CS229: Machine Learning

Event Identification in Continues Seismic Data

Please print out, fill in and include this cover sheet as the first page of your submission. We strongly recommend that you use this cover sheet, which will help us to get your graded homework back to you more quickly, as well as help us with tracking submissions.

Please mark the submission time clearly below. It is an honor code violation to write down the wrong time.

If you are submitting this homework late: Each student will have a total of **seven free late (calendar) days** to use for homeworks, project proposals and project milestones. Once these late days are exhausted, any assignments turned in late will be penalized 20% per late day. However, no assignment will be accepted **more than four days** after its due date, and late days cannot be used for the final project writeup. Each 24 hours or part thereof that a homework is late uses up one full late day.

On-campus (non-SCPD) students: Please either hand in the assignment at the beginning of class on Wednesday or leave it in the submission cabinet on the 1st floor of the Gates building, near/outside Gates 188 and 182.

Name1: Alex Hakso
SUNet ID: ahakso (05526504)

Name 2: Fatemeh Rassouli
SUNet ID: frasouli (05734544)

Section 1: Introduction

Injection of saline wastewater by the oil and gas industry has resulted in earthquake rate increases in much of the central and eastern United States. The changes have been significant: Oklahoma has experienced an approximately 200-fold increase in frequency of earthquakes greater than magnitude 4.0 since 2009 (Walsh & Zoback, 2015). The causal mechanism is poorly understood, complicating risk assessment and mitigation. The earthquakes themselves provide the primary window into the nature of the relationship between wastewater injection and induced seismicity. Accordingly, robust and efficient earthquake detection is an important component of addressing increased seismic risk associated with wastewater injection. The **input** to our algorithm is a large number of short time series amplitude recordings. We then use a variety of machine learning algorithms to **output** a predicted classification: the time series does/does not contain seismic energy originating from an earthquake.

Section 2: Related Work

Earthquake detection is a longstanding problem in geophysics, and the research comprises hundreds of approaches, spanning decades. Indeed, the features driving our algorithms draw directly from the literature. Early event detection algorithms searched for anomalous amplitudes over short time periods in the time series, using, for example, a ratio of short term to long term average amplitudes, as documented by Freiberger (1962). This method rarely produces false positives when the monitoring location is relatively free from impulsive noise; seismic monitoring locations are chosen largely to accommodate this consideration. A weakness of this approach is that low magnitude events are often below the detection threshold for this method, producing many false negatives.

To compensate for this difficulty, a particularly effective method of identifying low amplitude signals in a relatively high amplitude noise environment is template matching with cross correlation, known as a matched filter (Anstey, 1964). The shortcoming of this method is that the results are sensitive to the form of the master waveform used as a template. The form of the target wave is a product of the nature of the source producing the signal, as well as its location. Generally, these are unknown quantities, making the success of this method dependent on informed template selection. Templates are often chosen using a high amplitude event as a model or by developing a synthetic waveform based on the Green's function associated with an estimated earth model. Schaff & Waldhauser (2010) used a dataset near Parkfield, CA to demonstrate that a matched filter technique can allow for an improvement for earthquake detection threshold to approximately an order of magnitude lower energy seismic events.

Gibbons & Ringdal (2006) employed an array of seismometers in combination with cross correlation techniques to achieve a further improved detection threshold. When available, using multiple seismometers is advantageous. The magnitude of completeness threshold for our study area (the magnitude at which all events of this size or larger are detected) is estimated to be approximately MM 2.1 (Llenos & Michael, 2013). This provides a benchmark goal, albeit a rather conservative one. The goal of this research project is to build on existing techniques to achieve high event detection reliability using a single seismometer.

Section 3: Dataset and Features

The data used in this study was gathered by the United States Geological Survey in Guy Arkansas using a continuously deployed seismometer in Guy, Arkansas (network AG, station WHAR). The data was cleaned and low-passed to 20 Hz by Clara Yoon and Greg Beroza. Acceleration amplitudes were sampled at 100 Hz on 3 components: vertical, N-S, and E-W. The portion acquired for this study was 1 month of recording, although a week was found to be acceptable for training and testing the algorithm. This week contained 60,480,000 samples, which were divided into 4 seconds windows, which we overlapped by 3 seconds, resulting in a total of 604,797 candidate time windows. 2,878 events had been identified by Clara Yoon and Greg Beroza using their patented "Efficient Similarity Search of Seismic Waveforms" (Beroza et al., 2015) methodology. As discussed in section 2, the chosen four features drew on existing literature for event detection. They were: variance, ratio of short term average amplitude to long term average amplitude, ratio of low frequency energy to high frequency energy, and maximum match filter value.

