EE 58*J* Data Mining
Project 3

Oğuzhan Sevim
May 9, 2020

I. INTRODUCTION

In this project, we will implement support vector machine (SVM) classifiers on the precomputed CNN features of Vispera-SKU101 – 2019 dataset. We will design/train an SVM classifier for each of the five product categories. Different regularization parameters C and kernel functions will be experimented to find the optimal SVMs. Python programming language is preferred for this assignment.

II. DATA

Vispera-SKU101 – 2019 data is composed of approximately 10.000 arbitrary–sized images in jpeg format. Each image belongs to one of 101 classes, and is kept in the folder named with SKU (stock keeping unit) it belongs to. Each of these class folders contains approximately 100 instances/images. An example image is given in Figure 1 in which the image belongs to class of "33*cl* Fanta in can" with the class SKU of 8736.

Fig. 1. An example image with 75×224 pixels. The image belongs to "33*cl* Fanta in can" class with SKU of 8736

III. CATEGORY 1: ICE-CREAM

In this section, we will train and optimize an SVM for the classification of ice-cream category. First, the model is trained with radial basis function kernel. K–fold cross validation with $K = 10$ is used in all the models. When we train the model with different regularization parameter C and gamma of RBF function, cross validation accuracies are obtained as shown in Table I. Best cross validation accuracy is obtained to be 96.09% for the model with RBF kernel.

When we experiment with linear SVM, which requires less computations, we get the results shown in Table II. Since higher C makes misclassifications less tolerable, accuracy gets higher as we increase C. After a certain level, we get the maximum achievable accuracy of 96.09%.

TABLE I
10−FOLD CROSS VALIDATION ACCURACIES FOR DIFFERENT C (VERTICAL) AND GAMMA (HORIZONTAL) PARAMETERS

	1.00E-08	1.00E-06	1.00E-04	1.00E-02
1.00E-02	2.51	2.51	2.51	2.51
1.00E+00	2.51	2.51	93.17	92.35
1.00E+02	2.51	93.17	95.85	92.87
1.00E+04	93.17	95.68	95.91	92.87
1.00E+06	95.79	96.09	95.91	92.87

Since the same accuracies can be obtained by using linear and RBF kernels, choosing the simplest model (linear one) would be the wise move. When we train a linear SVM ($C = 10$) by using all the training set, we get the test accuracy of 97.20%. Mostly confused classes are *sku*.92 and *sku*.94.

TABLE II
ICE-CREAM DATA 10−FOLD CROSS VALIDATION ACCURACIES

C	CV acc
1E-06	2.51
0.00001	3.86
0.0001	91.35
0.001	94.92
0.01	95.91
0.1	96.09
1	96.09
10	96.09
100	96.09

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	20	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	19	0	1	0	0	0	0	0	0	0
12	0	0	0	0	1	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0
16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	19	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	18	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0
20	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	16

Fig. 2. Confusion matrix for ice-cream category where test accuracy is 97.20%. Mostly confused classes are *sku*.92 and *sku*.94 where there seems to be no particular relation between these two classes.

For the remaining categories, we will experiment only with the linear SVMs. 10−fold cross validation will be used for choosing the right model.

IV. Category 2: Soft drinks-I

TABLE III
SOFT DRINKS-I DATA 10−FOLD CROSS VALIDATION ACCURACIES

C	CV acc
0.0001	51.73
0.001	74.57
0.01	80.54
0.05	83.19
0.09	83.69
0.1	83.56
1	82.45
10	82.39
100	82.39

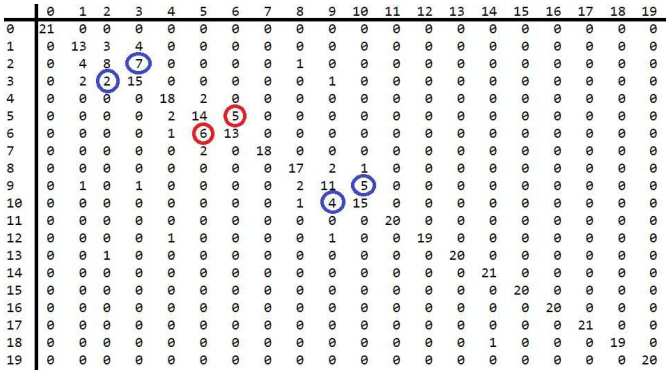

Fig. 3. Confusion matrix for Soft drinks-I category where test accuracy is 84.48%. Mostly confused classes are 8703 and 8706. They are Coca-Cola in plastic bottles of 2 and 2.5 liters. The second and third mostly confused class pairs are 8693 – 8697 (Coca-Cola in can) and 8720 – 8725 (Sugar-free Coca-Cola in can). Even with my eyes, I couldn't perform any better on distinguishing between these classes.

