Latent Structures in US Startup Funding

Kelvin Yu
Princeton University
zkyu@princeton.edu

Abstract

Institutional investing in early-stage startups, or otherwise known as venture capital (VC), has been a crucial asset class in American society for the past three decades. In 2018, US VC firms invested over $84 billion in over 8,000 companies, an all-time investment high since the dot-com bubble [1]. Venture capital is a game of winners; from 1981-2014, the top-performing quartile of funds have an average IRR of 25%, while the median fund returns 11.9% [2]. In contrast, the SP 500 has a 12.7% average IRR over that same time span, which means over half of VC firms return less than a public index. This asymmetry in returns leads us to seek ways to gain a competitive advantage over traditional venture capitalists.

1 Introduction

While the industry has fueled much technological innovation, VC itself remains a cottage industry largely untouched by data; many VCs still make investments based off "gut feel," but there is room to use data to aid in investment decisions. In this assignment we are interested in exploring which kinds of companies receive funding to better understand shifting technology trends and patterns of capital allocation within the VC industry. This is important because return on investment (ROI) is partially determined by the total amount of capital flowing into a sector. For example, if a market is oversaturated with capital, even if it is growing quickly, it may not yield as an attractive of a return as an undersaturated market with less growth. In fact, too much capital is a sign of a bubble, which the VC industry famously suffered with the tech bubble in 2001. If too many startups are being funded, like cleantech companies from 2005-2009, it is as much a warning sign as a sign of growth.

In addition, latent structures in funding dynamics can shed light on interesting ways we can segment the data to look for patterns in a given year, category, region, or more. In this paper we analyze the patterns extracted by two different unsupervised learning models, one with and one and without feature extraction. We note interesting patterns as well as explore extensions we could add to the original task and the differences the extensions could make.

1.1 Related Works

A number of similar studies have emerged over the past few years. Using supervised learning methods such as Support Vectors Machines, Bhm and Weking 2017 analyzed 181 startups from the USA and Germany to classify different business models in regards to their performance. They found 12 distinct business model clusters with different growth expectations and chances of survival, and claims to be able to predict a startup's survival with an 83.6% accuracy. In addition, they were able to assign weights to variables such as information about the founders and their prior knowledge of the industry, their degree of innovativeness, competition, founding team size, and patents, which all improved the model accuracy [3].

Belenzon, Chatterji, and Daley 2017, in "Eponymous Entrepreneurs" demonstrate that eponymy-firms being named after their ownersis linked to superior firm performance. They propose an explanation based on eponymy creating an association between the entrepreneur and her firm that

increases the reputational benefits/costs of successful/unsuccessful outcomes. Using a data set consisting of over 1.8 million firms, they found support for their hypothesis on a corresponding signaling model, which further predicts that these effects will be stronger for entrepreneurs with rarer names [4].

Last year Andrew Ng, Princeton Class of 2018, published his senior thesis: Data-Driven Investment: Formalizing the Early-Stage Venture Capital Process using Machine Learning, in which he utilized PCA dimensionality reduction, regression models, and cross validation to evaluate startup performance [5]. His data set was extracted from Crunchbase. However, like Bhm and Belenzon, Ng's primary goal was to predict startup success using regression, whereas our interests lie in the latent dynamics around venture capital financings.

While not utilized by most firms, some venture capitalists are beginning to realize the power of machine learning and incorporating them into their operations. Notable ones include Social Capital, 645 Ventures, SignalFire, Rocketship VC, and Correlation Ventures. However, even among them there are significant differences. 645 Ventures and SignalFire utilize most of their data science efforts into sourcing companies (i.e. identifying companies for potential investment), while Correlation and Rocketship rely on their models much more heavily to evaluate companies for investment. A common criticism against using data to evaluate companies is that early-stage VC success depends a lot on evaluating the founder(s), which only a human can do. Lastly, Social Capital uses machine learning primarily after investing to perform portfolio support, meaning they take their portfolio companies' data to help them gain insight on customer engagement, unit economics, and more.

