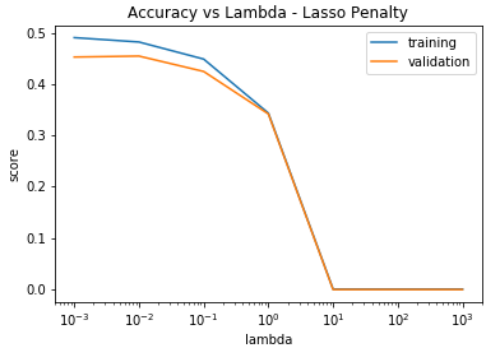

Figure 4: Lasso Regression Results

For the lasso model, the behaviour of the training score was similar to that in ridge, which is not surprising. There exists an optimal penalty ($\lambda = 0.01$) which also yielded an optimum R^2 score of about 0.46 on the validation set.
Following the same discretization procedure, the best lasso regression model achieved accuracy, sensitivity and specificity of 0.772, 0.829 and 0.726 respectively.

5.5 Logistic Regression with Regularization

We used L1 and L2 regularization for the logistic regression models. For each type of regularization, the regularization strength was varied from $\lambda = 0.001$ to $\lambda = 1000$. The ROCs are plotted in Figure 5 for the baseline and regularized models. The model using L1 regularization has the highest AUC of 0.872, as compared to baseline model (AUC = 0.857) and model using L2 regularization (AUC = 0.871).

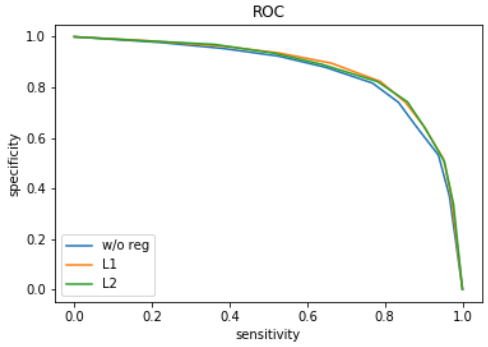

Figure 5: Regularized Logistic Regression Results

5.6 Support Vector Classification

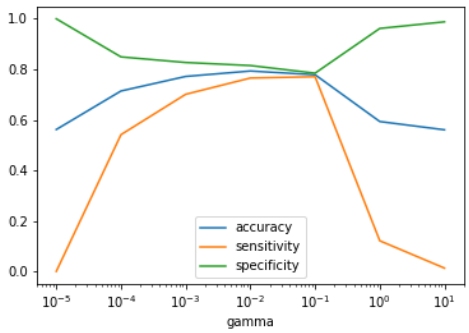

Figure 6: SVM Classifier Results (C=1)

The hyperparameter γ was varied from 10^{-5} to 10^{-1} and the regularization strength C was varied from 10^{-2} to 10^{2}. The results are plotted in Figure 6. The best SVM model used $\gamma = 0.01$ and $C = 1$, and achieved accuracy, sensitivity and specificity of 0.793, 0.765 and 0.815 respectively.

5.7 Comparison of Models

Table 2 summarizes the relative performance of the various models on the validation set. We note a few key observations. Firstly, regularized models tend to perform better, which is not surprising since the feature space is large. Secondly, since the objective is to perform a classification task, classification models generally work better than regression models. This is expected as they are more robust to outliers (outliers has a greater impact on the loss functions in regression models). Thirdly, simpler models such as logistic regression outperform more sophisticated models such as MLP. We attribute this to the noise in the dataset leading to the neural network overfitting to the training set and thus not generalizing as well as logistic regression.

5.8 Benchmarking of Results

The model with the best performance on the validation set is chosen as the best model. The best model is logistic regression using L1 Regularization. The results are benchmarked with those obtained by Sharabiani [5]. Table 3 shows the comparison.

Our best model outperformed Sharabiani's in all the metrics. The key reason is that we included genotypic features (CYP2C9 and VKORC1), which were found to have significant influence on warfarin dosing

Model	Accuracy	Sensitivity	Specificity	**P**
Linear Reg	0.766	0.823	0.715	0.734
Ridge Reg	0.773	0.827	0.728	0.748
Lasso Reg	0.773	0.829	0.726	0.747
MLP Reg	0.786	0.797	0.779	0.783
Log Reg	0.795	0.768	0.817	0.807
Log Reg (L2)	0.805	0.781	0.823	0.815
Log Reg (L1)	0.808	0.789	0.823	0.816
KNN	0.740	0.742	0.738	0.739
SVM Clf	0.793	0.765	0.814	0.804
MLP Clf	0.793	0.752	0.825	0.810
GPC	0.789	0.753	0.817	0.804

Table 2: Comparison of Model Performance on Validation Set

Metric	Best Model	Sharabiani's
Accuracy	0.769	0.66
Sensitivity	0.713	0.63
Specificity	0.812	0.73

Table 3: Benchmarking of Model Performance.

requirements through clinical trials [4]. Not incorporating these genotypic effects might have introduced some omitted variable bias.

6 Conclusion and Future Work

We have trained many predictive models that could reasonably predict the warfarin dosage class given the features of a patient, and logistic regression with regularization performed the best. For future work, we could further improve our models (e.g. neural network) and apply methods such as dropout to better cope with the over-fitting issue. We also plan to try decision tree and random forest on the dataset, as interpretability of the model can be helpful to the doctors for making decisions. Another interesting direction is to run unsupervised learning algorithms on this dataset to extract useful features, and see how that can improve the regression performance.

Contributions

Fengjun performed data pre-processing (standardization, imputation and one-hot encoding) and trained models such as MLP, Gaussian Processes and kNN. Chee Ching worked on logistic regression and SVM models, including the generation of confusion matrix and ROC to evaluate the models.

Note

Fengjun has worked on this dataset for his final project of CS234. In that project, the problem was formulated as a bandit problem, where the dataset was revealed sequentially to the decision maker. For this project, we formulate the problem differently (the training data is revealed to the algorithm at once) and apply different methods. For this project, we also use significantly more features than Fengjun used in his CS234 project.

References

[1] International Warfarin Pharmacogenetics Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

[2] Erdal Cosgun, Nita A Limdi, and Christine W Duarte. High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in african americans. *Bioinformatics*, 27(10):1384–1389, 2011.

[3] Ya-Han Hu, Fan Wu, Chia-Lun Lo, and Chun-Tien Tai. Predicting warfarin dosage from clinical data: a supervised learning approach. *Artificial intelligence in medicine*, 56(1):27–34, 2012.

[4] Nita A Limdi and David L Veenstra. Warfarin pharmacogenetics. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 28(9):1084–1097, 2008.

[5] Ashkan Sharabiani, Adam Bress, Elnaz Douzali, and Houshang Darabi. Revisiting warfarin dosing using machine learning techniques. *Computational and mathematical methods in medicine*, 2015, 2015.

[6] Ashkan Sharabiani, Adam Bress, William Galanter, Rezvan Nazempour, and Houshang Darabi. A computer-aided system for determining the application range of a warfarin clinical dosing algorithm using support vector machines with a polynomial kernel function. *arXiv preprint arXiv:1903.09267*, 2019.

[7] Ashkan Sharabiani, Houshang Darabi, Adam Bress, Larisa Cavallari, Edith Nutescu, and Katarzyna Drozda. Machine learning based prediction of warfarin optimal dosing for african american patients. In *2013 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 623–628. IEEE, 2013.

Link to Github

The codes used for data analysis is uploaded to Github. The link is https://github.com/FJYang96/CS229-Project