Chapter 17

Principal Components Analysis

In this last chapter, we introduced techniques for feature design. We focused primarily on manual heuristics that one can use to generate and select features. The drawback of manual feature design is that it involves substantial engineering work. In this chapter we will introduce the idea of learning representations, and we will start with the simplest approach, principal components analysis (PCA).

17.1 Projections to a Subspace

PCA relies on the idea of projecting our initial input features to a low-dimensional subspace. Here, we will assume that we already have some feature representation $\mathbf{x} \in \mathbb{R}^m$—which may have been generated using the feature design techniques discussed in the previous chapter. The goal in PCA is to find a d-dimensional subspace—where $d < m$—such that projections of the initial features into this subspace preserve only the most useful information. This idea is often referred to as *dimensionality reduction*.

To formalize the idea of projecting to a subspace, we must define the following quantities. First, we assume that we have an orthonormal projection matrix $\mathbf{U} \in \mathbb{R}^{m \times d}$; this projection matrix defines a map from the m-dimensional space $\mathbb{R}^m$ to the lower dimensional space $\mathbb{R}^d \subset \mathbb{R}^m$. Note that this matrix has orthonormal columns, which means that all the columns are orthogonal and that $\mathbf{U}^\top \mathbf{U} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We also define an origin vector $\boldsymbol{\mu}$, which defines the origin or center of the newly defined subspace. Typically, we assume that

$$\boldsymbol{\mu} = \frac{1}{\mathcal{D}} \sum_{\mathbf{x} \in \mathcal{D}} \mathbf{x}. \tag{17.1}$$

That is, we assume that the origin of the subspace is simply defined as the mean of the points in the dataset. This is equivalent to assuming that we have centered or normalized all our points before projecting them to the subspace.

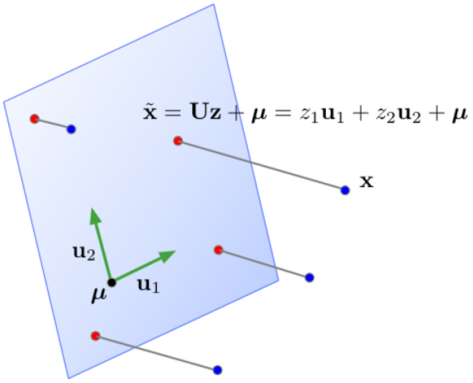
$\tilde{\mathbf{x}} = \mathbf{U}\mathbf{z} + \boldsymbol{\mu} = z_1\mathbf{u}_1 + z_2\mathbf{u}_2 + \boldsymbol{\mu}$
$\mathbf{x}$
$\mathbf{u}_2$
$\mathbf{u}_1$
$\boldsymbol{\mu}$

Figure 17.1: Illustration of a projection to a subspace.

Based on these quantities, we can define a projection of a point $\mathbf{x}$ on the subspace as follows:

$$\mathbf{z} = \mathbf{U}^\top(\mathbf{x} - \boldsymbol{\mu}). \tag{17.2}$$

The vector $\mathbf{z} \in \mathbb{R}^d$ is the low-dimensional projection of $\mathbf{x} \in \mathbb{R}^m$, and $\mathbf{z}$ is often called the *code* representing $\mathbf{x}$. Another important quantity is the *reconstruction* of $\mathbf{x}$ based on the code $\mathbf{z}$, which is defined as

$$\tilde{\mathbf{x}} = \mathbf{U}\mathbf{z} + \boldsymbol{\mu}. \tag{17.3}$$

The reconstruction $\tilde{\mathbf{x}}$ is the best approximation we can get of $\mathbf{x}$ using only the lower-dimensional information contained in $\mathbf{z}$. Note that information will tend to be lost when the low-dimensional projection is applied, so we will typically have that $\tilde{\mathbf{x}} \neq \mathbf{x}$ in general. Our hope, however, is that we can have that $\tilde{\mathbf{x}} \approx \mathbf{x}$. The idea of projecting to a subspace is illustrated in Figure 17.1.

17.2 What Makes a Good Projection?

The goal of PCA is not just to project data to an arbitrary subspace; the goal is to project data to a subspace that retains the most useful information about the initial features $\mathbf{x}$. We will introduce two ways to define the “goodness” of a projection, and then we will show that these two perspectives are actually equivalent.

