Project 3: Unsupervised Learning

Creating Customer Segments

Project Description

You've been hired by a wholesale grocery distributor to help them determine which changes will benefit their business. They recently tested out a change to their delivery method, from a regular morning delivery to a cheaper, bulk evening delivery. Initial tests didn't discover any significant effect, so they implemented the cheaper option. Almost immediately, they began getting complaints about the change and losing customers. As it turns out, the highest volume customers had an easy time adapting to the change, whereas smaller family run shops had serious issues with it--but these issues were washed out statistically by noise from the larger customers.

For the future, they want to have a sense of what sorts of different customers they have. Then, when implementing changes, they can look at the effects on these different groups independently. Your task is to use unsupervised techniques to see what sort of patterns exist among existing customers, and what exactly makes them different.

Language and libraries

For this project, you will need to have the following software installed:

- Python 2.7
- NumPy
- pandas
- matplotlib
- scikit-learn
- iPython Notebook

Template code

Download customer_segments.zip, unzip it and refer to the `README.md` file for further information on the dataset and instructions on how to open the iPython notebook (`customer_segments.ipynb`). Follow the instructions in the notebook to complete each step, and answer the questions asked.

Deliverables

- Fully implemented notebook (`customer_segments.ipynb`) with all the code blocks executed and showing output.
- Report in PDF format (you can simply save the notebook as PDF with all the answers typed in).

Questions and Report Structure

Component analysis

1. Reflection on PCA/ICA
    - What are likely candidates for early PCA dimensions?
    - What might ICA dimensions look like?
2. What proportion of variance is explained by each PCA dimension?
3. PCA dimensions
    - What are the first few components? What might they represent?
    - How can you use this information?
4. ICA
    - What are the components that arise?
    - How could you use these components?

Clustering

5. Decide on K means clustering or Gaussian mixture methods
    - What are the advantages and disadvantages of each?
    - How will you decide on the number of clusters?
6. Implement clusters
    - Sample central points of the clusters
7. Produce a graphic
    - Visualize important dimensions by reducing with PCA
    - Are there clusters that aren't very well distinguished? How could you improve the visualization?

Conclusions

8. Which of these techniques felt like it fit naturally with the data?
9. How would you use that technique to assist if the company conducted an experiment?
10. How would you use that data to predict future customer needs?

Evaluation

Your project will be reviewed by a Udacity reviewer against **this rubric**. Be sure to review it thoroughly before you submit. All criteria must "meet specifications" in order to pass.

Submission

When you're ready to submit your project go back to your Udacity Home, click on Project 3, and we'll walk you through the rest of the submission process.

If you are having any problems submitting your project or wish to check on the status of your submission, please email us at **machine-support@udacity.com** or visit us in the discussion forums.

What's Next?

You will get an email as soon as your reviewer has feedback for you. In the meantime, review your next project and feel free to get started on it or the courses supporting it!