2 Methods

2.1 Data Cleaning

This data set is a public spreadsheet of every single publicly announced VC investment in U.S. media from June 10th, 2016 to the present [6]. The data set consists of approximately 14,000 samples (investments), which are made globally and across all sectors. 16 variables are available, including headquarter location, industry category, key words, top investors in the round, round stage (seed, Series A, etc.), and more. The data is in an investment x investment features matrix. We drop features that do not have any correlation with other features, namely the "Link," "Website," and founder information features, which only contain links. We also drop any samples with unknown investors and rows containing "Week to Date" and "Month to Date" values, which are extraneous to our study.

Next, we transform the "Date" variable into two features, Month and Year, because venture funding can be cyclical so a given time period might correlate with higher fundraising numbers. We also split the "Location" variable into two separate features, City and Region, because VC is more concentrated in certain geographies like Silicon Valley.

In any given fundraising round there is a "lead" investor, which is the firm that contributes the most capital. We assume that the "lead" investor in the round is the first investor listed in the "TOP INVESTORS" feature since the most important investor is usually put first. To reduce the number of features, we drop all other investors and only add the first investor as a feature to our data set.

We are now ready to clean the data for the Category feature, which are the markets that the sample belongs to. Many categories are blend words, e.g. agtech (agricultural tech), fintech (financial tech), etc. Many of these have typos, inconsistent hyphenating (fintech vs. fin-tech), and inconsistent capitalizations, which we all address. In addition, we combine certain categories that we deem too similar to separate (e.g. we classify both 'machine learning' and 'artificial intelligence' as 'AI', and combine 'health' and 'biotech' categories as 'health.') Our data set is now 11,686 samples x 10 features, and we are ready to impute missing data.

2.2 Imputation

Many funding amounts are missing, so we impute them based on estimated funding rounds. In startup funding, the earliest rounds of funding are traditionally "Angel" rounds, where "angels" refer to high-net worth individuals that invest their own wealth into startups. These are followed by

the first form of institutional capital, deemed the "Seed" round, which is followed by "Series A," "Series B," etc. We calculate the mean amounts for each type of funding round, and then impute missing funding amounts based on the sample's funding round. Surprisingly, we found that the mean invested amount for angel rounds was actually greater than Seed or Series A rounds, which were $15.9 million, $3.4 million, and $16.0 million, respectively. Further analysis showed that twelve out of nineteen samples indicated as angel investments were from China. This makes sense because angel rounds are typically not publicly announced in the US unless it is an unusually large round, and assuming the same holds for China, the data for angel rounds will naturally be skewed towards the upper extreme.

Next, we impute the missing values for the "Region" feature by filling them with the sample's city. Lastly, we perform one-hot encoding on our categorical features, 'Category', 'Round', 'City', 'Region', 'Month', 'Year', and 'First Investor' to expand our feature space from 10 to 10250.

Lastly, we shuffle our sample ordering to obtain a random shuffling, then hold out a portion of the data via a 80/20 train/test split in order to evaluate our held-out latent structure for the general data set at the end.

2.3 Segmentation

For all data sets, we fit them with unsupervised learning models Latent Dirichlet Allocation and Gaussian Mixture Model (GMM).

To uncover interesting latent structures, we decided to control for some variables that we thought might influence macro-VC funding trends. Specifically, we chose to run our methods on the entire data set (the "General Case"), and also sub-data sets separated by year and region. After running our initial file cleaning and imputation scripts, we created sub-data sets based on these segmentations. For example, we ran LDA separately on year0.csv, year1.csv, year2.csv, and year3.csv, which are the samples from 2016, 2017, 2018, and 2019, respectively. For region, we created a separate data set containing only Chinese investments to compare Chinese latent structures with those of the US, since China is the only country in the world whose VC industry rivals that of the US's. For the first time in 2016, venture fundraising in China matched U.S. fundraising levels ($50 billion) [7].

2.4 Feature Selection

After our initial data processing, we drop the Notes variable, as it is not categorical and would not add much to our analysis. For GMM, we apply Principal Component Analysis as feature selection to reduce dimensionality and computation time. This performs an orthogonal transformation on our data onto four principal axes so that we can fit our models on data of reduced dimensions, thus vastly reducing computation time.