Reconstruction error

The first notion of “goodness” that we can use is the *reconstruction error*. This measures the distance between an original point $\mathbf{x}$ and the reconstruction of that point $\tilde{\mathbf{x}}$ based on its low-dimensional code $\mathbf{z}$:

$$\|\tilde{\mathbf{x}} - \mathbf{x}\|^2 = \|\mathbf{U}\mathbf{z} + \boldsymbol{\mu} - \mathbf{x}\|^2 \tag{17.4}$$

Retained variance

Another measure of "goodness" is the extent to which the low-dimensional codes $\mathbf{z}$ capture the variance in the data. Ideally, we want low dimensional codes that capture variance in the data. For example, in the worst case, when our codes capture no variance, we end up with constant codes that contain no information about the original data. We can quantify the amount of variance retained as

$$\sum_{j=0}^{d-1} \mathrm{Var}(\mathbf{z}[j]) = \frac{1}{|\mathcal{D}|} \sum_{j=0}^{d-1} \sum_{i=0}^{|\mathcal{D}|-1} (\mathbf{z}_i[j] - \bar{z}_j)^2 \tag{17.5}$$

$$= \frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|-1} \|\mathbf{z}_i - \bar{\mathbf{z}}\|^2 \tag{17.6}$$

$$= \frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|-1} \|\mathbf{z}_i\|^2. \tag{17.7}$$

Here, we use

$$\bar{\mathbf{z}} = \frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|-1} \mathbf{z}_i \tag{17.8}$$

to denote the mean of the low-dimensional projections of the data points. We leave it as an exercise to the reader to show that $\bar{\mathbf{z}} = 0$.

Equivalence between reconstruction error and retained variance

An important fact underlying PCA is that the objectives of minimizing the reconstruction and maximizing the retained variance are equivalent. In particular, we can show that

$$\frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|-1} \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 = -\frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|-1} \|\mathbf{z}_i\|^2 + C, \tag{17.9}$$

where C is a constant that is independent from the projection used to generate the $\mathbf{z}_i$ This equality implies that

$$\min_{\mathbf{U}} \frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|-1} \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 = \max_{\mathbf{U}} \frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|-1} \|\mathbf{z}_i\|^2, \tag{17.10}$$

which means that optimizing the projection matrix $\mathbf{U}$ according to either objective is equivalent.

To show this equivalence, first we can note that

$$\|\tilde{\mathbf{x}}_i - \boldsymbol{\mu}\| = \|\mathbf{U}\mathbf{z}\| = \|\mathbf{z}\| \tag{17.11}$$

for any code $\mathbf{z}$, since $\mathbf{U}$ is matrix with orthonormal columns. This equality in Equation 17.9 then holds as a consequence of the generalized version of the Pythagorean Theorem, which states that

$$\|\mathbf{q}+\mathbf{v}\|^2 = \|\mathbf{q}\|^2 + \|\mathbf{v}\|^2 \tag{17.12}$$

if the two vectors $\mathbf{q}$ and $\mathbf{v}$ are orthogonal. In our case, we can note that the vector $\tilde{\mathbf{x}}_i - \boldsymbol{\mu}$ is orthogonal to the vector $\mathbf{x}_i - \tilde{\mathbf{x}}_i$, since $(\mathbf{x}_i - \tilde{\mathbf{x}}_i)^\top(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}) = 0$ (which can be verified as an exercise). This allows us to apply the generalized Pythagorean Theorem

$$\|(\mathbf{x}_i - \tilde{\mathbf{x}}_i) + (\tilde{\mathbf{x}}_i - \boldsymbol{\mu})\|^2 = \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 + \|\tilde{\mathbf{x}}_i - \boldsymbol{\mu}\|^2 \tag{17.13}$$

$$\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 = \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 + \|\tilde{\mathbf{x}}_i - \boldsymbol{\mu}\|^2, \tag{17.14}$$

and since $\|\mathbf{x}_i - \boldsymbol{\mu}\|^2$ is a constant that does not depend on the projection matrix $\mathbf{U}$ we can say re-write this as

$$\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 = -\|\tilde{\mathbf{x}}_i - \boldsymbol{\mu}\|^2 + C \tag{17.15}$$

$$\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 = -\|\mathbf{z}_i\|^2 + C, \tag{17.16}$$

which implies Equation 17.9 when averaging over the datapoints.

