CS60050 Machine Learning - Weekly Report

Nisarg Upadhyaya (19CS30031)

Week 9: 6 and 8 October

1 Topics covered

- Dimensionality Reduction: Motivation and Approaches
- Subset Selection
- Principal Component Analysis
- Linear Discriminant Analysis

2 Summary

2.1 Dimensionality Reduction: Motivation and Approaches

Dimensionality reduction helps in decreasing the memory and computation cost. It also helps decrease the complexity of the inference algorithm during testing. Simpler models are more robust on small datasets and have less variance. It also becomes convenient to extract knowledge of the data, plot and visualise it. There are two major approaches:

1. Feature Selection: In this we find k of the d dimensions that give us the most information. Subset selection is a feature selection method.
2. Feature Extraction: In this we find a new set of k dimensions that are a combination of the original d dimensions using supervised or unsupervised techniques. Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) are feature extraction methods which project feature vectors to a lower dimensional space.

2.2 Subset Selection

Best subset which contains least number of dimensions that most contribute to accuracy, using a suitable error function. Let F be the set of input dimensions and $E(F)$ be the error in validation set when F is used as input. There are two greedy methods:

1. Forward selection: We start with $F = \phi$. At each step choose $j = argmin_i E(F \cup x_i)$ and add x_j to F if $E(F \cup x_j) < E(F)$. Repeat till no further additions are possible. This is a local search procedure and does not guarantee finding the optimal subset. If there are originally d dimensions the time complexity is $O(d^2)$.
2. Backward selection: We start with F as the set of all features. At each step choose $j = argmin_i E(F - x_i)$ and remove x_j from F if $E(F - x_j) < E(F)$. Repeat till no further removals are possible. This is also a local search procedure and does not guarantee finding the optimal subset. Its time complexity is also $O(d^2)$. Training a system with more features is more costly than training a system with fewer features, and forward search may be preferable especially if we expect many useless features.

2.3 Principal Component Analysis

Let x be the input feature vector of dimension d and w is a direction unit vector also of dimension d. The projection of x on the direction of w is $z = w^T x$. The first principal component is w_1 such that that its variance is maximum among all possible projections. We want to maximize $Var(z_1) = w_1^T \Sigma w_1$ subject to constraint $w_1^T w_1 = 1$. Formulating this as a lagrange problem we have $w_1 = argmax_w \{w^T \Sigma w - \alpha(w^T w - 1)\}$. Solving this lagrange problem we get w_1 as the eigen vector of Σ corresponding to the maximum eigen value. The second principal component w_2 should also maximize variance, be of unit length and orthogonal to w_1. Formulating this as a lagrange problem we have $w_2 = argmax_w \{w^T \Sigma w - \alpha(w^T w - 1) - \beta(w^T w_1 - 0)\}$. Solving this lagrange problem we get w_2 as the eigen vector of Σ corresponding to the second maximum eigen value.

2.3.1 PCA Algorithm

Given a set of data points in the d-dimensional space, it outputs the set of k eigen vectors. We compute the mean of the data points and translate all data points to their mean. We compute covariance matrix of the set and their eigen values and vectors. Choose k such that the fraction of variance accounted for is more than a threshold and use those k components for representing any data point.

2.3.2 Properties

PCA diadiagonalizes the data covariance matrix. We can write $\Sigma = CDC^T$ where D is a diagonal matrix and columns of C are unit eigen vectors. The components are uncorrelated. We can normalize components with their variances (eigen values), and use euclidean distance for classification.

2.3.3 Applications

It can be used for data compression. It provides an optimum set of orthonormal basis vectors for a set of data points. The transformation is called *Karhunen-Loeve Transform* (KLT). Color images in RGB space are highly correlated and PCA can be used for decorrelating components. Factor analysis highlights decorrelated factors which is useful for classification. For example, if we perform PCA on a large set of images of human faces cropped to the same size then any arbitrary face can be expressed as linear combination of them. The principal components can be used for classification and high level processing.

2.4 Linear Discriminant Analysis

It is a supervised method for dimensionality reduction for classification problems. It captures the direction of maximum separation between the groups of data points of differing labels. Consider a set of data points S with N_1 points in class w_1 and N_2 points in class w_2. Projection of data point x_i on u is given by $y_i = x_i^T u$. Let m_1 and m_2 be the mean of data points of the two classes and m_{y_1} and m_{y_2} be their projections on u. A measure of separation is $D = |m_{y_1} - m_{y_2}|$, however, it doesn't consider the variance of the data. The scatter of a data belonging to class C in direction u is defined as $s^2 = \sum_{x \in C}(x^T u - m_c)^2$ where m_c is the mean of the class C. We want to find a u that maximizes $J(u) = \frac{D^2}{s_1^2 + s_2^2}$. This ensures scatter of projected samples is small. Scatter matrix for samples of class C is given by $S_C = \sum_{x \in C}(x - m_c)(x - m_c)^T$. Defining withing the class scatter matrix $S_W = S_1 + S_2$ where S_1 and S_2 are scatter matrices for classes w_1 and w_2 respectively. Defining between the class scatter matrix as $S_B = (m_1 - m_2)(m_1 - m_2)^T$. We can rewrite $J(u) = \frac{u^T S_B u}{u^T S_W u}$. To maximise this we should have $u = cS_W^{-1}(m_1 - m_2)$ where c is some constant. Because only the direction matters we can just take $c = 1$.

3 Challenging concepts

Understanding how the various optimization equations in PCA and LDA are being solved took some time and reading to understand properly.

4 Interesting concepts

None

5 Concepts not understood

None

6 Ideas

Nowadays a lot of authentication systems are based on fingerprint detection. It is useful to have fast and accurate fingerprint sensors in devices such as phones and laptops. However, just like faces the complexity of fingerprints is also high and complex models can take significant time for verification. PCA can help in reducing the complexity which will in turn reduce the inference time for verification purposes, therby providing a faster experience to the user.

7 Quiz Feedback

Difficulty level: Moderate
Time given: A bit less
The quiz questions did enhance my understanding and brought more clarity to the topics taught in class. However, some questions were quite calculative in nature and so I feel a bit more time would have been better.