Math 156, Spring 2018
Harsh Gupta, David Nguyen, and Aviva Prins

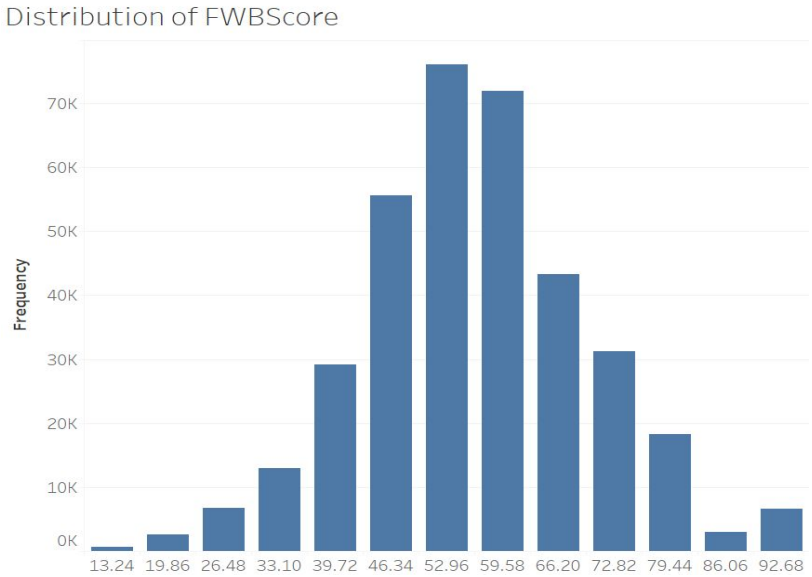
"How fat is your piggy bank?" - An exploration into the predictability of financial well-being using Support Vector Machines (SVMs).

Introduction

One of the most commonly used tools in public policy and social sciences are surveys. In the absence of directly observable actions, self-reported surveys become indispensable in identifying and classifying high-risk individuals. One such survey, which identifies individuals under financial duress, is the subject of our project.

The Consumer Financial Protection Bureau (CFPB)[2] conducts a National Financial Well-Being survey in order to determine the financial well-being. Financial Well-Being (FWB)[2] is defined as "a state of being wherein a person can fully meet current and ongoing financial obligations, can feel secure in their financial future, and is able to make choices that allow them to enjoy life" . In order to determine this scale, CFPB asks questions directly related to their financial well-being, and then uses an IRT scale to determine a financial well-being score (FWB Score)[Appendix 1]. Using this FWB Score we can determine whether an individual is at financial risk.

Figure 1: Distribution of Financial Well-Being Score



In addition to the questions needed to calculate the FWB Score, the survey asks a bank of questions regarding the Financial Knowledge, Education Level, Income and Employment, Family History, Financial Habits, Demographic Information and so on. A natural question that arises is whether we can use this indirect information about the individual to determine his financial well-being. That is, can we identify individuals in financial danger without directly asking these specific questions? This question, which focuses largely of the applicability of SVMs in answering real world problems, is the first part of our project. We select 150 questions

from the survey, obviously excluding the questions used to determine the FWB Score. We use 150 questions as our training and target variables to classify each individual.

The second aspect of our project focuses largely on the technique itself. Specifically, we look at how the performance of SVMs (measured in terms of accuracy) varies as we change the number of classes from two to three. It comes as no surprise that SVMs perform much better when classifying into two classes, rather than three classes. We provide some possible explanations for this in the context of our particular problem, and suggest some ways in which we can overcome this shortcoming. We also examine whether the questions we select are particularly important, or whether SVM does well irrespective of the nature of the features, as long as we choose a sufficient quantity of features.

Model

We choose Support Vector Machines (SVMs) rather than any of the other methods discussed in the course because SVM is a flexible maximum margin classifier. In general, we know that SVM performs better than the other supervised learning methods except neural networks. We choose the same kernel as in the homework, the Gaussian radial basis function (RBF) kernel.

