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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 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 drive PCA. There are two equivalent perspectives to understand PCA. PCA aims to either find

1. the directions of maximum variance, or
2. the projections of minimum reconstruction error

given a dataset.

(a) Gaussian MLE: Assume our data matrix $X \in \mathbb{R}^{n \times d}$ is mean centered. What is the mean and variance of the maximum likelihood estimate for a Gaussian distribution fitting our dataset?

(b) Given this Gaussian, how may we construct a k dimensional basis to project our data?

(c) Maximum Projected Variance: We would like the vector w such that projecting your data onto w will retain the maximum amount of information, i.e., variance. We can formulate the optimization problem as

$$\max_{w:\|w\|_2=1} \frac{1}{n}\sum_{i=1}^{n}\left(x_i^\top w\right)^2 = \max_{w:\|w\|_2=1} \frac{1}{n} w^\top X^\top X w \tag{1}$$

where x_i is the feature of ith sample, i.e., the ith row of the matrix X.

Show that the maximizer for this problem is equal to the eigenvector v_1 that corresponds to the largest eigenvalue λ_1 of matrix $X^\top X$. Also show that optimal value of this problem is equal to λ_1/n.

(d) Let us call the solution of the above part w_1. Next, we will use a *greedy procedure* to find the ith component of PCA by doing the following optimization

$$\begin{array}{ll} \text{maximize} & w_i^\top X^\top X w_i \\ \text{subject to} & w_i^\top w_i = 1 \\ & w_i^\top w_j = 0 \quad \forall j < i, \end{array} \tag{2}$$

where 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 v_i that corresponds to the ith eigenvalue λ_i of matrix $X^\top X$. Also show that optimal value of this problem is equal to λ_i.

(e) Show that the previous *greedy procedure* finds the global maximum, namely for any $k < d$, $w_1, w_2, \ldots, w_k$ is the solution of the following maximization problem

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{k} w_i^\top X^\top X w_i \\ \text{subject to} & w_i^\top w_i = 1 \\ & w_i^\top w_j = 0 \quad \forall i \neq j. \end{array} \tag{3}$$

(f) Minimizing Reconstruction Error: Our final perspective on PCA is 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 x onto the subspace spanned by a unit vector w is

$$P_w(x) = w\left(x^\top w\right). \tag{4}$$

Show that the minimizer w for the reconstruction error

$$\min_{w:|w|=1} \sum_{i=1}^{n} \|x_i - P_w(x_i)\|_2^2 \tag{5}$$

is as same as the w in Equation (1).