Homework 1

Submitted by Vishantan Kumar
February 20, 2024

Part 1: Short Answers

Ans 1. Features that are manually designed and not learned by the machine are hand-crafted. Some examples are:

- Histogram of Oriented Gradients (HOG): Distributions of intensity gradient directions. Capture the outlines by calculating the orientation of edges
- GABOR filters: These are defined by sinusoidal waves and analyze spatial frequencies in images, capturing texture information.
- Local Binary Patterns (LBP): Labels the image pixels by binary thresholding and calculates its histogram.

Ans 2. Features that are calculated and learned by a machine learning model by observing various features from the data during training. These aren't specified by humans and can include:

- Text embeddings: Representations of words in high-dimensional space. Can be vector such as in Word2Vec or other HD features.
- Autoencoder embeddings: Neural networks that learn image features by compression followed by reconstruction.
- Recurrent Neural Network embeddings (RNN): Used in sequential datasets such as audio or video and learns to capture temporal dynamics.

Ans 3.
Advantages:

- Efficient. Require less computational power since they are targetted and selective.
- More interpretable. Learned features might not be intuitively obvious to humans.
- Performance with smaller datasets. Since they are manually crafted to better fit the data they perform well even without much training data being available.

Disadvantages:

- Poor generalisation and adaptability. Since they are tailored towards one specific dataset, they might generalise poorly
- Complexity. Require a lot of careful calibration and specialisation
- Time taken to generate. Since they require manual experimentation, they might be a lot more time consuming to generate and require significant domain expertise

Ans 4. The two ways are:

- Singular Value Decomposition of Data Matrix - This formulation looks to minimize the reconstruction error when the original HD matrix is approximated from its LD representation
- Eigen Value Decomposition of Covariance Matrix - This maximizes the variance captured and identifies the directions that do the same i.e., the largest eigen values of the Covariance Matrix.

Ans 5. LDA is used for dimensionality reduction and. For a dataset with n classes it'll give $n - 1$ features.

Ans 6. Limitations of LDA are as follows:

- It can't reduce the number of features to less than number of classes-1
- Not optimal for highly correlated or non-linear features
- Assumes normal distribution. Performs poorly otherwise

Ans 7. The drawbacks of PCA are as follows:

- Assumption of linearity and orthogonality
- Primary emphasis on variance can lead to poor performance as a classifier since the direction of maximum variance might not be the best decision boundary.
- Data loses interpretability upon performing the transformation

Ans 8. Distance metrics used for histogram comparisons are as follows:

- Euclidean Distance: It is the l_2 norm and treats histograms as vectors in a multidimensional space
- Chi-Squared distance: Measures the divergence between 2 histograms and is particularly useful in comparing non-negative values such as pixel intensities.
- Bhattacharya distance: It measures the similarity between 2 histograms by considering them as 2 probability distributions and therefore, doesn't just use the bin differences.

Ans 9. The l_0 norm is equivalent to the number of non-zero elements in the data, which is also a metric of vector sparsity since sparsity is also dependant upon the number of non-zero elements. Hence, higher the norm, lower the sparsity

Ans 10. There are 2 primary reasons for this:

- The l_1 is convex as opposed to l_0 which makes it much more easy to optimize
- l_1 induces sparsity in the vector while l_0 measures the sparsity, hence for several classification based problems l_1 is better

Ans 11. The disadvantages of K means are as follows:

- Centroid Initialisation: The algorithm is heavily dependent on where the centroids are initialised. Non-accurate initialisation can lead to very poor results
- Sensitivity to outliers: For smaller values of k, even smaller number of outliers can produce large errors
- Assumption of sphericality: It assumes the clusters to be radially centralized which might not be the case always.

Ans 12. Nearest neighbours is a special case of kNN where $k = 1$. It assigns the classes based on the class of its nearest geometrical neighbour, while kNN does it as the modal class of the samples' k nearest neighbours

Ans 13. Visual Bag of Words (BoW) features are extracted by performing:

- Feature Detection: Detection of keypoints. Can use SIFT, SURF, ORB etc.
- Feature Description and clustering: Describing local features around the keypoints
- Feature Quantization: Descriptors extracted around keypoints from a new image. Maps the features to a visual word they might be most similar to
- Histogram Generation: Histogram that represents the frequency of each visual word
- Aggregation: Histograms from each image in the set and normalised to form a codebook

Ans 14. Cross-validation is a method of evaluating the performance of a dataset on the training set. It doesn't make use of the test set and is primarily used in hyper-parameter optimization. It can be done in several ways, with the most common being k-fold wherein the training set is divided into k subsets with the model being trained on $k-1$ subsets and k^{th} set being used for testing.

