Glass vs No-Glass Classification

Vaibhav *(EE-Department),* Neeraj *(CSE-Department),* Tanuj, *(EE-Department)*

Abstract

The course's final project requires us to build a model capable of classifying a dataset of images with glasses and images with no-glasses. The initial models used were Logistic Regression, Principal Component Analysis, LinearDiscriminant Analysis and Random Forest Classifier. The Data set is provided to us having approximately 6000 images. We analysed the data set. We analysed the performance of different classifiers on the data-set. We evaluated the entire pipeline with cross validation

I. Introduction

First the core concepts are described here that are applied in this project. Then further the whole written code is described. inally results and conclusion are obtained.

A. Core Concepts

Here we have our data set with different features of images given. We can analyse the whole data set first. Further, it can be seen how many images of our data set have images with sunglasses and how many images do not have sunglasses. This can be done using **matplotlib**. Let's understand in short about matplotlib.

*1) **Matplotlib**:* Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI tool-kits like Tkinter, wxPython, Qt, or GTK+

Then we can apply different classifiers on the given data set and compare their performances.First we used **Random Forest Classifier.**

*2) **Random Forest Classifier**:* Let's see what us Random decision forests.Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees. Second we used logistic regression.

*3) **Logistic Regression Classifier**:* Now, let's understand about logistic regression.It is an appropriate regression analysis to conduct when the dependent variable is binary. The logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. As we have binary data set, so, it becomes easy to apply this classifier. Then we applied LDA as a classifier

*4) **Linear Discriminant Analysis**:* Now, let's understand about Linear Discriminant Analysis Classifier. Logistic regression is a classification algorithm traditionally limited to only two-class classification problems. If you have more than two classes then Linear Discriminant Analysis is the preferred linear classification technique. That;s why we used this Classifier for our data set.

Then we applied the concept of dimensionality reduction using PCA.Let's understand both the terms.

*5) **Dimensionality Reduction**:* The number of input variables or features for a data set is referred to as its dimensionality. Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset.

More input features often make a predictive modeling task more challenging to model, more generally referred to as the curse of dimensionality.

High-dimensionality statistics and dimensionality reduction techniques are often used for data visualization. Nevertheless these techniques can be used in applied machine learning to simplify a classification or regression dataset in order to better fit a predictive mode

*6) **Principal Component Analysis**:* Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of "summary indices" that can be more easily visualized and analyzed

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

Finally we applied the concept of cross-validation and applied it on each above-mentioned classifier and compared their performances.

*7) **Cross-Validation**:* Cross-validation is a technique for evaluating a machine learning model and testing its performance. CV is commonly used in applied ML tasks. It helps to compare and select an appropriate model for the specific predictive modeling problem.

Now let's analyse the complete code in details and finally observe the results

II. Analysis of Code

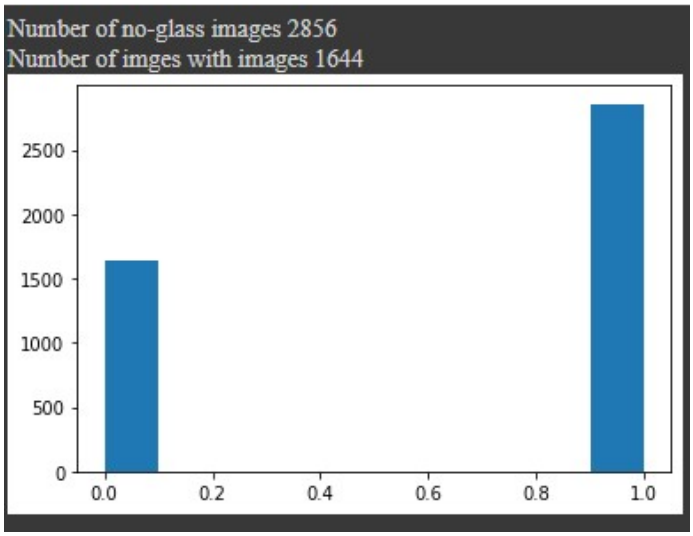
*1) **Data set**:* The data set is given with different features obtained from pixels of the images. Here are the images that are expected to be tested

Here is the csv format of our data set:

	id	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20	v21
0	1	0.37797	-0.94808	0.01346	0.17893	0.37795	0.63571	0.13943	-0.25607	-0.39341	1.08947	-1.36048	-1.31529	0.36119	-0.62857	-1.65290	1.47167	-0.88741	-0.25942	-0.34414	-0.38914	0.38425
1	2	0.07609	-0.09774	0.39666	-0.39026	0.10606	0.52774	0.07105	0.33720	0.69917	-0.02842	-0.56431	0.42060	-0.47533	1.60025	-0.02106	-2.30193	-0.31614	-0.08050	0.85041	-0.10574	-0.33177
2	3	1.19391	-0.68707	-0.68422	-0.36378	-0.60847	-0.40118	1.45432	0.00592	1.68940	-0.98205	0.67435	-1.27604	-1.37050	-0.91682	0.23617	0.53676	-0.26815	-0.66989	0.46076	-0.04117	1.34078
3	4	1.34949	-0.31498	-1.30248	0.50278	1.66292	-1.06094	-0.70835	-0.24237	-0.15509	-0.04532	0.97349	-0.21032	-0.71143	0.59725	-0.51849	0.00126	1.22219	0.57160	0.41212	0.90077	-0.80797
4	5	-0.03512	-0.34196	0.14230	1.50513	-0.14364	0.49429	0.07823	-0.04356	0.42009	-0.88828	0.13601	0.26917	0.11973	0.04378	1.06879	0.10060	-0.83331	-0.64776	0.26461	0.65249	-0.41807

5 rows × 514 columns

*2) **Data Visualisation**:* The data set was quite simple. So, we really did not require to do any kind of extra preprocessing. We just assigned the assigned the features to a variable and target to other and counted the images with glass and without glass. We got the following result:



III. Evaluation using Models

Once we had the visualized dataset we did train-test split of 85-15 We used the training dataset to train 4 models that are - Logistic Regression,Linear Discriminant Analysis and Random Forest Classifier

1) ***Logistic Regression****:* A model of Logistic Regression was trained with default parameters from sklearn. On the validation set, the model gave an accuracy of 0.86555589988888

2) ***Linear Discriminant Analysis****:* A model of Linear Discriminant Analysis was trained with default parameters from sklearn. On the validation set, the model gave an accuracy of 0.8755555555555555

3) ***Random Forest Classifier****:* A model of Random Forest Classifier was trained with default parameters from sklearn. On the validation set, the model gave an accuracy of 0.8355555555555555

IV. Dimensionality Reduction

As we can see we go very less accuracy. So to improve the accuracy score we applied the concept of dimensionality reduction using Principal Component Analysis. Here are the accuracies we got after Applying PCA:

Models	Random Forest Classifier	Linear Discriminant Analysis	Logistic Regression
Accuracies	0.9322222222222222	0.9744444444444444	0.9822222222222222

V. Contributions

Vaibhav Meena- Visualization of datasets, added comments and contributed in report writing.
Neeraj- Trained the models with all the three Classifiers and Documented the same.
Tanuj Bhardwaj - Applied Cross Validation and Dimensionality Reduction on the dataset using PCA

VI. References

1. https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html
2. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
3. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

A. Pipeline of the complete project