Lecture 3

3 Introduction to (linear) dimensionality reduction

We now shift to unsupervised learning via linear dimensionality reduction. As we shall see, this is (at least tangentially) related to our dot product kernel $k(x, x') = \langle x, x' \rangle$ defined previously.

3.1 Principal component analysis

This section is essentially [2, Chapter 1.1], with minor modificaitons. See also [3] for another reference.

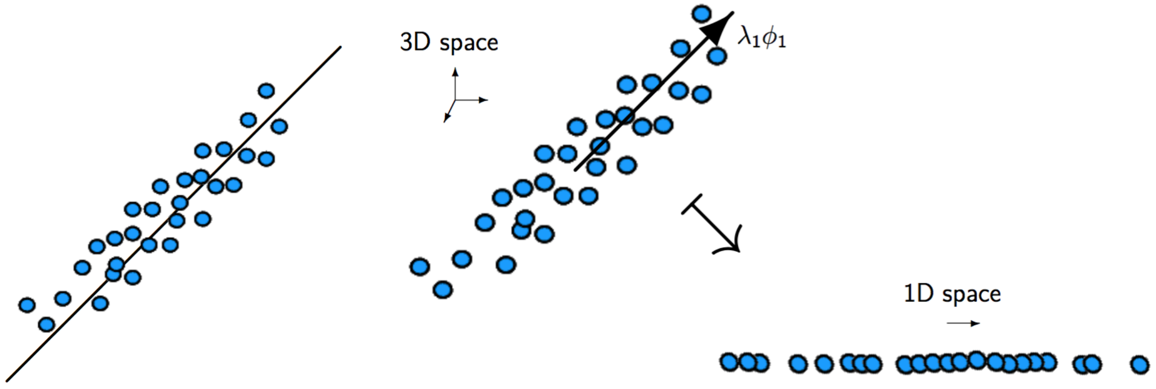
Once again suppose $\mathcal{X} = \mathbb{R}^p$ and that we have n samples $\mathcal{X}_n = \{x_i\}_{i \leq n}$ (but no labels, or we are ignoring the labels). If the dimension p is large, it is natural to try to reduce it to the "intrinsic" dimension of $\mathcal{X}$. This is often useful for visualizing the data in two or three dimensions; additionally, the coordinates in the reduced dimensional space can (sometimes) give insight into the underlying variables that generated the data (even if the reduced dimension is larger than three). Dimension reduction can be done by linear projections, or a nonlinear map. In this section we focus on the most common linear projection, which is principal component analysis (PCA) and which goes back to a paper written by Karl Pearson in 1901 [4].

Suppose we want to linearly project $\mathcal{X}_n$ to $d < p$ dimensions. Two possible ways of doing this are:

1. Finding the d-dimensional affine subspace for which the projections of $\mathcal{X}_n$ best approximate the original points in a suitable sense.

2. Finding the d-dimensional projection the preserves as much variance of the data as possible.

As we shall see, these two goals are in fact equivalent. Figure 2 illustrates the second goal (and hence the first as well).



(a) Data set with line along direction of maximum variance.

(b) One dimensional projection of data onto the direction of maximum variance.

Figure 2: PCA illustration: Projecting data onto it its first principal component.

Before examining either, define the sample mean as

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the sample covariance as:

$$\Sigma_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_n)(x_i - \mu_n)^T$$

where $x \in \mathbb{R}^p$ is considered a $p \times 1$ column vector and x^T is its transpose, a $1 \times p$ vector. Notice that if $X_n = [x_1 \cdots x_n]$ is the $p \times n$ matrix whose columns are the data samples x_i, then

$$\Sigma_n = \frac{1}{n-1}(X_n - \mu_n\mathbf{1}^T)(X_n - \mu_n\mathbf{1}^T)^T$$

where $\mathbf{1}$ is the $p \times 1$ vector consisting of all ones. In particular, if $\mathcal{X}_n$ has zero mean, then $\Sigma_n = (1/(n-1))X_nX_n^T$. Notice that the dot product kernel, defined in Section 2.1, was given by $K = X_n^TX_n$. Up to the scaling factor $1/n-1$, these two matrices are related through the singular value decomposition (SVD). We will come back to this towards the end of this section.

We also remark that if data points in $\mathcal{X}_n$ are independently sampled from a distribution, then μ_n and Σ_n are unbiased estimators for the mean an covariance of the distribution, respectively.

3.1.1 PCA is the best d-dimensional affine fit:

We want to approximate each x_i by:

$$x_i \approx \mu + \sum_{j=1}^{d} \beta_i[j] v_j \tag{4}$$

where $\mathcal{V} = \{v_i\}_{i \leq d}$ is an orthonormal basis (ONB) for the d-dimensional subspace, $\mu \in \mathbb{R}^p$ is the translation, and $\beta_i \in \mathbb{R}^d$ are the coefficients of x_i in $\mathcal{V}$. Let $V = [v_1, \ldots, v_d]$ be the $p \times d$ matrix of $\mathcal{V}$. Then (4) can be rewritten as:

$$x_i \approx \mu + V\beta_i$$

with $V^T V = I$ since $\mathcal{V}$ is an ONB.

