Fifth practice exercises in Machine learning 1 – 2024 – Paper 1

1 Principal component analysis (September)

Suppose we have a data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of D-dimensional vectors, which have a zero mean for each dimension. Assume we perform a complete eigenvalue decomposition of the empirical covariance matrix $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$. You are interested in only a single projection of your data such that the variance of this projection is maximized. Let $\mathbf{u}_i$ be the direction vector of a particular projection. Assume that $\mathbf{u}_i^T\mathbf{u}_i = 1$.

(a) What is the projection z_{ni} of a given point $\mathbf{x}_n$ under the particular vector $\mathbf{u}_i$?

(b) What is the empirical mean of the projection z_i across all points $\mathbf{x}_n$?

(c) What is the empirical variance of the projection z_i? Provide your answer in terms of the empirical covariance matrix $\mathbf{S}$

(d) Replace $\mathbf{S}$ with its eigenvalue decomposition and simplify the aforementioned expression. What is the variance now?

with $\mathbf{e}_i$ to be a vector with zeros except the position with index i.

(e) Suppose that you are interested in reducing the dimensionality from D to K, such that 99% of the variance is maintained. How can you select an appropriate K?