17.3 Principal Components Analysis

The goal of PCA is thus to find a projection matrix $\mathbf{U}$ that maximizes the retained variance (or equivalently that minimizes the reconstruction error). To do so, the PCA method defines the projection matrix $\mathbf{U}$ based on the eigendecomposition of the empirical covariance matrix. We describe this idea in detail and justify it below.

17.3.1 Defining PCA

First, we can define the empirical covariance matrix as

$$\boldsymbol{\Sigma} = \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top. \tag{17.17}$$

The eigendecomposition of this matrix is given by

$$\boldsymbol{\Sigma} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top, \tag{17.18}$$

where $\mathbf{Q}$ is an orthonormal matrix containing the eigenvectors as columns and Λ is a matrix containing the eigenvalues $\lambda_0, ..., \lambda_{m-1}$ on the diagonal. We assume that the eigenvalues are ordered in descending order. Note that empirical covariance matrices are known to be positive semi-definite, which implies that $\lambda_j \geq 0$ for all the eigenvalues.

The key idea in the PCA approach is that we choose the projection matrix $\mathbf{U}$ to be the first d columns of $\mathbf{Q}$. In other words, we project our data based on the d principal eigenvectors of the covariance matrix.

17.3.2 Deriving the PCA approach

The PCA approach provably maximizes the retained variance (and thus minimizes the reconstruction error). We will show the full derivation for the case where $d = 1$, and the generalization of this result to $d > 1$ is known as the Courant-Fischer Theorem. Suppose we set $d = 1$ and define the projected representations as

$$z = \mathbf{u}^\top(\mathbf{x} - \boldsymbol{\mu}). \tag{17.19}$$

(Note that in this special one-dimensional case the projection matrix is just a vector). Let us consider the retained variance for these projected representations. We have that

$$\begin{aligned}
\frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1}(z_i)^2 &= \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1}(\mathbf{u}^\top(\mathbf{x}_i - \boldsymbol{\mu}))^2 && (17.20)\\
&= \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1}\mathbf{u}^\top(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top\mathbf{u} && (17.21)\\
&= \mathbf{u}^\top\frac{1}{|\mathcal{D}|}\left(\sum_{i=0}^{|\mathcal{D}|-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top\right)\mathbf{u} && (17.22)\\
&= \mathbf{u}^\top\boldsymbol{\Sigma}\mathbf{u} && (17.23)\\
&= \mathbf{u}^\top\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top\mathbf{u}. && (17.24)
\end{aligned}$$

Now, we can define a new vector $\mathbf{a} = \mathbf{Q}^\top\mathbf{u}$ and we have that the maximization

$$\max_{\mathbf{u}} \mathbf{u}^\top\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top\mathbf{u} \tag{17.25}$$

is equivalent to the problem

$$\max_{\mathbf{a}} \mathbf{a}^\top\boldsymbol{\Lambda}\mathbf{a} \tag{17.26}$$

with $\mathbf{u} = \mathbf{Q}\mathbf{a}$. To solve this maximization problem we can see that

$$\max_{\mathbf{a}} \mathbf{a}^\top\boldsymbol{\Lambda}\mathbf{a} = \sum_{j=0}^{m-1}(\mathbf{a}[j])^2\lambda. \tag{17.27}$$

Finally, we can note that $\|\mathbf{a}\|^2 = 1$ since $\mathbf{a}$ is an orthonormal transformation of a unit vector $\mathbf{u}$. And, we can see by inspection that the optimal solution to Equation 17.27 will occur when we set $\mathbf{a}[0] = \pm 1$ and $\mathbf{a}[j] = 0, j = 1, ..., m-1$, since this puts all the weight of the sum on the largest eigenvalue λ_0.

Thus, we can see that the optimal projection for retaining the maximum amount of variance is when $\mathbf{a} = [1, 0, ..., 0]^\top$, which implies that $\mathbf{u} = \mathbf{Q}\mathbf{a}$ corresponds to the fist column of $\mathbf{Q}$, i.e., the principal eigenvector.