8/26/2015

Weston Buck

Data Science Final Project

**Problem**: Locating accurate crime data when determining living locations.

**Hypothesis**: I predict the KNN algorithm will accurately predict the crime most likely to occur at a specific latitude / longitude area.

The goal of this algorithm is to correctly predict the most likely crime to happen at an exact latitude / longitude. The data was downloaded from [www.sfdata.org](http://www.sfdata.org), with the full address referenced on the sources page. The data was ten years and 26 columns initially but was truncated down to five columns and one year to speed up the algorithm.

This pre-processing was key to test multiple feature iterations quickly and efficiently till the perfect balance was found. During the feature modification I learned that certain features initially though strong had no bearing on the overall accuracy. It turns out the only features needed, of the ~20 total initial options, were the latitude / longitude.

Data visualizations were weakly explored as the initial concept had to be changed two weeks prior to the due date. The website offered an online visualization tool that created a heat map of the SF crimes, included in the power point. The initial concept was tabled as I wasn’t able to scrape permit data in large enough quantities to warrant its inclusion. The permit data had addresses that needed lat/long conversions and the host sites allowed only 70 pulls per minute. This challenge solidified my decision to use KNN as the entire data set revolves around locational data that KNN could easily sort through.

The algorithm’s initial concept had real world applications for real estate price forecasting. This could have been achieved by plotting the crime data and finding trailing housing permit data overlaid with Zillow housing data. This initial concept assumes money being pumped into a set location will reduce the crime and increase the housing values at x months later. This wasn’t feasible with the permit data set missing the lat/long data.

In conclusion, the primary takeaway is that data cleaning takes far longer that initially though and excess features can create noise and reduce accuracy.

Sources:

<https://data.sfgov.org/Public-Safety/SFPD-Incidents-from-1-January-2003/tmnf-yvry>

Data Dictionary:

|  |  |  |
| --- | --- | --- |
| **Item** | **Type** | **Description / Definition** |
| IncidntNum | integer | Incident Identifier |
| Category | Object | Crime Type |
| Descript | Object | Crime Description |
| DayOfWeek | Object | Day of the Week |
| Data | Object | Date of the crime |
| Time | Object | Time of the crime |
| PdDistrict | Object | Region of the city |
| Resolution | Object | Arrested, Booked or None |
| Address | Object | Address of the crime |
| X | integer | Latitude |
| Y | integer | Longitude |
| Location | Object | Latitude and Longitude |
| PdId | integer | Police ID |