Ans 15. Sparse Coding is the process of representing data as a sparse combination of the training set elements. Dictionary learning is the method of *learning* the best dictionary for sparse representation. These 2 are complementary processes with dictionary learning being a precursor to sparse coding, however, in sparse coding the dictionary can be learned or hand crafted whereas in Dictionary Learning it is learned.

Part 2: Face Recognition (kNN)

2.1.1 Error as a function of Training Set size

Following is the plot of the test error against the training size. We can see that the test error decreases from 55% to almost 25% as the size of the training set increases from 10 to 50 samples per class. This is pretty much expected since the error should decrease as the number of training samples increases since more information can be captured.

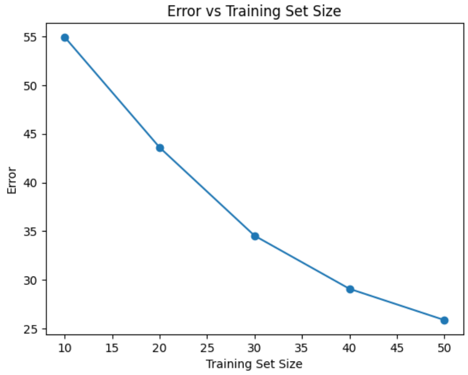

Figure 1. Test error vs Training Size

Summarising : Lowest error is 25% at 50 training samples per class.

2.1.2 Error as a function of k

Following is the plot of the test error against the training size for different values of k. From the figure, we can see

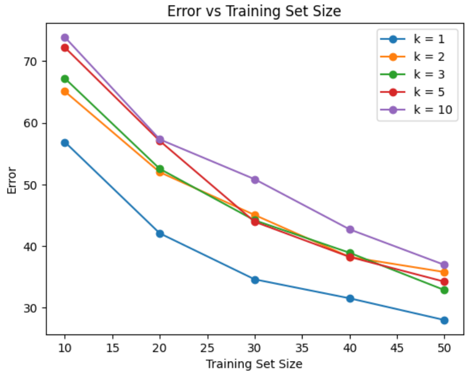

Figure 2. Test error for various k

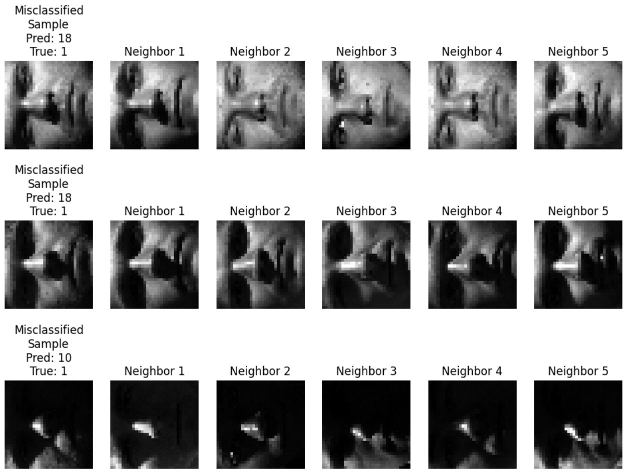

Figure 3. Misclassified samples and their 5 nearest neighbours

that the performance is best for $k = 1$ and progressively decreases as we increase the number of nearest neighbours. This shows that as in many cases of k-nn, as k increases, the bias increases and thus leads to the model underfitting. We can also see from Figure 3 some of the misclassified samples along with their nearest neighbours. We can see that in most cases the nearest neighbour is the same class however, the others belong to a different class, leading to higher probability of misclassification

Summarising: Lowest error is for $k = 1$ which is around 30% for 50 training samples per class.

2.1.3 Error as a function of Minkowski Distance Metric

We next analysed the error vs the Minkowski distance metric p. Minkowski distance is defined as:

$$D(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p} \tag{1}$$

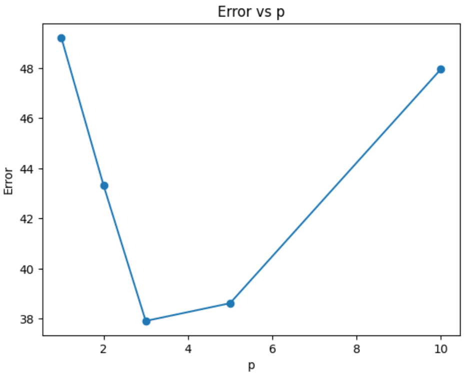

Figure 4. Error vs Minkowski Parameter

Where p is a hyper-parameter. Subsituting $p = 1$ gives us the l_1 or Manhattan Distance while $p = 2$ gives us the l_2 or Euclidean distance. From 4 we can see that the relation between the error and p is non monotonic. The error decreases as p increases from 1 to 3 but starts increasing as p is increased further.