1. Match Filter Value

Matched filters are an effective method of identifying signals buried in noise, providing a reasonable approximation to the target signal is used. To calculate this feature, each window is cross-correlated against a set of templates. This work employs a set of templates that is highly representative of the waveforms produced by earthquakes in the area. The templates were developed using **unsupervised machine learning** to cluster the training events based on similarity of waveforms. **Agglomerative hierarchical clustering** was used to group the events into maximally

dissimilar groups based on a dissimilarity matrix produced by cross correlating each event against every other. The three components provided three dissimilarity dimensions, which were weighted by the signal to noise ratio as determined to STA/LTA. Once the event had been clustered by similarity of waveform, the templates were formed by stacking several representative events within each group at maximum cross correlation value to reduce the impact of noise. Once each template had been cross correlated with a window, the maximum matched filter value was taken. Viewing the values assigned by classification clearly shows that the feature calculated by this rather involved method produced exceptional results for discriminating between events and non-events.

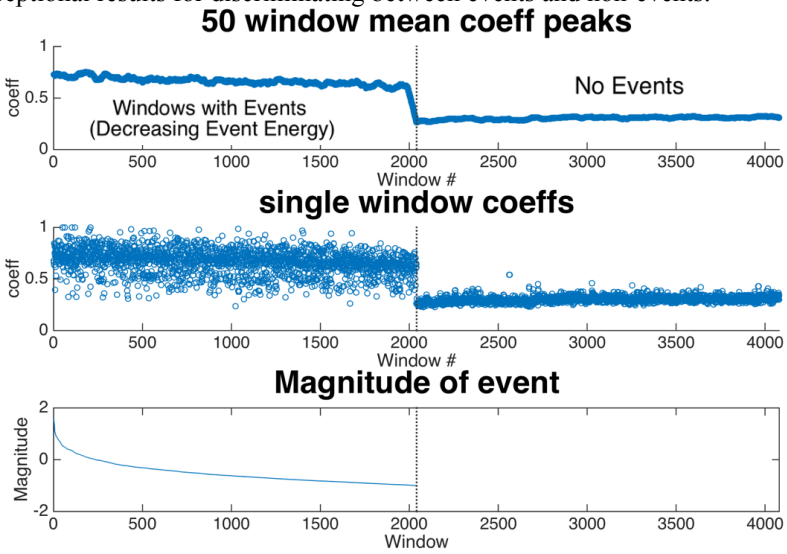

Figure 1: Event windows of the left. Non-event windows on the right are a small, random subset of total non-event windows.

2. Variance

Fundamentally, seismic energy arriving on seismometers increases the magnitude of the amplitude recorded. Variance is a convenient measure of the average amplitude, and the squared component of the term brings it to a 1:1 correspondence with energy in the ambient seismic field, as energy is proportional to $\sqrt{amplitude}$. The arrival of seismic energy originating with an earthquake is one of many causes which may cause an increase in amplitude. A common examples is nearby vehicle or foot traffic.