The decision to use the RBF kernel, $K(x_m, x_n) = exp(-\gamma||x_m - x_n||^2)$ is largely becauses our data is not linearly separable. We see this is true for 2-dimensions using PCA (Figure 2), and the RBF kernel nonlinearly maps features to labels. In addition, there are only two parameters γ and C, which hopefully reduces overfitting.

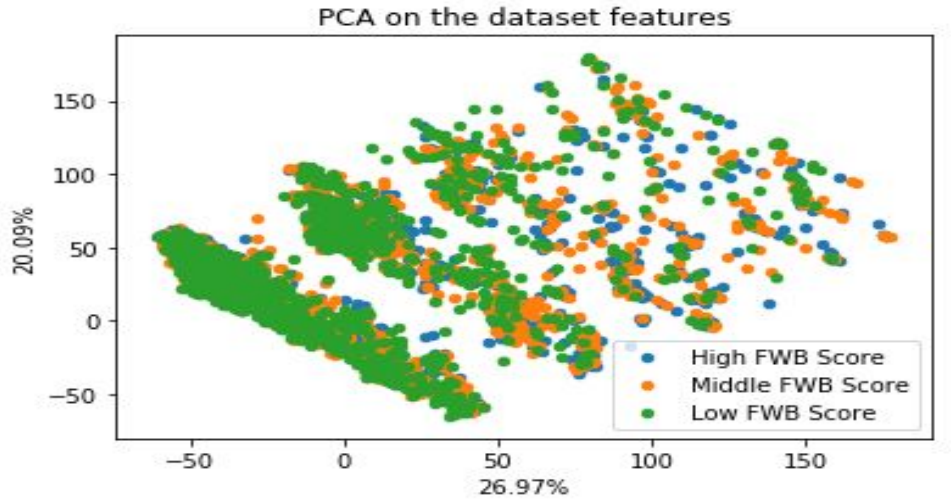
SVM uses the nonlinear decision boundary $\sum_{n=1}^{N} y_n \alpha_n K(x_n, x) + b = 0$ which comes from minimizing $L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{n=1}^{N} \alpha_n \{y_n(w^T k(x_n) + b) - 1 + \varepsilon_n\} + C \sum_{n=1}^{N} \varepsilon_n - \sum_{n=1}^{N} \mu_n \varepsilon_n$. The variables are defined as follows: $K(x_m, x_n) = exp(-\gamma||x_m - x_n||^2)$ and $k(x) \varepsilon R^N$ where each element $k_n(x) = K(x, x_n)$, $\alpha_n \geq 0$ are Lagrange multipliers, ε_n are slack variables penalizing points on the wrong side of the classifier, C is a regularization term that scales the error, and μ_n are also Lagrange multipliers.

In order to extend this to three classes we make three binary classifiers for each pair of classes and classify an instance based on majority vote. If there is a tie in the votes, in this case each binary classifier classifies it differently, we classify based on the greatest magnitude of $\sum_{n=1}^{N} y_n \alpha_n K(x_n, x) + b$. This approach is referred to as the 'one-versus-one' method.

For two classes, we choose boundary FWB score to be 49. The choice of this decision comes from a qualitative description CFBP[3]. CFPB considers someone with a score lower than 49 as relatively financially insecure. The histogram of FWB scores resemble a Gaussian distribution, as mentioned previously. This is not uncommon in social science applications, so therefore we take the distribution into account when choosing how to divide the dataset into three classes. For three classes, we use the 33rd percentile and 66th percentile as our cutoff, corresponding to FWB scores 50 and 62, respectively. We reserve two thirds of our dataset,

distributed equally among the classes, for training our data. The total sample size of our data is 6,390.

Figure 2: PCA on the 150 utilized features of the dataset



In order to get a better understanding of our dataset, we linearly project it onto two dimensions using Principal Component Analysis (PCA). As shown in Figure 2, a linear projection does not clearly separate our dataset into three groups by Financial Well-Being. The first two principal components account for 47% of the variance in our data. These results, as well as the clustering due to an unknown source, indicates to us that a nonlinear classification method is necessary. We also see that a low parameter C, allowing misclassification, is more likely to yield good results. Unfortunately, we will not be able to use PCA to visualize the differences between our SVM output and the original labels. The fact that the data is not linearly separable in lower dimensions reinforces our belief that SVM with a radial kernel is an appropriate choice. From the PCA it is clear that we require higher dimension to cluster data into classes. Situations like these are where SVM is most effective and useful. The lack of linear separability emphasizes the need for a non-linear transformation which is a defining feature of kernel methods like SVM.