2.5 Latent Structure Analysis

We perform two forms of latent structure analysis: Gaussian Mixture Model (GMM) with PCA dimensionality reduction and Latent Dirichlet Allocation (LDA). To implement these methods, we use Sci-Kit Learn's Python libraries. All three procedural components are described below.

1. *Latent Dirichlet Allocation* (LDA): A generative probabilistic model for collections of discrete data that can discover abstract topics from a collection of documents. We chose to fit our LDA model to five topics for all the data sets by comparing log-likelihood scores across varying topic numbers from one to ten and seeing that five was the optimal number.
2. *Gaussian Mixture Model* (GMM): A probablistic model that uses the expectation-maximization (EM) algorithm and assumes the data is generated from a mixture of a finite number of Gaussian distributions. Depending on which data set we were fitting, we used different numbers of mixture components based on their log-likelihood scores.
3. *Principal Component Analysis* (PCA): Decomposes a multivariate dataset in a set of linearly uncorrelated variables (principal components) that explain a maximum amount of the variance. We chose to decompose our data into 4 principal components, as that captured over ninety percent of the variance.

3 Spotlight Method: Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative probabilistic model that fits a probability distribution to a group of documents and topics. Each topic, or category, describes a latent structure hidden in the data, and the data is made up of the set of documents, or samples. Each document can be thought of as a mixture of topics, hence it is also known as an admixutre mixture or mixed membership model. The key assumption made by LDA is that the topic distribution is assumed to have a spare Dirichlet prior, which implies the documents only cover a small set of topics and that topics use only a small set of words frequently [8].

The probablistic model estimated by LDA consists of two matrices. The first describes the word distribution per topic, where the columns are all the words in the samples and each row describes the probability of selecting each word in a given topic. The second matrix describes the probability of a given topic existing in a document, e.g. a document may contain 0.76 of topic one, 0.22 of topic two, and 0.02 of topic three.

The model is described as follows [8]:

$$\pi_{\mathbf{i}}|\alpha \sim \text{Dir}(\alpha \mathbf{1_K}) \text{ Draw document-specific distribution } \pi_i$$

$$q_{il}|\pi_{\mathbf{i}} \sim \text{Cat}(\pi_{\mathbf{i}}) \text{ Assign every word its own topic, drawn from the document-specific distribution}$$

$$b_k|\gamma \sim \text{Dir}(\gamma \mathbf{1_V}) \text{ Find the distribution of topics over each word } \mathrm{b_k}$$

$$y_{il}|q_{il} = k, \mathbf{B} \sim \text{Cat}(\mathrm{b}_k) \text{ Draw a sample from a Dirichlet distribution to find the probability of each topic per document.}$$

To simplify the model into one equation, We can marginalize out the q_i variables to create the following:

$$p(y_{il} = v|\pi_i) = \sum_k p(y_{il} = v|q_{il} = k)p(q_{il} = k) = \sum_k \pi_{ik} b_{kv}$$

where $p(y_{il})$ is a document, each vector b_k defines a distribution over V words, each k is a topic, and each document vector pi_i defines a distribution over K topics [9]. Thus, we can model each document as an admixture over topics. Visually, this can be represented as

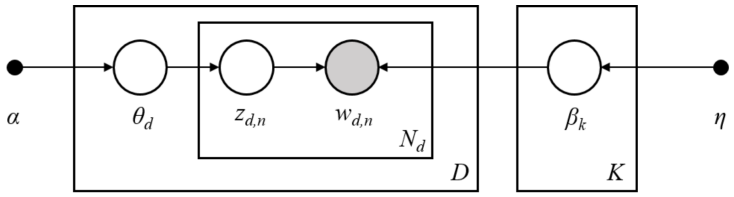

Figure 1: LDA model for D documents, N words, and K topics.