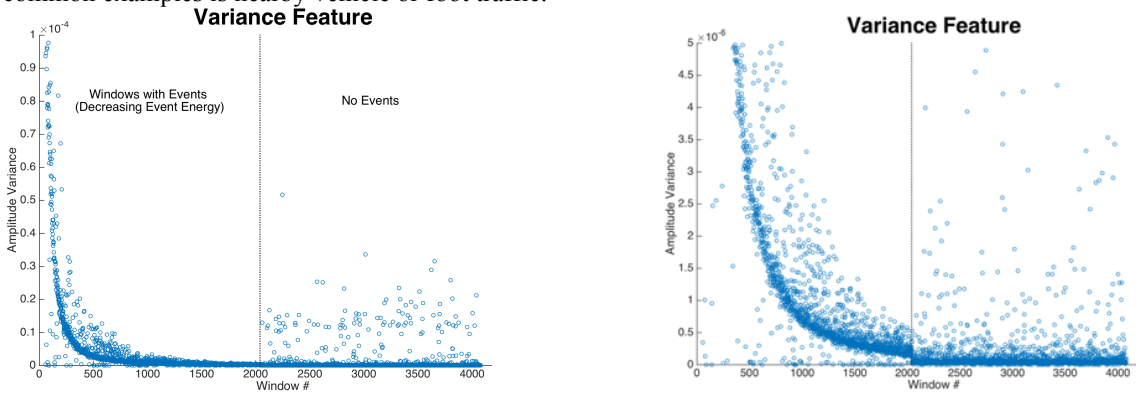

Figure 2: Left: Large events have unambiguously high variance, while smaller events are distributed more similarly to non-event windows. Right: y range is reduced

3. STA/LTA

This feature identifies anomalous amplitudes relative to immediately adjacent time signals. This feature is designed to distinguish between long-term high amplitude noisy periods and impulsive transient amplitude increases (Freiberger, 1962).

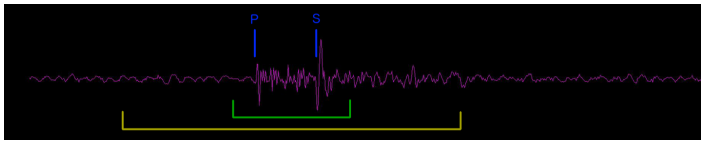

Figure 3: Here, the average abs (amplitude) of the window indicated by the green bracket is substantially higher than the average abs(amplitude) of the window indicated by the yellow bracket. This yields a high STA/LTA value.

4. Spectral Content

Noise on seismometers located on the surface of the earth is generally relatively high frequency. Using $\frac{\sum_{f=0}^{10} log\, E_f}{\sum_{f=12}^{20} log\, E_f}$ where E_f is the energy in the f Hz range, a measure of the relative energy in the signal is obtained. A simple fft with a Hamming window is employed in MatLab to calculate spectral content.

Section 4: Methods (An Introduction to Statistical Learning; James et al.; Springer; 2014)

In this work, we used classification methods, including "Logistic Regression", "Naïve Baye's", "KNN" and "SVM," as our learning methods. These methods are briefly described as follows:

1- Logistic Regression

In this method, $p(X) = Pr(Y=1|X)$ is calculated using a logistic function (Eq 1), outputting a value between 0 and 1 for all values of X. Maximum likelihood is then used to fit the model (Eq 2).

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{1}$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = e^{\beta_0+\beta_1 X} \tag{2}$$

2- Naïve Bayes

This model follows Bayes rule (Eq 3) to model $p(X|Y)$ using the strong assumption that the features are conditionally independent given y (Eq 4).

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)} \tag{3}$$

$$p(x \mid y) = \prod_{i=1}^{n} p(x_i \mid y) \tag{4}$$

3- KNN (K- Nearest Neighbor)

Since conditional distribution of label Y given a feature X is mostly unknown for real data, computing the Bayes classifier is impossible. Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability. One such method is the K-nearest neighbors (KNN) classifier. Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0 in feature space. This set is termed N_0. KNN then estimates the conditional probability for class j as the fraction of points in N_0 whose response value equals j:

$$\Pr(Y = j \mid X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i = j) \tag{5}$$

Finally, KNN applies Bayes rule and assigns the test observation x_0 to the class with the largest probability.