V. Category 3: Soft drinks-II

TABLE IV
SOFT DRINKS-II DATA 10−FOLD CROSS VALIDATION ACCURACIES

C	CV acc
0.00001	4.16
0.0001	98.84
0.001	99.14
0.01	99.45
0.1	99.45
1	99.45
10	99.45
100	99.45

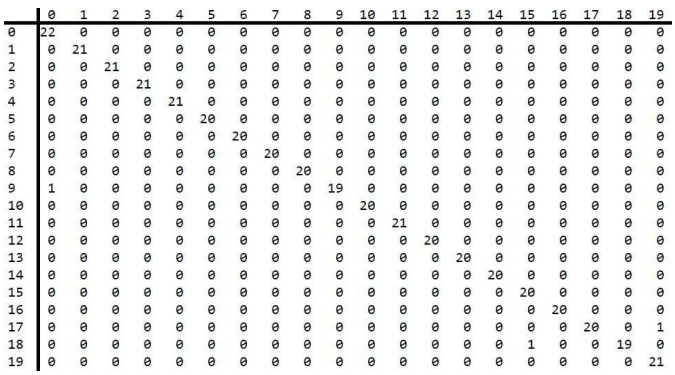

Fig. 4. Confusion matrix for Soft drinks-II category where test accuracy is 99.27%. There is only few mistakes in the test set. So, no useful conclusion can be drawn from mostly confused classes.

VI. Category 4: Laundry

TABLE V
LAUNDRY DATA 10−FOLD CROSS VALIDATION ACCURACIES

C	CV acc
0.00001	7.63
0.0001	50.54
0.001	80.41
0.01	89.30
0.06	89.72
0.1	89.18
1	88.22
10	88.22

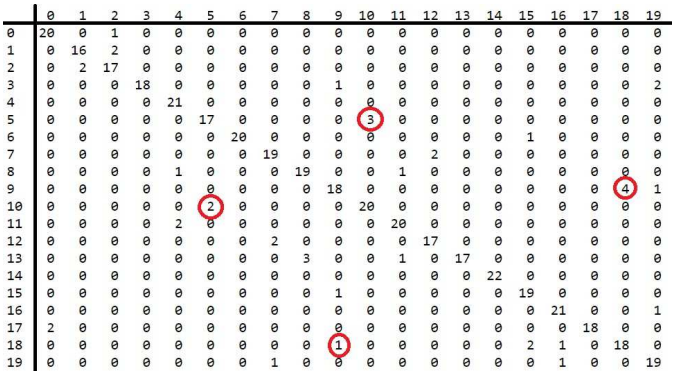

Fig. 5. Confusion matrix for Laundry category where test accuracy is 90.17%. Mostly confused classes are 10988 – 9686 (both Ariel in green box) and 9684 – 9921 (both Ariel parlak renkler).

VII. Category 5: Confectionery

TABLE VI
CONFECTIONERY DATA 10−FOLD CROSS VALIDATION ACCURACIES

C	CV acc
0.00001	3.57
0.0001	74.28
0.001	92.98
0.01	93.34
0.02	93.40
0.1	92.44
1	92.13
10	92.13

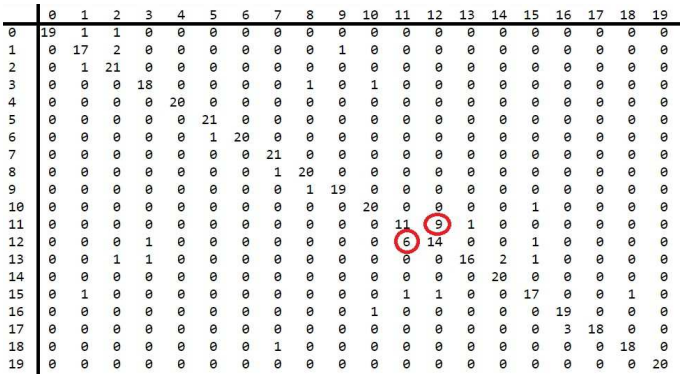

Fig. 6. Confusion matrix for Confectionery category where test accuracy is 90.17%. Mostly confused classes are 6866 and 6867 where both classes are the same product: Altınbaşak Grissini.