Project 3
Classification

Due: Monday, November 28

Analyzing Housing Trends on the Central Coast

Over the course of the past month and a half, one of the instructors have been collecting (as it becomes available) the data on the state of the real estate market on the Central Coast. This data is made available to you in a form of a CSV file.

The data was collected from two pages:

http://www.slocountyhomes.com/newlistex.php

and

http://www.slocountyhomes.com/pricedrops.php

The first page shows new real estate listings in the Central Coast area. The second page shows the changes in prices in the existing listings.

Please note, that the data collected from these two pages does not reflect the current state of the market, i.e., it does not show all the houses that are currently on the market in the Central Coast area. Therefore, it is an imperfect snapshot. However, as a sign of the current state of the market, this dataset may provide useful insight.

Analytical Questions to study. In class we demonstrated that certain subsets of the collected dataset can be subjected to classification in a straightforward way: for example, for the available market slice of San Luis Obispo and Paso Robles houses with the size from 2000 to 2500 square, we can predict the location of the house with high accuracy using linear classifiers (such as SVM). The key question to ponder is broad:

What can we learn from this dataset about the Central Coast real estate market using classification methods?

To successfully answer this question you need to do the following:

- Continue collecting data. Your final analysis should be based on at least two extra weeks of data.
- Choose the subsets of data you want to analyze. Despite relatively few attributes in the dataset, the dataset is complex - there are a lot of geographic areas covered, a lot of types of houses included, and so on. You should look at subsets of data where you may hope to achieve good predictions.
- Ask classification questions. For each subset of data, identify the class variable you want to predict.
- Analyze data. Use a variety of classification techniques to see how accurate your classifiers can get. You can use both individual classifiers as well as ensembles.
- Compare, contrast, conclude. For each subproblem you are studying, produce a comparative analysis of the different classifiers that you have tried, and determine which classifiers worked best.
- Report. Create a report documenting your findings.

Notes. When performing this set of tasks, please be aware of the followiong things:

- Data acquisition. You can acquire the extra data in any way you want. The conservative way is to simply continue getting the data from the two web pages indicated above and grow the dataset that way. You may also select other ways of procuring the data (as long as it is current!) from alternative web sites, or from alternative web pages for the SLO County Homes web site. In fact, if you are able to acquire more data from a different source, you do not even need to use the data from the dataset we are providing (assuming that your own data collection can collect the vast majority of the listings in the instructor's dataset).

  Additionally, while you should collect *at least* the features provided to you in the dataset (they are fairly self-explanatory), you may collect more features. This is also left up to you.
- Data Cleaning. The dataset given to you may need cleaning. First, some attribute may be inconsistent (see the listing date attribute for example). Second you may choose to remove certain listings from the dataset as out-of-area (or incomplete) noise.

- Subset selection and question asking. This is a chicken-and-egg situation. Try to see if you can ask meaningful analytical questions of your data (e.g., "how can we compare the Atascadero and Paso Robles markets?" or "Can we predict the number of bedrooms in the houses in the Five Cities area?"). Then figure out how to produce a slice of the dataset that would help you answer this question. Visualising your data slices is important because you may be able to see from the visualization whether or not your slice of data can be successfully partitioned by a classifier.

- Data Analysis. While we (the instructors) are interested in they questions you ask and the conclusions you reach, the real purpose of the lab is for you to compare the results of different classifiers on the same problem, and to try to make you form some understanding of why certain classifiers behave in certain ways. We want you to build some intuition concerning the classification methods, their strengths and weaknesses, and the means of overcoming the latter.

  We want you to try both individual standalone classifiers, as well as to examine one or more ensemble methods in order to determine the lift you get from them.

- Classifiers. You can use a combination of the classifiers implemented in the various data mining libraries (e.g, in `scikit-learn`), and the classifiers you implement yourself (like the SVM classifier that you are working on right now). While relying on existing implmentations may give you more robust performance, you can extract much more information out of the classifier you have built yourself.

- Observations and Conclusions. We want you to report on pretty much all the work you have completed and all analyses you have ran, but we also want you to eventually converge on classifiation problems where the results can be robust (i.e., have fairly high accuracy). This means that you should try working on individual questions for this dataset, until you hit on one or more questions where you can obtain sufficiently accurate classification results.

- Automation. One way to approach this part of the project (although not the only one) is to build a small analytical suit of methods that can be run on any subset of data. This will front-load the develpment work and leave the analytical work until the very end, but it may allow you to essentially pass a SQL query (or a Pandas selection condition) to your auomated suit and run all your analyses on the subset that gets selected.

Deliverables.

You have two types of deliverables: code and project reports.

Code deliverable is for archival purposes only. Submission instructions will be provided to you closer to the submission deadline.

Reports shall be text-processed and submitted in PDF format. The reports shall contain your team's narrative for each of the parts of the project. We want to know what you have tried, what worked and what did not. We also want to see, the most accurate predictions you were able to make highlighted in your report.

The reports can be written in an informal style (i.e., do not treat them as academic papers), but they should be as complete as possible. You should find ways to visualize important insight you gained and include these visualizations into the report.

Finally, each team will present their observations on November 30 (the official end of the project date is November 28, but we want to give you two days after coming back from the Thanksgiving break to complete the report and to work on the presentations).