4- SVM (Support Vector Machine)

For this project, we used SVM methods with a variety of Kernel and objective functions, as listed below:

$$\min \quad \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \qquad \text{C-Classification} \quad (6)$$

$$s.t. \quad y_i(w^T\phi(x_i)+b) \geq 1-\xi_i \text{ and } \xi_i \geq 0, i=1,...,N$$

$$\min \quad \frac{1}{2}w^T w - \upsilon\rho - \frac{1}{N}\sum_{i=1}^{N}\xi_i \qquad \text{Nu-Classification} \quad (7)$$

$$s.t. \quad y_i(w^T\phi(x_i)+b) \geq \rho-\xi_i \text{ and } \xi_i \geq 0, i=1,...,N \text{ and } \rho \geq 0$$

$$K(x_i,x_j) = (x_i^T x_j + 1)^q \qquad \text{Polynomial Kernel} \quad (8)$$

$$K(x_i,x_j) = e^{-\gamma\|x_i-x_j\|} \qquad \text{Laplace Radial Kernel} \quad (9)$$

Section 5: Experiments/ Results/ Discussion

Excluding the agglomerative hierarchical clustering associated with feature calculation, which was carried out in Matlab, we performed all of our data analyses using R language. Since our data is very skewed, we used precision and recall functions to compare the results of different methods.

$$True\ Positive / (True\ Positive + False\ Positive) \qquad \text{Precision (10)}$$

$$True\ Positive / (True\ Positive + False\ Negative) \qquad \text{Recall (11)}$$

The first challenge was picking the right size of test and training datasets. To find the best size of the data, we used "e1071" package in R to run Naïve Bayes model on different sizes of training data set ranging from 67% to 99% of the total data set. The purpose of running the Naïve Bayes is that, because of its strong assumptions, this method has a low computational cost. The calculated precision and recall for this method is shown in Figure 4. Note that in this figure, we tested two types of features: original and normalized.

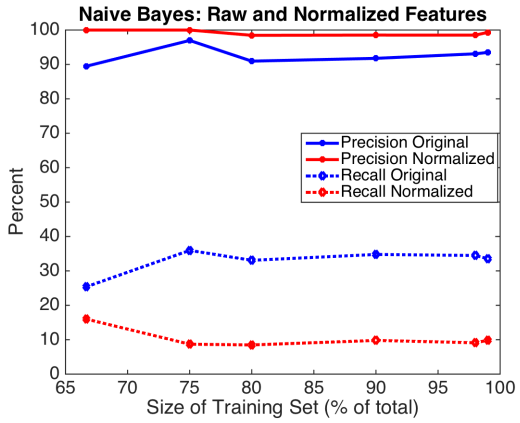

Figure 4: Naïve Bayes model fitting over different sizes of training datasets, using original and normalized features.

The most changes are observed for 67%, 75% and 80% of the total data. We used these fractions for the rest of our study. Also, the values of recall for the predicted labels from normalized features are very small. Considering the trade off between precision and recall values, we decided to use our data with original features.
Next, we tested various SVM methods. Figure 5 shows the results of our classification using C-classification SVM and Polynomial kernel equation. Here, we have changed both the training data subset number and the degree of polynomial. The recall of these methods is still low, showing that the model is not detecting a good fraction of events. This is in fact not acceptable in our context, as false negatives are costly in characterizing earthquake swarms. Attempting to improve recall, we varied the kernel functions. Additionally, we implemented nu classification SVM, which has to the potential to handle skewed data more robustly. Results of these methods, in addition to the results of logistic regression, are provided in Figure 6.

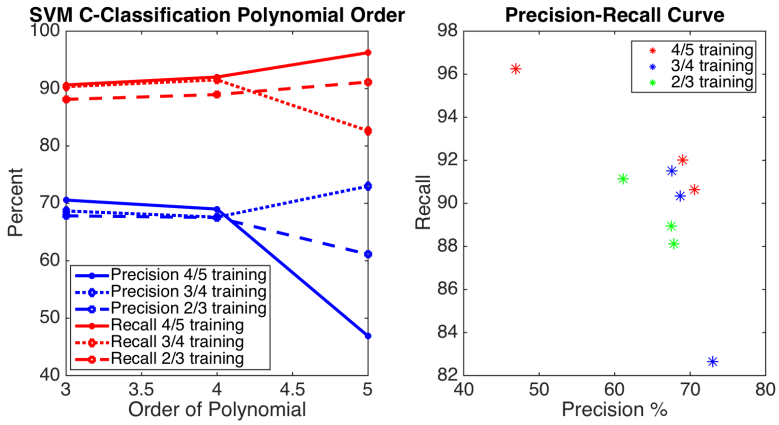

Figure 5: Polynomial SVM classification over different subsets of training dataset.