Accuracy is scored with 10 fold cross validation. In this scheme, the entire data set is randomly partitioned into 10 partitions. The model is trained on 9 of the partitions and tested on the one left over. This is done 10 times with a different partition used for testing. The accuracy is then averaged over the 10 iterations for a final accuracy score.

Algorithm

To train an SVM classifier $\sum_{n=1}^{N} y_n \alpha_n K(x_n, x) + b$, we need to calculate α_n for all $n = 1, ..., N$ and b. For α we will minimize $L(w, b, \alpha)$'s equivalent dual representation $L`(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m K(x_n, x_m)$ for convenience. This is done using Stochastic

Gradient Descent because it is often too costly to find an exact solution and the approximation generally performs well. Below is the pseudocode.

Set $\alpha = 0 \in R^N$
For epoch = 1,...T
    For x_i in the dataset
        if $\sum_{n=1}^{N} y_n \alpha_n K(x_n, x) < 1$
            $\alpha = \alpha - \gamma \nabla L`(x_i)$
Return α

Here γ is our learning rate chosen to be "small" and **∇** is the gradient function. The bias term b is then calculated as described in *PRML* (Bishop 334)[4] after finding α.

---

set $b = 0$
For x_i in the dataset
    if α_i < C **//x_j lies on the margin**
        set innerSummation = 0
        For x_j in the dataset
            if $y_j * \sum_{n=1}^{N} y_n \alpha_n K(x_n, x_j)$ = 1 **//x_j is a support vector**
                innerSummation += $\alpha_j * y_j * K(x_i, x_j)$
        b += y_i - innerSummation
Return b / number of vectors on the margin

To find γ and C that allows the classifier to perform well, we use a grid search algorithm. The algorithm is supplied with a set of γ's and C's to perform an exhaustive search by pairing every γ and C and scoring them against a test data set.

Results

As shown in Table 1, we find that we are able to correctly classify survey respondents into two groups with an accuracy of nearly 86% using 150 features. We find in the two-class case that values of γ near 1 and $C \ll 1$ produce optimal results. This is indicative of the fact that we are allowing our model to have a large variance, and permitting a high degree of misclassification. We believe this is desirable as it prevents overfitting, and allows for a greater accuracy when testing on new data. Furthermore, from a qualitative perspective, this indicates that our variables vary a lot across different individuals. This makes sense as survey responses are highly subjective, and our model is able to capture this.

Table 1: Best performance after tuning on 150 feature set.

Classes	γ	Cost C	**Accuracy**
2	1	1	73.43%
2	1	0.0067	85.89%
3	1	1	33.60%
3	0.0001	1000	66.67%

In the three-class case, using all 150 features, our best results occur when $\gamma \to 0$ and $C \to \infty$. Even then, we observe an accuracy of only 66.67%. The parameters indicate that SVM is trying very hard to fit the data, and we are experiencing overfitting. Interestingly, this quite contrary to what we see for two classes. Perhaps the one-versus-one extension of a binary classifier is the cause of the lack of accuracy. More likely, it is due to the format of our data.

We did not scale our data, so even though most of our questionnaire data is on the magnitude 1-10, we have some features, such as life expectancy, that range from 1-100. While performing our analysis, we did not consider that features with greater ranges may dominate the analysis. This is something we might consider implementing in future for the three class case to get better results.(Chih-Wei et. al.).

Another reason why the three class projection may not perform well is the curse of dimensionality. To test this, we utilize only 10% of the 150 features, but the performance of three class SVM does not improve. In fact, it does slightly worse, according the results in Table 2. Thus we conclude that the decrease in performance is not due to the large quantity of features.