One advantage of using LDA for natural language processing is that the prior distributions on multinomials are on a k-1 simplex, which can handle ambiguity as opposed to Euclidean space [8]. A common issue in NLP is polysemy, in which words have multiple means; for example, "bounce" can be used in the context of "bouncing a ball," but can also be used as slang for leaving e.g. "let's bounce!" In LDA, we can choose how many topics we want, and so we can have multiple topics that generate the word "bounce," thus reflecting this ambiguity. Looking at a word in isolation might be confusing, but if we analyzed the other words in a latent structure or document then it clears up the ambiguity.

Since it was first published in 2003, numerous extension have been proposed to address its shortcomings, such as its inability to capture correlation between topics and its assumption that topics are static. We will discuss them below.

It is reasonable to assume that any document with the "health" topic might also have the "biotech" topic, but LDA does not take advantage of this correlation. This is because the use of a Dirichlet prior for π_i, which is only characterized by a mean vector and a strength parameter with a fixed

covariance. To solve this, Blei and Lafferty 2007 proposed the correlated topic model (CTM), whereby they replaced the Dirichlet prior with a logistical normal distribution. This is similar to categorical Principal Component Analysis, except the CTM covariance is represented by stochastic matrix whereas in PCA we have an unconstrained matrix.

It is also reasonable to assume that some topics evolve over time. For example, venture capital is a field that is heavily influenced by Fearing Of Missing Out (FOMO), and investors are always trying to brand themselves as unique, which works for a time until everyone else starts copying them. For example, the term "value-added investor" was pioneered by Andreessen Horowitz in 2009 to describe how they brought more than just capital to their investments by helping portfolio companies hire talent, get media attention, etc. Today, every VC describes themselves as "value-add" even though many of them aren't. To model topics dynamically in LDA we can use a dynamic logistic normal model, devised by Blei and Lafferty 2006b. This assumes the topic distributions evolve according to a Gaussian random walk, then map these distribution vectors to probabilities via the softmax function:

$$
\begin{aligned}
b_{t,k}|b_{t-1,k} &\sim \mathcal{N}(b_{t-1,k}, \sigma^2 1_V) \\
{\pi_i}^t &\sim \text{Dir}(\alpha 1_K) \\
{q_{il}}^t|{\pi_i}^t &\sim \text{Cat}({\pi_t}^t) \\
{y_{il}}^t|q_{il}^t = k, B^t &\sim \text{Cat}(\mathcal{S}(b_k^t))
\end{aligned}
$$

We use the *LatentDirichletAllocation* package from SciKitLearn Python to implement the Latent Dirichlet Allocation model on the data set.

4 Results

4.1 General Case

4.1.1 LDA

In the general case (no segmentation of data based on year or region), we found that LDA with five topics (i.e. latent users) gave us the most robust results. In particular, topics 3 and 5 had extremely high user proportions, meaning samples that contained the 3rd and 5th latent structures contained a lot of their features.

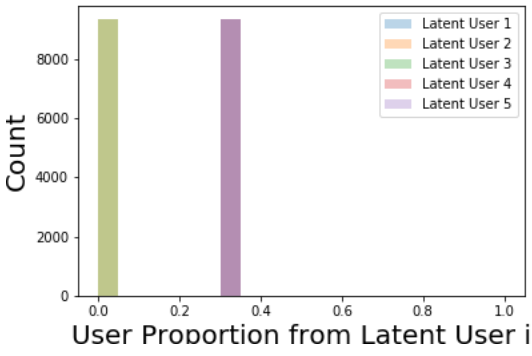

Figure 2: User proportions for different latent users.

Latent user 5 contained health, consumer internet, and enterprise category startups based in San Francisco that raised Series A, B, and/or C rounds. Latent user 4 was very similar to user 5, also containing the same three categories based in San Francisco with Series A, and Series B funding, although it also included Angel funding. Latent user 1 was the exact same as user 5, except instead of San Francisco the location was New York City. These results suggest enterprise and biotech/health and B2B software companies coming out of Silicon Valley and NYC, the two biggest tech hubs in

the US, are likely to raise multiple rounds of funding, which is reasonable given how cash-intensive they typically are.

