ANALYTICS X PRIZE

JOHN MYLES WHITE

1. Naive Models

The first thing we should do is to consider a series of progressively more meaningful naive models to get a sense of the accuracy of the lowest possible quality predictions we can give. This allows us to see how low the baseline RMSE is for this problem without clever models.

At a minimum, it is safe to assume that we must start with the empirical proportions of homicides that have occurred in each of Philadelphia's 47 zipcodes for a sample of recent years. Fake data of this sort is available in the `fake_input_data` directory. These CSV files provide data that looks like the following:

Table 1. Historical Homicide Data

Zip	Year	Probability
11911	2009	0.5
...	...	...
11911	2007	0.5

1.1. Naive Model 1.

With this historical data, we can construct our most naive model: we assume that the probability of a homicide occurring in zipcode Z in 2010 is equal to the probability of a homicide in zipcode Z in 2009:

$$p(H|Z, Y = 2010) = p(H|Z, Y = 2009)$$

More generally, we assume that

$$p(H|Z, Y) = p(H|Z, Y - 1)$$

1.2. Naive Model 2.

One obvious way in which such a model might fail is that uses only the most recent year's data. Unless a sort of Markov assumption holds for our data, it would be better to use all of the historical data we have access to. Of course, we want to weight the more recent data more strongly. A naive model is to weight every year in our data set, D, by its distance from our target year, T:

$$p(H|Z, Y = T) = \frac{\sum_{Y \in D, Y < T} \frac{1}{T-Y} p(H|Z, Y)}{\sum_{Y \in D, Y < T} \frac{1}{T-Y}}$$

Date: January 16, 2010.

1.3. **Naive Model 3.** A still more sophisticated model is to assume that the homicide rate in each zipcode exhibits a linear trend over our data set: either increasing or decreasing monotonically over time. At base this amounts to the assumption that

$$p(H|Z, Y) = \alpha_Z + \beta_Z Y,$$

modulo a normalization to insure that the resulting values are actual probabilities when pooling across the zipcodes in our dataset:

$$p(H|Z, Y) = \frac{\alpha_Z + \beta_Z Y}{\sum_{Z \in D} \alpha_Z + \beta_Z Y}$$

These linear models are fit separately within each zipcode, which can lead to overfitting.

1.4. **Naive Model 4.** A better linear model is to fit using all of the training data, by including dummy variables for each zipcode:

$$p(H|Z, Y) = \alpha_Z + \beta Y.$$

In the process, we abandon varying slopes for the year regressor.

1.5. **Naive Model 5.** Still better is to fit using all of the training data, including dummy variables for each zipcode and interacting them with the year variable to get specific slopes for each zipcode:

$$p(H|Z, Y) = \alpha_Z + \beta_Z Y.$$

1.6. **Naive Model 6.** Building upon the previous simple linear models, we can further assume that the α_Z's and β_Z's are normally distributed and construct a hierarchical regression model of the sort described in Gelman and Hill's ARM textbook:

$$p(H|Z, Y) = \frac{\alpha_Z + \beta_Z Y}{\sum_{Z \in D} \alpha_Z + \beta_Z Y}$$

$$\alpha_Z \; N(\mu, \sigma)$$

$$\beta_Z \; N(\mu, \sigma)$$

This solves some of the overfitting of the earlier models.

1.7. **Naive Model 7.** Here is the simplest model that I can think of that uses something like k-nearest neighbors to generate predictions. We use Drew's graph distance metric as a distance between two zipcodes, z_1 and z_2, as a function $\delta(z_1, z_2)$. With this distance metric, we generate weights as follows:

$$w(z_1, z_2) = \frac{1}{1 + \delta(z_1, z_2)}$$

With these weights we generate predictions as,

$$p(H|Z_i, Y = T) = \frac{\sum_{j=1}^{47} w(z_i, z_j) p(H|Z_i, Y = T - 1)}{\sum_{j=1}^{47} w(z_i, z_j)}$$

modulo a possibly required normalization.

Once this works, we can improve it by using optimization techniques to select a function over δ that weights the distances more heavily:

$$w(z_1, z_2) = \frac{1}{1 + f(\delta(z_1, z_2))}$$

Taking $f(\delta) = \delta^a$ for an a fitted by least squares would be the easiest approach.

Afterwards, we can try pooling historical data in lieu of the previous year's data as the inputs for weighting.

1.8. **Ensemble Technique.**

Once you predictions from a series of models, you can pool them using ridge regression to get a better prediction.

1.9. **Real Models.**

Beyond this, it is clear that we can use external predictors variables for each zipcode, such as the average income in the zipcode, as well as year level predictors. Moreover, non-linear models such a k-nearest neighbors might easily be used. Finally, Drew has suggested using spatial regression, which, from what little I know, might be a substantial improvement over k-nearest neighbors while employing a similar pooling of data across neighboring zipcodes (modulo a metric space definition for our data set).