

1 Motivation: Dimensionality reduction

In this problem sheet we explore the motivation for general dimensionality reduction in machine learning and derive from first principles why projection on the first eigenvectors of the covariance matrix of the data has some favorable properties. A deeper understanding on the advantages of PCA and other dimensionality reduction methods is conveyed in the homework.

In general, we assume the following scenario: Suppose we are given n points $\mathbf{x}_1, \dots, \mathbf{x}_n$ in \mathbb{R}^d and the dimension of the feature vectors is d (very big, like 10^3). By dimensionality reduction, we refer to a mapping $\psi : \mathbb{R}^d \mapsto \mathbb{R}^k$ that maps vectors from \mathbb{R}^d to \mathbb{R}^k with $k \ll d$.

- (a) (Motivation) Given n feature vectors of d dimensions, in which regimes of n, d and why would you want to reduce the dimensionality in practical machine learning applications? Think about the concept of regularization studied extensively in the past few weeks.
- (b) (Computational aspect) Revisit this in the context of linear regression. What is the computational complexity of performing a linear regression of n data points in d dimensions with $n > d$ (say by solving the normal equations when $\mathbf{X}^\top \mathbf{X}$ is invertible)? If the projection was given to you for free, approximately how many operations would you save if you reduced the dimension from $d = 10^3$ to $d = 10$?

2 Derivation of PCA

PCA is often used as a tool in data visualization and reduction of computation load and noise. PCA can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after removing the mean from the data matrix for each feature/column. In this question we will derive PCA. There are two equivalent perspectives to understand PCA. PCA aims to either find

- the directions of maximum variance, or
- the projections of minimum reconstruction error

given a dataset.

- (a) In the first part, we will derive PCA from the perspective of *maximum variance*. You want the line such that projecting your data onto this line will retain the maximum amount of information, i.e., variance. Assuming that the feature matrix \mathbf{X} has zero mean across each of its

column, we can formulate the optimization problem as

$$\max_{\mathbf{w}: \|\mathbf{w}\|_2=1} \sum_{i=1}^n \left(\mathbf{x}_i^\top \mathbf{w} \right)^2 = \max_{\mathbf{w}: \|\mathbf{w}\|_2=1} \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} \quad (1)$$

where \mathbf{x}_i is the feature of i th sample, i.e., the i th row of the matrix \mathbf{X} .

Show that the maximizer for this problem is equal to the eigenvector \mathbf{v}_1 that corresponds to the largest eigenvalue λ_1 of matrix $\mathbf{X}^\top \mathbf{X}$. Also show that optimal value of this problem is equal to λ_1 .

- (b) Let us call the solution of the first part \mathbf{w}_1 . Next, we will use a *greedy procedure* to find the i th component of PCA by doing the following optimization

$$\begin{aligned} & \text{maximize} && \mathbf{w}_i^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}_i \\ & \text{subject to} && \mathbf{w}_i^\top \mathbf{w}_i = 1 \\ & && \mathbf{w}_i^\top \mathbf{w}_j = 0 \quad \forall j < i, \end{aligned} \quad (2)$$

where $\mathbf{w}_j, j < i$ are defined recursively using the same maximization procedure above. Show that the maximizer for this problem is equal to the eigenvector \mathbf{v}_i that corresponds to the i th eigenvalue λ_i of matrix $\mathbf{X}^\top \mathbf{X}$. Also show that optimal value of this problem is equal to λ_i .

- (c) Show that the previous *greedy procedure* finds the global maximum, namely for any $k < d$, $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k$ is the solution of the following maximization problem

$$\begin{aligned} & \text{maximize} && \sum_{i=1}^k \mathbf{w}_i^\top \mathbf{X}^\top \mathbf{X} \mathbf{w}_i \\ & \text{subject to} && \mathbf{w}_i^\top \mathbf{w}_i = 1 \\ & && \mathbf{w}_i^\top \mathbf{w}_j = 0 \quad \forall i \neq j. \end{aligned} \quad (3)$$

- (d) Finally, we will show that PCA from the perspective of *minimizing the reconstruction error* from projection, i.e., minimizing the perpendicular distance between the principle component subspace and the data points. Let's say we want to find the best 1D space that minimizes the reconstruction error. The projection of the feature vector \mathbf{x} onto the subspace spanned by a unit vector \mathbf{w} is

$$P_{\mathbf{w}}(\mathbf{x}) = \mathbf{w} \left(\mathbf{x}^\top \mathbf{w} \right). \quad (4)$$

Show that the minimizer \mathbf{w} for the reconstruction error

$$\min_{\mathbf{w}: \|\mathbf{w}\|=1} \sum_{i=1}^n \|\mathbf{x}_i - P_{\mathbf{w}}(\mathbf{x}_i)\|_2^2 \quad (5)$$

is as same as the \mathbf{w} in Equation (1).