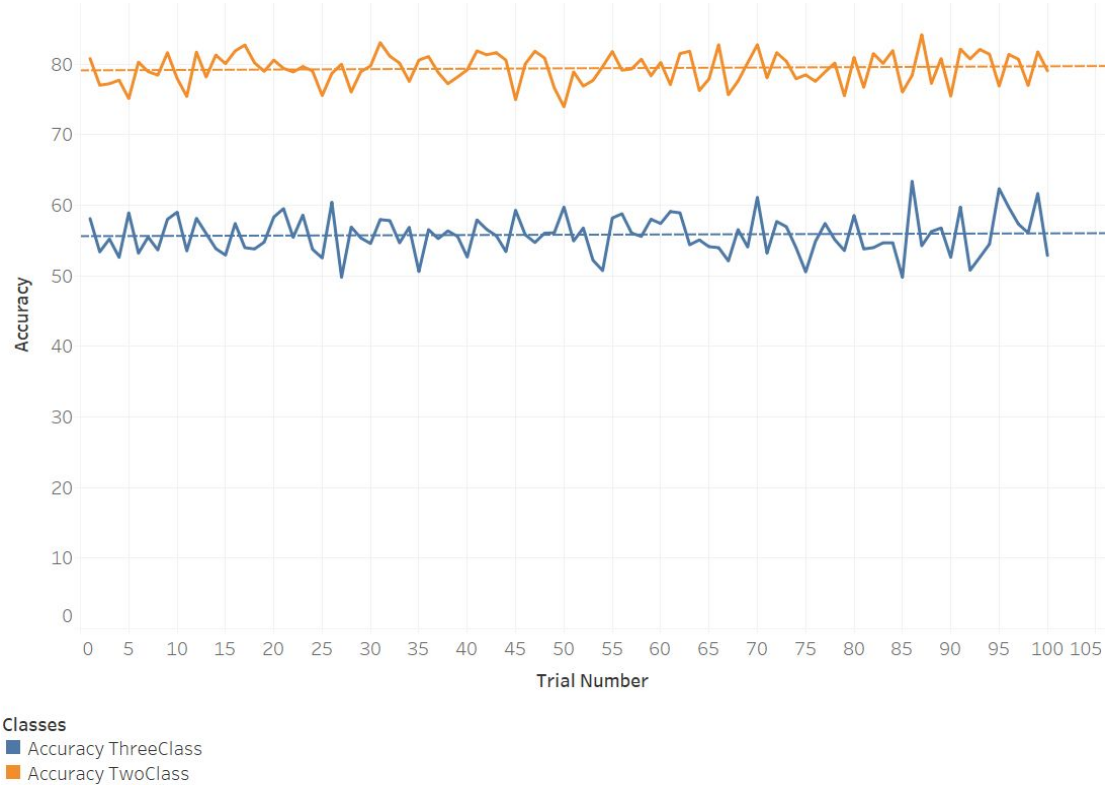
In the second part of project, we want to examine whether there are particular features which help classify better than others. In order to do so, we randomly choose 15 features and trained an SVM model (optimized using tuning and 10-cross fold method as described previously). The accuracy results and summary statistics are reported above. This gives us two interesting results. Firstly, even though the accuracy is lower compared to when we use all 150 features, we see that the decrease in accuracy is not in the same proportion. That is, by using only 10% of features we can get fairly close estimates of the results of the full feature space. By doing this we significantly reduce computational cost without perversely affecting accuracy. Thus, the performance of the SVM model does not drastic drop as we reduce dimensionality.

The second important result that we arrive at from this is that the accuracy does not fluctuate depending on the features we use as long as the dimensionality in the same. For both 2 and 3 class classification the standard deviation across random trials is approximately 2%. This implies that as long as we ask sufficient appropriate question to people taking the survey, it is not of tantamount importance which questions we ask. This again enables us to tremendously reduce computation cost. Furthermore, it provides an interesting insight on the data set and the use of SVM itself. It highlights that for this data set as long as we provide SVM with the same dimension feature space it can perform classification. The choice of features is not very important.

