Machine Learning Engineer Nanodegree

Capstone Proposal

Levi King
January 13th, 2017

Proposal

Domain Background

The earliest soap-like substance is at least 4500 years old; people have been making soap for a very long time. Not only that, they've been writing down how to make soap for nearly as long, with soap recipes being found from many ancient civilization. Soap is just alkali salt and fat, but why would you need a recipe?

In the most basic sense, you need a recipe for soap because it uses caustic ingredients that can irritate or burn your skin if not made correctly, and not just any oil can be used as some oils will not saponify (the word for the chemical process of becoming soap). Traditionally, the oils used for soap came from local plants or animals that the people used in their daily lives either as byproducts for cooking or processing plants for other uses.

As time went on and industrialization and transportation improved, people stopped making soap and started buying it and global supply chains lead to a race to the bottom as well as niche markets as with most goods people

use frequently.So our soap evolved from whatever was around to whatever was cheapest. Like most other products however, as the internet allows more people exposure to markets and distribution channels and people get wealthier they become more discriminating and markets have become much more clustered based on preferences. A lot of soap these days isn't even technically soap (though it does clean).

I personally hate feeling sticky after using hotel soap, and mostly I'm just very picky about things like soap. But, what always bothered me was I don't know why I like the soap I like. I can't articulate what it is that bothers me or know what kind of words to use to describe it. It's a lot like wine - you need to know a certain amount about wine before you can even describe what you like about it and buy. So I began to learn a lot more about soap and why I like and dislike certain soaps. Organic chemistry is a complicated field that governs soap making and soaps are never pure substances. They always have a lot of ingredients in the tail of their molecular composition. In addition, people have a lot of different oils on their skin, which is what the soap is interacting with. To make matters even more complicated that skin oil - soap interaction is taking place in water that itself has yet more chemicals related to what's in the water when the bathing is taking place.

My problem is that I take a shower when travelling, in a hotel, at a friend/family members, or at an airbnb and I do not feel clean after bathing or I feel sticky. I never want to have to experience that, but the complex organic chemistry makes that difficult to quantify. I don't want to be someone who asks for a water sample before visiting a friend. It doesn't make sense to buy a spectrometer to observe the differences in my soap and observe the minute differences in it's reaction to various skin oils. It's just too big and complicated of a problem to solve analytically.

Problem Statement

I want to take a person, a water, and a soap and classify it as either satisfactory or unsatisfactory. People's tastes certainly change, but I have talked to many people and I certainly share with at least some number of people the desire to have a good soap. Can I define a good soap? On a few dimensions I can certainly say "this is a good lather" or "this is sticky" but it's difficult to put an objective value to it. This has a lot of analogs to music, so basically I'm making the Pandora of soap.

Datasets and Inputs

Unfortunately, soap is mostly unexplored by machine learning (or it's been supressed by results of machine learning being applied to web services). If there exists what I am looking for, it is dilluted. So, I'm going to test soap on myself and friends and family for a small dataset - get some range of values and just randomly generate more values based on that distribution to do a proof of concept.

Since I'll be bootstrapping a recommendation system and do not have thousands of friends, an adaptive content based filter will probably be my best bet (since soaps have objective attributes). The inputs will be personal information and a seed soap, and the output will be a soap recommendation. To continue the Pandora analogy I will have to create a basic Soap Genome Project where we identify a vector that is some number of "genes" of soap, the MVP will likely be just using the fatty acid composition of the base ingredients. The soap wikipedia page identifies seven fatty acids and the amount of those fatty acids in traditional oils used in soap making.

Solution Statement

In this section, clearly describe a solution to the problem. The solution

should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

The end solution to me is a black box soap recommendation engine, where you fill out a survey and it spits out a soap formula. It's really a recommender system, which can either use collaborative or content based filtering. Unfortunately, such a database does not yet exist and soap companies aren't exactly forthcoming about their specific recipes, and I don't currently have a long running wildly popular database of people's soap use patterns. This means I'm mostly without a good reference point. The content based filtering is much easier to bootstrap though so that's what I'll start with as described in the dataset and input section.

Benchmark Model

Since we are building a model from scratch, existing datasets do not exist. This means we really only need to beat a random selection to be performant. The other benchmark model could be trivial selection, or just choosing the same recommendation every time. Seeing significant improvement over both of these is probably a good indication we are on the right track.

Evaluation Metrics

Ultimately, we are trying to predict which soap a person would like as it compares to random selection. For this reason we can just use the r^2 of the model against the r^2 of choosing randomly.

Project Design

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

First, I will make a few single oil soaps such as 100% olive oil, 100% coconut oil, and 100% peanut oil for example and note the fatty acid composition of those oils. Then I will find willing test subjects to fill out a questionaire and use the soaps rating whether or not they like the soaps (binary yes/no), to rank/rate the soaps (regression), and to do the same with attributes of the soap (stickiness, lather, hardness, cleanliness, drying-ness, etc). When combined with the "genes" of the soap and the soap that the testers buy I should be able to estimate what combination of soap they would like and make that.

Next, I will look for trends with the results. And try to determine what's important to each tester. This analysis should yield a personal preference matrix for each tester. And then I can make a second batch of soaps that are a linear combination of oils for each testers preferences (50% olive oil, 20% coconut oil, 30% peanut oil for example). Then, I can reevaluate to see if that's consistent with our expectations - ideally the combinations should do better than the pure oil soaps.

From there, I will have a good base, some primary "genes" of the soap and some fairly naiive preferences. I can then build a system that takes a seed

soap, recommends a soap, gathers ratings from customers and iterates adding complexity and improving accuracy as time goes on (likely with help from me).

Once the dataset is significantly large we can look at using techniques such as Neural Networks or Support Vector Machines to form more general cases for recommending to new people. It may also be possible to use something basic like a decision tree if it turns out the attributes leading to certain preferences are in fact very well defined. It is difficult to know exactly what to do until we have a better idea of how the exploration turns out.