We measure the fit of (4) in terms of the squared ℓ^2 error, meaning we want to solve:

$$\min_{\substack{\mu, V, \beta_i \\ V^T V = I}} \sum_{i=1}^{n} \|x_i - (\mu + V\beta_i)\|^2. \tag{5}$$

We start by solving for the optimal value μ^* of μ. Recall that the minimum of a multivariate function occurs where all partial derivatives are zero. Taking the gradient with respect to the μ variables, we get:

$$\begin{aligned} \nabla_\mu \sum_{i=1}^{n} \|x_i - (\mu + V\beta_i)\|^2 = 0 &\Leftrightarrow \sum_{i=1}^{n} (x_i - (\mu + V\beta_i)) = 0 \\ &\Leftrightarrow \left(\sum_{i=1}^{n} x_i \right) - n\mu - V \left(\sum_{i=1}^{n} \beta_i \right) = 0. \end{aligned}$$

But without loss of generality we can assume $\sum_{i=1}^{n} \beta_i = 0$ (check this!), and so we have:

$$\mu^* = \frac{1}{n} \sum_{i=1}^{n} x_i = \mu_n.$$

Thus we have reduced (5) to solving:

$$\min_{\substack{V, \beta_i \\ V^T V = I}} \sum_{i=1}^{n} \|x_i - (\mu_n + V\beta_i)\|^2.$$

Now let us solve for $\{\beta_i\}_{i \leq n}$. Notice that the minimization decouples over these coefficients, meaning that we can solve for each β_i separately:

$$\min_{\beta_i} \|x_i - \mu_n - V\beta_i\|^2 = \min_{\beta_i} \left\| x_i - \mu_n - \sum_{j=1}^{d} \beta_i[j] v_j \right\|^2 .$$

Since $\mathcal{V}$ is an ONB, it is easy to see the solution is:

$$\beta_i^*[j] = v_j^T (x_i - \mu_n) \Rightarrow \beta_i = V^T (x_i - \mu_n).$$

Thus (5) is now reduced to:

$$\min_{V^T V = I} \sum_{i=1}^{n} \|(x_i - \mu_n) - VV^T (x_i - \mu_n)\|^2.$$

By using $\|x\|^2 = \langle x, x \rangle$ and the fact that $V^T V = I$, we have (check this!):

$$\|(x_i - \mu_n) - VV^T(x_i - \mu_n)\|^2 = (x_i - \mu_n)^T (x_i - \mu_n) - (x_i - \mu_n)^T VV^T (x_i - \mu_n).$$

Since $(x_i - \mu_n)^T (x_i - \mu_n)$ does not depend on V, solving (5) is equivalent to:

$$\max_{V^T V = I} \sum_{i=1}^{n} (x_i - \mu_n)^T VV^T (x_i - \mu_n).$$

Using properties of the trace of a matrix (check this!), we can arrive at:

$$\sum_{i=1}^{n} (x_i - \mu_n)^T VV^T (x_i - \mu_n) = (n-1) \mathrm{Tr}(V^T \Sigma_n V).$$

Thus we have show that (5) is equivalent to:

$$\max_{V^T V = I} \mathrm{Tr}(V^T \Sigma_n V).$$

Exercises

Exercise 5. Three times in the above proof I wrote “check this!” Go back through the proof and prove the first statement.

Exercise 6. Three times in the above proof I wrote “check this!” Go back through the proof and prove the second statement.

Exercise 7. Three times in the above proof I wrote “check this!” Go back through the proof and prove the third statement.

3.1.2 PCA preserveres the most variance:

We now switch to the second goal, which is finding the d-dimensional projection that preserves the most variance of the data. This means we want to find an ONB $\mathcal{V} = \{v_1, \ldots, v_d\}$ such that the projection of $\mathcal{X}_n$ onto $\mathcal{V}$ has the most variance. The projection of x_i on to $\mathcal{V}$ is given by $V^T x_i$, and so we want to maximize the variance of the points $\{V^T x_i\}_{i \leq n}$.

Recall that the total variance of $\mathcal{X}_n = \{x_i\}_{i \leq n}$ is:

$$\text{Total Variance}(\mathcal{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \mu_n\|^2 = \frac{1}{n} \sum_{i=1}^{n} \left\| x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right\|^2 .$$

Thus, if we want to maximize the variance of $\{V^T x_i\}_{i \leq n}$, we want to solve:

$$\max_{V^T V = I} \sum_{i=1}^{n} \left\| V^T x_i - \frac{1}{n} \sum_{j=1}^{n} V^T x_j \right\|^2 .$$

But note that:

$$\sum_{i=1}^{n} \left\| V^T x_i - \frac{1}{n} \sum_{j=1}^{n} V^T x_j \right\|^2 = \sum_{i=1}^{n} \|V^T (x_i - \mu_n)\|^2 = (n-1) \mathrm{Tr}(V^T \Sigma_n V).$$

Thus the two goals are in fact the same!

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[2] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. MIT course *Topics in Mathematics of Data Science*, 2015.

[3] Jon Shlens. A tutorial on principal component analysis. arXiv:1404.1100, 2014.

[4] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901.

[5] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):507–536, 1967.

[6] J. Baik, G. Ben-Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[7] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.