Summarising: The minimum error is achieved at $p = 3$ and is equal to 38%

2.1.4 Using LBP and HOG features

We then compared the efficiency of the model in the case of using HOG and LBP features as opposed to image intensity for 30 training samples per class.

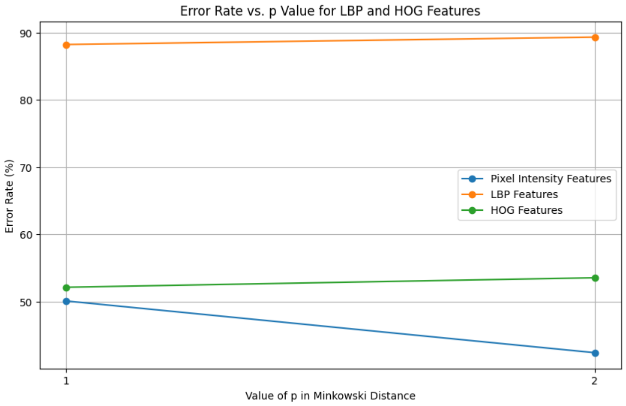

Figure 5. Classification error for LBP, HOG and Intensities

From Figure5, we can see that both LBP and HOG errors remain pretty much constant at 89% and 52% respectively as p increases while the pixel intensity feature decreases slightly from 50% to 45%. This can be explained as follows:

- LBP features are primarily helpful in determining textural differences. Since all the images are facial datasets, they have similar textures and hence cannot be well differentiated by LBP. Hence, HOG and intensity have a much higher accuracy than LBP
- Since the errors are constant, we can assume that the feature space is low dimensional, leading to not significant differences between l_1 and l_2 distances and resulting in similar error rates.

Summarising: LBP has much higher errors (90%) than HOG (52%) and Intensity $(50-45\%)$ with there being little difference between Minkowski parameters.

2.1.5 Minimum Error

The minimum error obtained throughout was around 25% with $k = 1$, number of training samples per class = 50, and Distance metric = *Euclidean*.

2.2 Cross Validation

For the Cross validation we implemented a GridsearchCV for $k = [1, 2, 3, 5, 10]$ and $p = [1, 2, 3, 5, 10]$. Upon check-

Number of Neighbours (k)	1
Minkowski Metric (p)	5
Cross-Validation Score	72.5%
Test-set error	24%

Table 1. Cross Validation Results

ing the best parameters we got the results as highlighted in Table 1. The best accuracy was found with $k = 1$ and $p = 5$, where the calculated error was around 24%. We could've implemented it as 1-fold cross-validation, however we tried a 5-fold cross-val which gives better accuracy.

Part 3: Face Recognition (Other Algorithms)

3.1 Principle Component Analysis

We performed dimensionality reduction using PCA followed by NN classification. The error ranged from 62% for 10 samples to 33% for 50 samples.

3.2 Linear Discriminant Analysis

Dimensionality reduction using LDA was performed followed by NN classification. The error ranged from 20% for 10 training samples to 4% for 50 training samples, with an abrupt increase to 26% error for 30 samples.

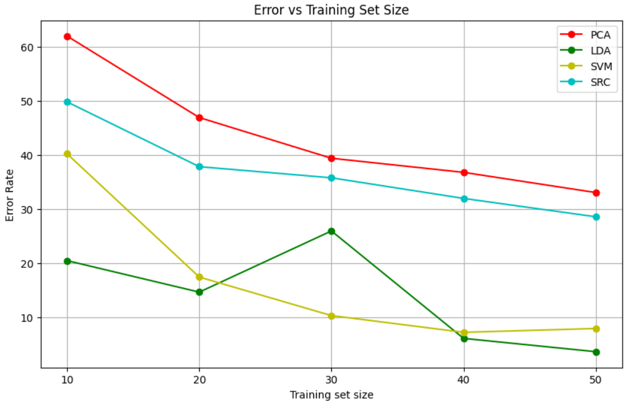

Figure 6. Test errors vs Training Set size for other methods

3.3 Support Vector Classifier

SVC with GridSearchCV to find optimal hyperperameters yielded the following results. The error decreased from 40% for 10 samples to around 8% for 50 samples. The optimal parameters were found to be: $C = 100$ and $\gamma = 0.01$

3.4 Sparse Representation Classification

We implemented an SRC based classification using Orthogonal Matching Pursuit (OMP) to find the sparsest representative vector and finding the class label which most accurately reconstructs the original image. The error was on a higher end ranging from 50% for 10 samples to 29% for 50 samples.

3.5 Results

The summary is that LDA provided the best accuracy, with just 4% error for 50 samples, however, SVM shows a much better performance for 30 samples at 90% accuracy while LDA showed a dip in performance. In totality though, LDA was the best performing algorithm for thie classification.