Narrative of Analysis for Time Series Consideration in Diabetes Predictors

August 23, 2017

1 The Question, The Process

We seek to explore the efficacy in considering the timing of tests, diagnoses, and conditions in health insurance claims on predicting whether a patient has diabetes or not. Approaching this subject will involve gradual steps. Working with the data from ≈ 43,000,000 people who have health insurance claims through Aetna, we seek to extract predictors of diabetes from the variety of ICD and CPT filings in these records.

1.1 Time Independent Diabetes Predictors

Before trying to extract time-related information regarding whether a patient has diabetes, we want to establish that there are inherent predictors in the Aetna database for having diabetes. We want to establish which of these predictors is the most effective in suggesting diabetes.

1.1.1 Data for initial steps

We will create a sparse feature matrix across subjects that will indicate the issuance or non-issuance of certain tests/screenings. We will take a population of patients such that there are some with diabetes and some without, and seek to work with hetergenous data – not specific to location, race, gender, etc. Each patient will be a row in the feature matrix, and each column will indicate the presence or absence of one of the tests.

1.1.2 Model Fitting

Since the data are quite categorical, and the final conclusion is as well, perhaps a logistic regression or SVM/margin based classifer would be useful for this classification (for soft margin/hard margin info: here).

- The relative quality of the logistic regression model can be measured with an Akaike information criterion (AIC) for it, defined as: AIC $= 2k - 2\ln L$, where k is the number of features and L is the likelihood function. We seek to maximize the log likelihood of the data (for computation purposes this is easier) where L is:

  $$\ln L(\mathbf{w}|\mathbf{X}) = \prod_i \ln P(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_i \ln(\sigma(\mathbf{w}, \mathbf{x}_i)^{y_i}(1 - \sigma(\mathbf{w}, \mathbf{x}_i))^{1-y_i}, \tag{1}$$

  $$\text{where} \quad \sigma(\mathbf{w}, \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}} \tag{2}$$

  One can either maximize the this function, or minimize the negative of it, essentially asking how likely are the regression parameters **w** given the data **X**?

- The SVM approach with a soft margin involves minimizes the loss defined as:

  $$\mathcal{L} = \frac{1}{n}[\sum_{i=1}^{n} \max 0, 1 - y_i(\mathbf{w} \cdot \mathbf{x} - b)] + \lambda||\mathbf{w}||^2 \tag{3}$$

  where b is the coordinate intercept for hyperplanes at $\mathbf{w} \cdot \mathbf{x} - b = 0$ and the λ is penalizing term for overfitting weight coefficients.

Both of these can be optimized with gradient descent to find the optimum weights to (a) maximize likelihood or (b) minimize loss.

The steps above help to optimize the weights for maximum predictability. We can see which features have the most influence on finding the maximum likelihood or minimum loss by withholding them iteratively and seeing how that influences the optimization. One thing to consider is that the features could be correlated, and removing one will not have as great an impact as removing them both, for example. It would be hard to quantify which are correlated with this method. Will have to think more about this.

1.2 Initial Test of Time Dependence

We want to see if some of these ICD and CPT filings can be used ahead of time to predict that diabetes will be onset later. As an initial test, we can see if there is a correlation between future diabetes diagnosis and the presence of some of the highly indicating features we derived from above appearing earlier in their health insurance claims.

1.2.1 Data

We want to look at patients with continuous health insurance claims from 2012-2013. We want to mitigate enrollment bias and bias from patients that appear in the claims file right as they are diagnosed with diabetes, for example (nothing to really predict there, but they are useful for the previous step).

1.2.2 Tests

Eventually, we would like to implement a method that infer what state the patient is in (diabetic, non-diabetic) as a latent variable. Using a hidden markov model, information about the observables of a patient at an earlier time can be used to infer what ”state” the patient is in at later times or what state they were in at earlier times. We can employ this technique in intervals of years or whatever the smallest continuous increment of health insurance claims are with a viable data set size. The quality of this model can be measured using the forward and backward algorithms on test data to maximize likelihood.