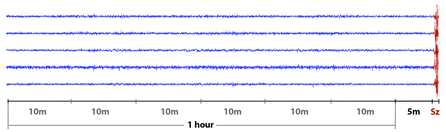
Predicting Seizure Onset in Epileptic Patients Using Intracranial EEG Recordings

Janet An, Amy Bearman, Catherine Dong

Introduction & Data

Intracranial EEG (iEEG)-based monitoring systems have the potential to predict seizure onset. However, even with this data, *preictal* (1 hour before seizure) states are often difficult to distinguish from *interictal* (non-seizure) states. Our goal is apply ML classification methods to this EEG data to more accurately model the probability that a patient is in a preictal state.

The iEEG data is organized into 10-minute clips labeled as preictal or interictal. Each data clip contains a (# electrode channels × # readings) matrix with 15 electrode channels and ~240,000 iEEG readings for each clip. We used 5 hours of interictal and 5 hours of preictal data for training.



Features

In addition to univariate features, we computed two bivariate features of EEG synchronization on 2 channels over 1-sec windows: maximal cross-correlation (MCC) and Euclidean distance. Cross-correlation is a linear measure of dependence between two channels, which allows for fixed time delays to account for the propagation of brain waves. We retain the maximum cross-correlation:

$$C_{a,b} = \max_{\tau}\left\{\left|\frac{C_{a,b}(\tau)}{\sqrt{C_a(0)\cdot C_b(0)}}\right|\right\} \text{ where } C_{a,b}(\tau) = \begin{cases} \frac{1}{N-\tau}\sum_{t=1}^{N-\tau} x_a(t+\tau)x_b(\tau) & \tau \geq 0 \\ C_{b,a}(-\tau) & \tau < 0 \end{cases}$$

as well as the non-delayed cross-correlation between the two signals.

Euclidean distance measures the distance in state-space between the trajectories of two EEG channels. First, each channel was time delay-embedded into a trajectory with a time delay of 6 readings and an embedding dimension of 10. Then, we compute the distance of each time-delay embedded vector to its *K* nearest neighbors in state space:

$$\frac{1}{N}\sum_{t=1}^{n}\frac{\frac{1}{K}\sum_{k=1}^{K}\|x_a(t)-x_a(t_k^a)\|_2^2}{\frac{1}{K}\sum_{k=1}^{K}\|x_a(t)-x_a(t_k^b)\|_2^2}$$

Model

MODEL SELECTION: We trained our data on various logistic regression (LR) models (using LIBLINEAR), SVMs with different kernels (using LIBSVM), and K-nearest-neighbors. The results of our testing are shown in the table below. We found that L2-regularized LR performed the best (after parameter selection) in terms of accuracy and speed.

Model	Training Accuracy	5-Fold CV Accuracy	# Iterations
L2-reg LR (primal)	86.2472	85.7194	16
L2-reg, L2-loss, linear kernel SVM (dual)	—	—	reached max iters
L2-reg, L2-loss, linear kernel SVM (primal)	85.9361	85.6694	15
L2-reg, L1-loss, linear kernel SVM (dual)	—	—	reached max iters
L1-reg, L2-loss, linear kernel SVM	86.2639	85.8444	52
RBF kernel SVM	100	50	1800
Sigmoid kernel SVM	50	50	1800
KNN	100	54.2	—

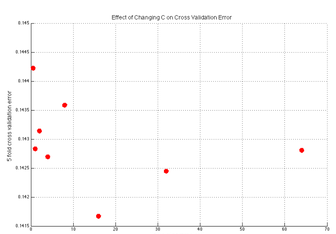
LOGISTIC REGRESSION: L2-regularized logistic regression solves the following unconstrained optimization problem:

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\log(1+e^{-y_i\boldsymbol{w}^T\boldsymbol{x}_i})$$

LIBLINEAR uses Newton's method to optimize the weights:

$$\theta := \theta - H^{-1}\nabla_\theta \ell(\theta)$$

PARAMETER SELECTION: For regularized logistic regression, the only parameter to select is the value of *C* in the formulation for the optimization problem. We first performed coarse grid search by running the algorithm using exponentially growing sequences of values for *C* and comparing the cross validation errors. After determining a "better" range of *C*, we performed fine grid search and deduced that *C* = 16 was the optimal value, although the improvement was marginal.