4.1.2 GMM

With GMM/PCA, we found that the principal component reduction captured the most variance between the features effectively, with GMM clusters being formed around years and location. In the general case, we saw in every year (2016-2019), California was the dominant region feature, with San Francisco coming in as the leading city feature. Health and enterprise were consistently the top categories, with Series A and Series B funding rounds being the most common. We also saw a cluster for New York, the second leading region, with enterprise and health also dominating.

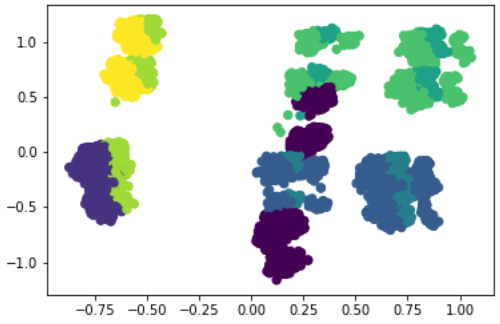

Figure 3: Clusters for the general case with PCA dimensionality reduction and GMM clustering.

4.2 Segmented Data

4.2.1 LDA

Even with segmentation by year, we see that our topics shared many similarities. The top latent structure in each year generally shared the same categories, type of fundraising rounds, and locations. In particular, California and San Francisco were in the top three features for each topic in every year except 2018, confirming that Silicon Valley is still the epicenter of startup life in the world. No meaningful topics were found for the held-out data set of Chinese investments. This is because it contained a relatively small feature set (405) compared to the others. Below are the user proportion distributions per topic, or latent startups.

Year0	Pseudocounts	Year1	Pseudocounts	Year2	Pseudocounts	Year3	Pseudocounts
California	431	California	778	Health	387	California	385
San Francisco	179	Undisclosed round	436	Round B	373	April	190
Seed	155	San Francisco	339	California	356	San Francisco	170
Round A	123	Health	230	Enterprise	312	Round A	142
June	112	Round A	204	Angel	297	March	137
Enterprise	112	March	184	Massachusetts	255	Feburary	115
Consumer Internet	89	Seed	181	UK	242	Round B	103
December	77	Enterprise	167	October	182	Janurary	87
September	73	Janurary	157	Round C	169	Angel	78
Undisclosed round	65	Consumer Internet	153	London	163	UK	75

Figure 4: Top ten features for the topic with the highest proportions in the users. Pseudocounts are the number of times a feature appears in topic 1 (i.e. latent startup 1).

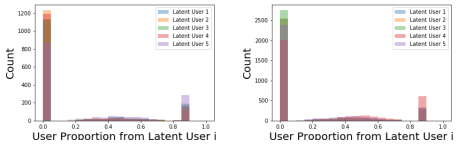

Figure 5: 2016 and 2017, respectively

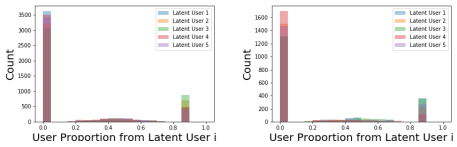

Figure 6: 2018 and 2019, respectively

4.2.2 GMM

After running GMM on the segmented yearly data, some previously undiscovered patterns emerge. In 2016, startups that raised undisclosed rounds in San Francisco or New York were primarily clustered with the enterprise category, suggesting B2B startups are more ”secretive” than consumer-facing businesses. Like in the general case and LDA, we still see that health and enterprise startups are the most common, although the 2017 cluster for NYC companies also contained consumer internet and fintech companies.

Cluster 1	Cluster 1 Proportion	Cluster 2	Cluster 2 Proportion	Cluster 3	Cluster 3 Proportion	Cluster 4	Cluster 4 Proportion
California	0.91	Round A	0.43	California	0.83	Undisclosed Round	0.89
San Francisco	0.46	New York	0.25	Undisclosed Round	0.83	Enterprise	0.29
Round A	0.34	NYC	0.26	San Francisco	0.38	New York	0.23
Enterprise	0.25	Health	0.23	Enterprise	0.30	NYC	0.22
Round B	0.22	Enterprise	0.23	October	0.22	October	0.20
Health	0.21	Round B	0.20	December	0.21	December	0.19
Seed	0.20	June	0.19	November	0.21	November	0.19
July	0.17	September	0.16	Seed	0.15	September	0.17
June	0.17	July	0.16	September	0.14	August	0.17
October	0.16	August	0.15	August	0.13	Health	0.14