Table 2: Trials with 15 random features

Classes	Average Accuracy	Standard Deviation
2	79.31%	2.21%
3	55.734%	2.73%

Figure 3: Accuracy of Trials using 15 random features



In conclusion, we were able to use related information from the dataset such as financial knowledge and habits to classify respondents financial well being for the two class case. SVM did not perform as well for the three class case, although there are some other measures that can be taken to improve the accuracy. In terms of the feature space, we see that using 150 features is ideal, however we can get similar results with just 15 features. This has important implications in reducing computational complexity. Furthermore, the classification accuracy does not change depending on which of the 150 features we select. Unlike other applications of supervised classification, such as on image data, it appears that we do not need to ask a barrage of questions in order to correctly classify an individual. Thus, SVM is an appropriate tool for analyzing and classifying high-risk individuals using short surveys.

Bibliography

1. Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003): 1-16.

2. “Financial Well-Being Survey Data.” *Consumer Financial Protection Bureau*, Consumer Financial Protection Bureau, www.consumerfinance.gov/data-research/financial-well-being-survey-data/.

3. “Measuring Financial Well-Being: A Guide to Using the CFPB Financial Well-Being Scale.”*Consumer Financial Protection Bureau*, Consumer Financial Protection Bureau, www.consumerfinance.gov/data-research/research-reports/financial-well-being-scale/.

4. Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2009.

Appendix 1

CFPB FINANCIAL WELL-BEING SCALE

Scoring worksheet

NAME OR NUMBER

1. Select the person's answers, record the response value in the right hand column and add up the total values for each part of the questionnaire.

This statement describes me	Completely	Very well	Somewhat	Very little	Not at all	Response value
1. I could handle a major unexpected expense	4	3	2	1	0	
2. I am securing my financial future	4	3	2	1	0	
3. Because of my money situation, I feel like I will never have the things I want in life	0	1	2	3	4	
4. I can enjoy life because of the way I'm managing my money	4	3	2	1	0	
5. I am just getting by financially	0	1	2	3	4	
6. I am concerned that the money I have or will save won't last	0	1	2	3	4	

Part 1 subtotal: _____

This statement applies to me	Always	Often	Sometimes	Rarely	Never	Response value
7. Giving a gift for a wedding, birthday or other occasion would put a strain on my finances for the month	0	1	2	3	4	
8. I have money left over at the end of the month	4	3	2	1	0	
9. I am behind with my finances	0	1	2	3	4	
10. My finances control my life	0	1	2	3	4	

Part 2 subtotal: _____

Total response value: _____

2. Find the financial well-being score

How old is the person?

☐ 18-61 ☐ 62+

How did the person take the questionnaire?

☐ Self-administered

☐ Administered by someone else

Because scores vary based on age and how the questionnaire was administered, you must convert the total response value to a financial well-being score.

a. Find the row that corresponds to the total response value.

b. Follow that row across to the column that corresponds to the person's age and how the questionnaire was administered.

c. Record the final score.

Financial well-being score: _____

Total response value	Questionnaire self-administered		Questionnaire administered by someone else	
	18-61	62+	18-61	62+
0	14	14	16	18
1	19	20	21	23
2	22	24	24	26
3	25	26	27	28
4	27	29	29	30
5	29	31	31	32
6	31	33	33	33
7	32	35	34	35
8	34	36	36	36
9	35	38	38	38
10	37	39	39	39
11	38	41	40	40
12	40	42	42	41
13	41	44	43	43
14	42	45	44	44
15	44	46	45	45
16	45	48	47	46
17	46	49	48	47
18	47	50	49	48
19	49	52	50	49
20	50	53	52	50
21	51	54	53	52
22	52	56	54	53
23	54	57	55	54
24	55	58	57	55
25	56	60	58	56
26	58	61	59	57
27	59	63	60	58
28	60	64	62	60
29	62	66	63	61
30	63	67	65	62
31	65	69	66	64
32	66	71	68	65
33	68	73	70	67
34	69	75	71	68
35	71	77	73	70
36	73	79	76	72
37	75	82	78	75
38	78	84	81	77
39	81	88	85	81
40	86	95	91	87

Learn more at
consumerfinance.gov/financial-well-being

cfpb Consumer Financial Protection Bureau