Project Title: Predicting Medicare/Medicaid Cost Fluctuations Over Time
Project Category: Finance and Commerce
Group: Jonathan Lin (jolituba), Michael L. Smith (msmith11), Aileen Wang (aileen15)

Motivation:

In 2015, Medicare and Medicaid spending totaled nearly 1.2 trillion dollars[1]. However, spending per beneficiary per condition is not uniformly distributed across states (and counties). For example, in 2012, Medicare would pay about $11,400 in Kansas for a heart attack (acute myocardial infarction) for a given beneficiary. In Oregon, in 2012, Medicare would pay 12,500. In 2015, the cost per beneficiary in Kansas was $10,800 (a decrease of ~6%) whereas in Oregon, the cost was $14,800 (an increase of 18%). We can compare this to the national average increase (over the span 2012-2015) of treating heart attacks at 4.7%. We've found that cost percentage increase disparities can be found across various health conditions based off our preliminary study of Medicare/Medicaid data, and we are interested in classifying which hospitals/counties/states can be expected to have cost percentage increases (or decreases). We believe that hospitals, state and local governments may benefit from being able to input their own situational features and receive a classification, with which leaders could make a more informed decision on how to mitigate large increases in Medicare spending.

We found a useful dataset for tackling this problem here: https://data.cms.gov/mapping-medicare-disparities. The user interface allows us to download data for a particular year, area, condition, age range, gender, etc, which will be useful in partitioning our data into different segments.

We also looked at past project from CS 229 to look for similarities and differences between our project and previous ones.
http://cs229.stanford.edu/proj2016spr/report/029.pdf
http://cs229.stanford.edu/proj2015/291_report.pdf
What we found is that previous projects focused on estimating general Medicare costs based on features such as state, county, age, gender, race, etc. However, our project primarily aims to look at percentage increase in cost from one year to the next, rather than the general Medicare cost, which is a new angle to a semi-familiar problem.

Method and Intended Experiments:

The features for our project will be the ones given to us in the dataset (age range, gender, county, state, etc.). Given these features, we will try to predict the percentage increase in

---

[1] https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html

healthcare cost from one year to the next. Our goal is to examine the increase for a given county so that patients can make informed decisions about their financial/medical future, and so that non-patient parties (insurance providers, hospitals, etc.) can see where improvements can be made.

We plan to use primarily the supervised learning techniques that we have covered in this course. Since we plan on being able to predict the increase in cost based on features, linear regression will likely be the first technique to try, followed by polynomial regression and more complex models. Furthermore, we plan on being able to classify whether the cost will increase or decrease by a "large" or "small" percentage relative to the national average percent fluctuation, where "large" is above the national average and "small" is below. This is a multi-bucket classification problem, so softmax regression seems appropriate here. Obviously, these are basic techniques that can be improved greatly, but they will be our starting points for our algorithms. Once we evaluate the performance of these algorithms, we can decide where and how to improve. Finally, we hope to be able to perform a comparative study on this problem between classical machine learning and neural networks. However, since neural networks are not the primary focus of this class, this will not be a priority and may not be a task that we tackle.

Since we have a fairly large dataset, we should be able to use hold-out validation to test our algorithm. More detailed and specific strategies for testing should become more evident after subsequent lectures and as we begin to write our algorithms.

To improve our classification results, we want to find the optimal set of features that give us the best prediction of medical cost percent increase. We will use this to reduce the number of attributes we use and optimize our supervised learning algorithms; furthermore, hospitals may be able to analyze controllable features to decrease costs.