Cluster 1	Cluster 1 Proportion	Cluster 2	Cluster 2 Proportion
California	0.93	Health	0.23
San Francisco	0.43	Round B	0.20
Round A	0.31	Round A	0.19
Enterprise	0.22	Enterprise	0.18
Health	0.18	New York	0.18
Round B	0.17	Round Undisclosed	0.18
Undisclosed Round	0.15	New York City	0.17
May	0.12	Consumer Internet	0.17
June	0.12	January	0.12
Seed	0.11	Fintech	0.11

Figure 7: 2016 and 2017 clusters, respectively

We were surprised to see that a cluster for China existed in 2018 because LDA did not reveal any meaningful latent structures even for the held-out China data. As you can see in column one and two, ”China” appeared in a cluster that had a very high proportion of startups that that raised Series A and B funding, suggesting that only large Chinese investments are covered by the US media. It also appeared to be a good year for consumer internet startups, as they appeared in all three clusters.

Cluster 1	Cluster 1 Proportion	Cluster 2	Cluster 2 Proportion	Cluster 3	Cluster 3 Proportion
Round B	0.94	Round A	0.95	California	0.94
Enterprise	0.24	Health	0.25	San Francisco	0.49
Health	0.24	Enterprise	0.24	Seed	0.34
China	0.13	New York	0.15	Round C	0.23
June	0.11	NYC	0.15	Angel	0.16
January	0.10	April	0.13	October	0.15
May	0.10	China	0.10	Health	0.15
April	0.10	November	0.10	Enterprise	0.13
July	0.10	May	0.10	November	0.11
Consumer Internet	0.10	Consumer Internet	0.10	Consumer Internet	0.11

Cluster 1	Cluster 1 Proportion	Cluster 2	Cluster 2 Proportion	Cluster 3	Cluster 3 Proportion	Cluster 4	Cluster 4 Proportion
January	0.99	March	0.96	April	0.87	February	0.96
Round A	0.26	Enterprise	0.34	Health	0.36	California	0.32
Seed Round	0.23	Seed	0.27	Round A	0.30	Seed	0.30
China	0.17	California	0.25	Round B	0.28	Round A	0.17
Health	0.17	Angel	0.17	Angel	0.19	Angel	0.16
Round B	0.15	Round A	0.15	Fintech	0.11	San Francisco	0.15
Enterprise	0.15	Round B	0.15	Consumer Internet	0.09	Enterprise	0.15
California	0.12	San Francisco	0.14	Massachusetts	0.09	Round B	0.14
Angel	0.12	Round C	0.11	NYC	0.09	Health	0.11
Fintech	0.11	Fintech	0.10	NY	0.09	Fintech	0.10

Figure 8: 2018 and 2019 clusters, respectively

The 2019 cluster, only consisting of startups funded thus far in the year, was clustered in different months, which allowed to see which startups raised capital in each month so far this year.

5 Discussion and Conclusion

For GMM, the average log-likelihood per data point in our training set versus test set was $7.64e-06$ and $-1.16e-03$, respectively. For LDA, the comparison was $-43.02, -54.33$. The relative variance in GMM log-likelihood ratios can be attributed to the fact that we fit the training and test sets to 8 and 7 clusters, respectively. Over-fitting with too many clusters causes the Expectation-maximization algorithm to converge to a local optimum that was not be truly representative of the data set.

