Adv Machine Learning Sheet 1

Amol Kapoor

May 2017

1 Eigenfaces

1. **Reconstruct a picture not in the database with eigenfaces. Is the reconstruction good?**

   No, it isn't. Eigenfaces excels when the images are all around the same standard for pixel values - in other words, the faces are all in the same general orientation with the same lighting. The picture I took of myself has a notably different background and very different lighting. Thus, even though the face alignment was around the same spot, the pixel values were functionally inverted - my head was mostly in shadow while the background was well lit. This type of image doesn't have a good eigenface representation, and so the vector basis that was created using the dataset could not accurately reconstruct the image.

2. **What are the advantages and disadvantages of this method?**

   Advantages

   Speed. Eigenfaces is very fast with no training time and fairly good results. No previous knowledge of faces is required, can learn directly from data provided.

   Intuition. Every step of the eigenfaces procedure has an intuitive explanation that can be visualized.

   Compression. Eigenfaces takes a dataset of 36000 bytes and compresses it down to a far smaller amount of floats.

   Disadvantages

   As mentioned above, eigenfaces is unable to handle differences in lighting, background, and head alignment (and any kind of transposition within the image, including rotation and inversion). In general, the more different the image is from the data set, the worse eigenfaces will perform. In addition, other limitations include:

Scaling. Images must be provided with the right scale for the system to be able to properly process and categorize.

Expression. In general the eigenfaces basis vectors represent different facial structures, and a wide variety of faces can be constructed from various amounts of each basis vector. If a facial structure is changed, eigenfaces can have trouble dealing with the input for the same reason it struggles with different backgrounds and lighting.

Cannot learn on-line. There is no easy way to quickly update the vectors used by eigenfaces to account for new faces added to the data set.

Only useful when the data set is smaller than the actual number of pixels in each image. For really large data sets, eigenfaces actually decreases compression. Further, large data sets with similar images can result in increased likelihood of conflating different faces, which can lower the recognition rate.

2 Word2Vec

1. **Discuss observations on results. Explain if the plots met expectations.**

   Observations: The words are broadly clustered by semantic domain. Three domains were used (food, transportation, and office supplies). Within each domain, words that could have multiple meanings (sub meaning sandwich or sub meaning submarine) were also included. Words with multiple meanings were less likely to be clustered at the edges of the plot - for example, sub was plotted somewhere between the three clusters. Some words were unexpectedly centered, for example the words 'staples' and 'carrot', but for the most part the plots show the expected clustering. The MDS and PCA plots show slightly different word locations, but this is expected as MDS uses a random seed to start minimizing the strain between various points.

   Thus, both plots met expectations.

2. **What similarities and differences did you find between both plots?**

   As briefly mentioned above, some of the relative locations between points changed - for example, in the PCA plot the word 'eraser' is to the left of 'highlighter', while in the MDS plot the word is to the right of 'highlighter. For the most part these differences are minor and are constrained to individual clusters. When looking at the clusters as a whole, or even eyeballing individual word distances between clusters, the two plots appear to be very similar. Again, this matches expectations, as MDS attempts to minimize strain given a set of euclidean distances and does not care about the objective mapping locations, whereas PCA attempts to create a new vector basis to plot on and then plot objectively.