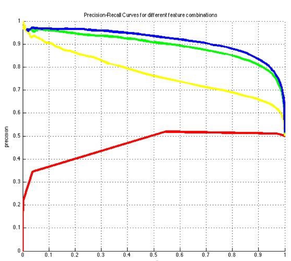
Results & Analysis

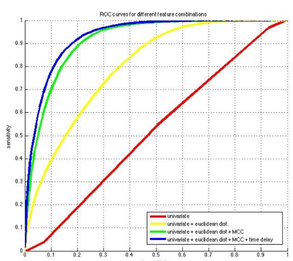
SINGLE FEATURE PERFORMANCE: The table below shows the accuracies from running L2-regularized LR on each of the feature types individually. It includes the baseline features (univariate), linear features (Euclidean, MCC), and nonlinear features (similarity, Spearman correlation, Pearson correlation).

Feature Type	Training Accuracy	5-Fold CV Accuracy
Univariate	52.0222	51.3694
Euclidean distance	70.4472	70.1194
MCC	78.6972	78.4000
MCC with time delay	54.9806	54.2750
Similarity	52.6000	52.2167
Spearman	76.7472	76.4944
Pearson	77.1833	76.9444

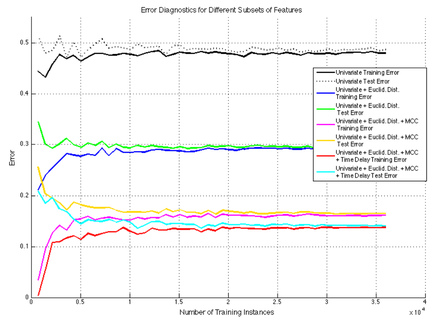
ABLATIVE ANALYSIS: Adding Euclidean distance to the baseline univariate features significantly improved accuracy by 37%. Concatenating cross-correlation over the maximal time delay also improved accuracy by 22%. Adding the Pearson and Spearman correlation coefficients did not improve accuracy, which makes sense, because we would not expect that adding similar measurements of the same relationship would improve test accuracy. Therefore, the features that had the greatest effect on accuracy were maximal cross-correlation and Euclidean distance, so we retained these features.

Feature Combination	Training Accuracy	5-Fold CV Accuracy
Univariate + Euclidean + MCC + MCC w/ time delay + Pearson + Spearman	85.9917	85.4944
Univariate + Euclidean + MCC + MCC w/ time delay + Pearson	86.0111	85.6583
Univariate + Euclidean + MCC + MCC w/ time delay	85.9500	85.7250
Univariate + Euclidean + MCC	83.7639	83.2917
Univariate + Euclidean	70.4472	70.1194
Univariate	52.0222	51.3694





ERROR DIAGNOSTICS: To diagnose the bias vs. variance of our learning algorithm, we plotted the training error and the test error for different training set sizes, starting with just the univariate features. Because the training error and test error both plateaued with barely any gap between them, we confirmed that we did not have high variance, and a larger training set would not increase accuracy. However, because we had high bias, we knew that adding more features could help. By concatenating together four of our features types, we were able to reach more desirable performance.



Future Work

We would like to determine the optimal threshold at which logistic regression marks a data segment as preictal instead of interictal (the default threshold is 0, or 50% probability), in order to achieve optimal precision and recall. We would also like to try different models and feature combinations on multiple subjects, since we did not get a chance to experiment with more than one subject and the International Epilepsy Electrophysiology Portal has many more intracranial EEG datasets.

References

C.-C. Chang and C.-J. Lin, LIBSVM : A Library for Support Vector Machines.

Forrest Sheng Bao, Xin Liu, and Christina Zhang, "PyEEG: An Open Source Python Module for EEG/MEG Feature Extraction."

Mirowski, P., Madhavan, D., LeCun, Y., and Kuzniecky, R. Classification of Patterns of EEG Synchronization for Seizure Prediction.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification.