CS342 Coursework

Shayan Borhani Yazdi

December 2023

Introduction

This report contains the various proofs and discussions related to different part of the Machine Learning coursework.

Task 1

In this task we want to show that we can apply PCA to $\bar{Z} = \phi(X)$ without explicitly computing $\bar{Z}$

Variable Names:
1) X Our data
2) $\bar{Z}$ The data points mapped to a higher-dimensional space by kernel trick
3) $\bar{K}$ The centered Kernel Matrix
4) Z' the projection matrix
5) $W = V^T$ The Principal Components matrix

Our Known Equations:
1) $\bar{Z} = Z'W$
2) $X = U\Sigma V^T$ The formula for $SVD(X)$
3) $w_c = \sigma_c^{-1} X^T u_c$ where σ_c is the c^{th} singular value and u_c is the c^{th} column of U
4) $\bar{K} = \bar{Z}\bar{Z}^T$
5) $\lambda_c = \frac{\sigma^2}{n-1}$ where n is the number of data points and λ_c corresponds to the c^{th} eigenvalue of $\bar{K}$
6) eigenvalues of AB = eigenvalues of BA, for two given matrices A and B
7) eigenvalues of $\bar{Z}^T\bar{Z}$ = eigenvalues of $\bar{Z}\bar{Z}^T$
8) $\bar{K}_{ij} = \kappa(x_i, x_j) - \frac{1}{n}\Sigma_{l=1}^{n}\kappa(x_i, x_l) - \frac{1}{n}\Sigma_{l=1}^{n}\kappa(x_j, x_l) + \frac{1}{n^2}\Sigma_{k=1}^{n}\Sigma_{l=1}^{n}\kappa(x_l, x_k)$, The formula for centering our K

By applying PCA to $\bar{Z}$, we aim to find $\bar{Z} = Z'W$ However we do not have $\bar{Z}$ or ϕ. What we do have is the Kernel matrix $\bar{K} = \bar{Z}\bar{Z}^T$. Using the properties of the eigenvectors of $\bar{K}$ we have:

$$\bar{K}u_c = \lambda u_c$$

Substituting $\bar{K}$ with $\bar{Z}\bar{Z}^T$:

$$\bar{Z}\bar{Z}^T u_c = \lambda u_c$$

having $w_c = \sigma_c^{-1}\bar{Z}^T u_c$ we can multiplying both sides by $\bar{Z}$:

$$\bar{Z}w_c = \sigma_c^{-1}\bar{Z}\bar{Z}^T u_c = \sigma_c^{-1}\bar{K}u_c$$

Note that σ_c^{-1} is a scalar and can be moved. We then replace $\bar{Z}$ with $Z'W$:

$$Z'Ww_c = \sigma_c^{-1}\bar{K}u_c$$

Now we know that W is orthanormal and w_c is its c^{th} column. This means that Ww_c is a column matrix, where every element is the product of two orthogonal vectors, which is 0, except for the c^{th} element which is equal to $||w_c||^2$. As W is orthonormal, w_c is a normal vector and $||w_c||^2 = 1$. Therefor Ww_c is a column matrix in the form

$$\begin{bmatrix} 0 \\ \ddots \\ 1 \\ \ddots \\ 0 \end{bmatrix}$$

where all elements are 0 except for the i^{th} which is equal to 1. We denote this column matrix I_c. The equation then becomes:

$$Z'I_c = \sigma_c^{-1}\bar{K}u_c$$

where $Z'I_c$ is just the c^{th} column of our Z', denoted as z'_c.

$$z'_c = \sigma_c^{-1}\bar{K}u_c$$

We also know that $\lambda_c = \frac{\sigma^2}{numberofdatapoints-1}$ where λ_c is the c^{th} eigenvalue of $\bar{Z}^T\bar{Z}$ which is the same for $\bar{Z}\bar{Z}^T$ which is our $\bar{K}$. We can now calculate σ_c as follows:

$$\sigma_c = \sqrt{(\lambda_c)(n-1)}$$

where λ_c is the c^{th} eigenvalue of $\bar{K}$ and n is the number of our data points.

Substituting this result in the z'_c formula we have

$$z'_c = \sigma_c^{-1}\bar{K}u_c = \frac{\bar{K}u_c}{\sigma_c} = \frac{\bar{K}u_c}{\sqrt{(\lambda_c)(n-1)}}$$

We can then compute Z':

$$Z' = \bar{K}Udiag(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_n})$$

This is particularly computable because:
1) We can construct $\bar{K}$ using our kernel function to construct K, and then centering K using equation 8.
2) We can compute U as it is the matrix of eigenvectors of our $\bar{K}$.
3) We can compute λ_cs as they are the eigenvalues of the matrix $\bar{K}$
4) We can compute σ_c and then the diagonal matrix, as we can compute λ_c

We have successfully used Principal Components to calculate Z' which is the projection of our data, i.e. the goal of the task.

Task 2

Questions to answer:
A) What kind of projection can be achieved with the Homogeneous Polynomial kernel and with the Gaussian kernel?
Answer: Polynomial The projection achieved with the Polynomial kernel corresponds to a higher-dimensional feature space where the data is mapped into a space of polynomial combinations of the input features up to the specified degree. This way we can capture complex relationships between features.
Answer: Gaussian The projection achieved with the Gaussian kernel corresponds to a feature space that is infinite-dimensional. The data is mapped into a space where each data point has an infinite number of features, and the influence of each feature decreases with its distance from the reference point. The Gaussian kernel is capable of capturing complex patterns, when the decision boundary is highly non-linear.

B) How can one relate the kernel width (σ) to the data available?
Answer: The parameter σ in the Gaussian kernel controls the width of the kernel and influences how sensitive the kernel is to the distance between data points. Specifically, σ determines the scale of the Gaussian function used to calculate the similarity between data points.The larger the value of σ, the smoother and more slowly the similarity decreases with increasing distance..

C) What is the influence of the degree (d) of a Homogeneous Polynomial kernel?
Answer: the degree d in a Polynomial kernel controls the flexibility and complexity of the decision boundary. Larger values of d lead to higher-degree polynomial terms in the kernel expansion, the decision boundary becomes more complex and capable of capturing more complex relationships between features and the model becomes more expressive but is also more prone to overfitting.
For small values of d, the model is less prone to overfitting but may struggle to capture complex patterns in the data.