As previously discussed, we found latent structures using our various models that illuminated capital allocation patterns, especially in regards to geography and category. Specifically, we saw that enterprise and health startups dominated investment rounds as a share of total fundraising events in both New York and California, which were two most represented regions in funding. These trends have remained constant over time, and while capital is flowing abundantly into these sectors, new investors should look to invest elsewhere where they can more easily build a competitive advantage. Over time, we saw that seed financing rounds decreased after 2016, but is starting to pick up again in 2019. It appears as a feature with a small proportion in 2017-2018 in only one cluster per year, but in 2016 and 2019 it appeared in multiple clusters with a high proportion in each. In 2018, the third most popular investment category behind enterprise and health was consumer internet, which appeared as tenth feature in all three 2018 clusters, but in 2019 this trend shifted overwhelmingly to fintech, with this category taking tenth place in every cluster in 2019. The fact that fintech startups are raising a significant amount of capital in 2019 compared to its historical levels suggests a new development in the market. Lastly, both our LDA and GMM models showed that the number of deals with undiclosed fundraising amounts drastically decreased over time. It was a prominent feature in our 2016 and 2017 models, but completely disappeared in 2018 and 2019. This suggests startups have grown more comfortable with telling the media about how much they raised.

We saw unexpected results as well. Due to the stereotype of Boston as center of biotech/health, we expected most health startups to come out of the MA region, but every latent structure containing the 'health' category had higher proportions in NYC or SAn Francisco. We also did not expect to see such a robust California startup-ecosystem outside SF, as cluster 1 in 2017, cluster 3 in 2018, clusters 1 and 3 in 2016, and every LDA topic all have 'California' in $85\%+$ of their samples, but only less than half containing SF. This is explained by a trend of startups moving away due to exorbitant real estate prices, but we did not expect the ratio to be so low given the historical dominance of Silicon Valley.

6 Extensions

An extension that we are currently in the process of working on is a predictive model using a probabilistic classifier like the Naive Bayes classifier to predict a given trait of a startup. For our purposes of augmenting the VC business, we are interested in predicting where a startup is likely to originate from given they've raised a certain type of fundraising, because that would give us insight as to which locations lack capital for certain investment rounds. For example, Norway may have an abundance of Series A/B funds but have a shortage of seed funds. We do this by computing the posterior distribution

$$\max_A f_{A|B}(A|B) = \frac{\max_A f_{B|A}(B|A)\pi_A}{f_B(B)}$$

for every A and B, where A is a city, B is the type of fundraising round, $\pi_A = \frac{\text{Number of startups in location A}}{\text{Total number of startups}}$, and f is a probability distribution function.

References

[1] Cook, John. Venture Capitalists Poured $84 Billion into Startups in 2017, a Massive Tally Not Seen since the Dot-Com Boom. GeekWire, 16 May 2018, www.geekwire.com/2018/venture-capitalists-poured-84-billion-startups-2017-massive-tally-not-seen-since-dot-com-boom-era/.

[2] VC 101: The Angel Investor's Guide to Startup Investing. FundersClub, fundersclub.com/learn/guides/vc-101/understanding-venture-capital/.

[3] Bhm, Markus Weking, Jrg Fortunat, Frank Mueller, Simon Welpe, Isabell Krcmar, Helmut. (2017). The Business Model DNA: Towards an Approach for Predicting Business Model Success.

[4] Belenzon, Sharon, Aaron K. Chatterji, and Brendan Daley. 2017. "Eponymous Entrepreneurs." American Economic Review, 107 (6): 1638-55.

[5] Ng, Andrew. Data-Driven Investment: Formalizing the Early-Stage Venture Capital Process Using Machine Learning. Princeton Senior Thesis, Apr. 2018.

[6] Funding Roundup. docs.google.com/spreadsheets/d/1kwbnPdEtHdhNjVdNRKWZIZnEAgXhyHZ462ASUodpc/editgid=0. Google Sheets, 2019.

[7] Hardin, Tim, and Silicon Valley Bank. China Now Rivals U.S. in VC Investments. VentureBeat, VentureBeat, 14 Oct. 2017, venturebeat.com/2017/10/14/china-now-rivals-u-s-in-vc-investments/.

[8] Kevin P. Murphy. Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series). The MIT Press, 2012.

[9] Sci-Kit Learn API Reference. scikit-learn v0.20.3, scikit-learn, 2019, https://scikit-learn.org/stable/modules/classes.html.

[10] Lu, Jonathan. UpdatedAnalysis.ipynb. 2019, UpdatedAnalysis.ipynb