The recall value of all these methods ranges from 58% to 80%, showing the low performance of these methods.

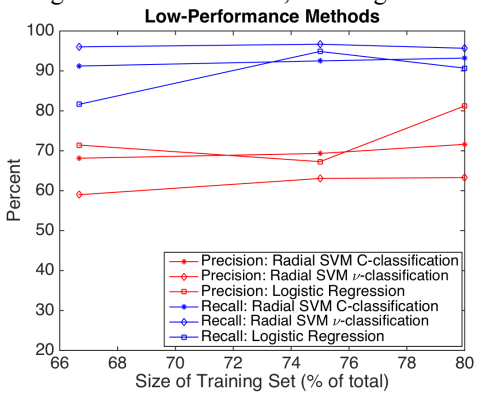

Figure 6:Precision and recall of the low performance methods

Finally, we ran KNN method over various subsets of the dataset. We chose K values of 4, 6, 8 and 10. Remarkably, in each case, the precision was 100%. The recall value as a function of chosen K is shown in Figure 7. Note that for each value of K, the recall value is in excess of 99.5%. Since we achieved impressive recall and precision with K=6, which incurs acceptable computational cost, we concluded that the optimal model is found using KNN with K=6.

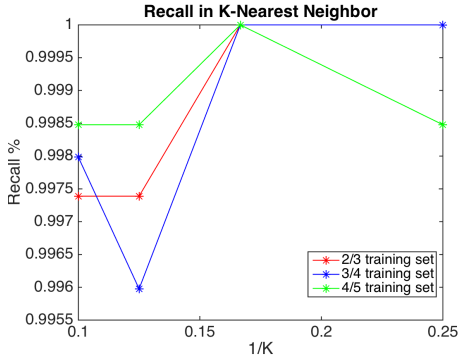

Figure 7: The resulting recall from performing KNN classification.

Section 6: Conclusions/Future Work

Applying KNN with the features outlined has been demonstrated to be an efficient and practical approach for robust earthquake detection, exceeding the baseline detection threshold in Arkansas by three orders of magnitude. These results have been well received in the geophysics department, and we expect to present internally next quarter. Our methodology holds promise for enabling a more detailed analysis of induced seismic events in Oklahoma and Arkansas. Moving forward, we intend to explore data normalization effects more thoroughly, and fine-tune parameters involved in feature calculation.

References

Anstey, N. (1964). Correlation Techniques–a Review*. *Geophysical Prospecting*, *12*(4), 355–382. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2478.1964.tb01911.x/abstract

Beroza, G. C., O'Reilly, O. J., Yoon, C. E., & Bergen, K. (2015). Efficient Similarity Search of Seismic Waveforms. United STates.

Freiberger, W. F. (1962). *An approximate method in signal detection*. DTIC Document.

Gibbons, S. J., & Ringdal, F. (2006). The detection of low magnitude seismic events using array-based waveform correlation. *Geophysical Journal International*, *165*(1), 149–166. doi:10.1111/j.1365-246X.2006.02865.x

Llenos, A. L., & Michael, A. J. (2013). Modeling earthquake rate changes in Oklahoma and Arkansas: Possible Signatures of induced seismicity. *Bulletin of the Seismological Society of America*, *103*(5), 2850–2861. doi:10.1785/0120130017

Schaff, D. P., & Waldhauser, F. (2010). One magnitude unit reduction in detection threshold by cross correlation applied to parkfield (California) and China seismicity. *Bulletin of the Seismological Society of America*, *100*(6), 3224–3238. doi:10.1785/0120100042

Walsh, F. R., & Zoback, M. D. (2015). Oklahoma ' s recent earthquakes and saltwater disposal. *Science Advances*, (June), 1–9. doi:10.1126/sciadv.1500195

The facilities of IRIS Data Services, and specifically the IRIS Data Management Center, were used for access to waveforms, related metadata, and/or derived products used in this study. IRIS Data Services are funded through the Seismological Facilities for the Advancement of Geoscience and EarthScope (SAGE) Proposal of the National Science Foundation under Cooperative Agreement EAR-1261681

The concatenated data was provided by Clara Yoon.