PCA gradient derivation

December 14, 2021

1 Gradient derivation

We start with the objective function for PCA given by

$$F_{\text{PCA}} = ||L(E^Y - E^X)||_F^2$$

where $L = I - \frac{\mathbb{1}\mathbb{1}^T}{N}$ is the centering matrix for N data points. We define the pairwise distance matrix $E^Y = \text{diag}(G_Y)\mathbb{1}^T + \mathbb{1}\text{diag}(G_Y) - 2G_Y$, where $G_Y = Y^TY$ is the Gram matrix of pairwise inner products.

The methods employing gradient descent for PCA currently perform the search over principal component directions [1]. We instead aim to search the space of points Y, so our gradient derivations will be with respect to those points.

$$\begin{aligned}
F_{\text{PCA}} =& ||L(E^Y - E^X)||_F^2 \\
=& \text{tr}\left((L(E^X - E^Y))^T L(E^X - E^Y)\right) \\
=& \text{tr}\left((E^X)^T L^T L E^X\right) - \text{tr}\left((E^X)^T L^T L E^Y\right) - \\
& \text{tr}\left((E^Y)^T L^T L E^X\right) + \text{tr}\left((E^Y)^T L^T L E^Y\right)
\end{aligned}$$

We now replace L^TL by L and take the derivative with respect to Y to obtain

$$d||L(E^Y - E^X)||_F^2 = \text{tr}\left((dE^Y)^T L(E^Y - E^X) + (E^Y - E^X)^T L\, dE^Y)\right)$$

The trace is unaffected by transposes, giving us:

$$d||L(E^Y - E^X)||_F^2 = 2\text{tr}\left((E^Y - E^X)^T L\, dE^Y\right)$$

Notice that L, $(E^Y - E^X)$ and dE^Y are all symmetric, so the trace is unaffected by commutative switches. This gives us

$$d||L(E^Y - E^X)||_F^2 = 2\text{tr}\left(L\left((E^Y - E^X)\, dE^Y\right)\right)$$

Plugging in the definition of the centering matrix $L = I - \frac{\mathbb{1}\mathbb{1}^T}{N}$ and rearranging gives us the resulting expression for the gradient:

$$d||L(E^Y - E^X)||_F^2 = 2\mathrm{tr}\left((E^Y - E^X)\, dE^Y\right) - 2\mathrm{tr}\left(\frac{\mathbb{1}\mathbb{1}^T}{N}(E^Y - E^X)\, dE^Y\right) \quad (1)$$

where the second term is the sum of column means of $(E^Y - E^X)\, dE^Y$. Assuming $dE^Y \neq 0$, we see that F_{PCA} can be minimzed by two methods:

1. Find Y such that $E^Y - E^X = 0$
2. Find dY such that $\mathrm{tr}\left((E^Y - E^X)\, dE^Y\right)$ equals the sum of column means of $(E^Y - E^X)\, dE^Y$

In essence, the first bullet implies that we can minimize PCA's objective function by taking gradient steps in the direction of $E^Y = E^X$ while the second bullet requires that our steps keep $(E^Y - E^X)dE^Y$ at zero mean. This leads us to the following gradient descent algorithm:

1. For epoch t in range T
   (a) For each $y \in Y$, collect the gradients:
   $dY = \alpha \nabla_Y$ towards $E^Y = E^X$
   (b) Re-center dY such that $(E^Y - E^X)dE^Y$ is zero-mean
   (c) Apply the step with $Y_{t+1} = Y_t + dY$

References

[1] Ohad Shamir. "Convergence of stochastic gradient descent for PCA". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 257–265.