ML Reading Group (Fall 2015) **Lecture: 5**

Learning Feature Representations with K-means

Lecturer: Krishnan Srinivasan *Scribe: Jonathon Cai*

1 Today

– Paper Discussion

2 Derivation of $\arg\max$ term

We would like to compute

$$\min_{s^{(i)}} ||Ds^{(i)} - x^{(i)}||_2^2$$

given $D \in \mathbb{R}^{n \times k}$, $x^{(i)} \in \mathbb{R}^n$, and $s^{(i)} \in \mathbb{R}^k$. The j-th column vector of D is denoted by $D^{(j)}$, where $\forall j$, the l_2 norms of $D^{(j)}$ are 1 and $s^{(i)}$ has at most one nonzero coordinate. It turns out that both these assumptions – the restriction on $D^{(j)}$ and the fact that $||s^{(i)}||_0 \leq 1$ – are important for the analysis.

For the moment assume that $s^{(i)}$ has a nonzero coordinate at $r = s_j^{(i)}$, where r is a constant number. Then $Ds^{(i)} = rD^{(j)}$. So we would like to solve the equation:

$$\frac{\partial}{\partial r} ||rD^{(j)} - x^{(i)}||_2^2 = 0$$

Expanding yields

$$\begin{aligned}
\frac{\partial}{\partial r} \sum_k (rd_k - x_k)^2 &= 0 \\
\sum_k 2(rd_k - x_k)d_k &= 0 \\
r \sum_k d_k^2 &= \sum_k x_k d_k \\
r &= \sum_k x_k d_k && \text{normalized } D^{(j)} \text{ assumption} \\
r &= {D^{(j)}}^T x^{(i)}
\end{aligned}$$

Now we will consider the minimum of the norm over all j. Let r_j denote the r corresponding to j. Note that:

$$\begin{aligned}
\sum_k (r_j d_k - x_k)^2 &= \sum_k r_j^2 d_k^2 + x_k^2 - 2 r_j d_k x_k \\
&= r_j^2 + \sum_k x_k^2 - 2 r_j^2 \\
&= C - r_j^2 && C = l_2 \text{ norm squared of } x^{(i)}
\end{aligned}$$

Thus, for $||s^{(i)}||_0 \leq 1$, the norm is minimized when we maximize $|r_j|$, or when $j == \arg\max_j |r_j|$.

Consider the geometric meaning of this statement. ${D^{(j)}}^T \cdot x^{(i)} = |D^{(j)}||x^{(i)}| \cos\theta = K \cos\theta$, where K is a constant. Hence, by maximizing $\cos\theta$, we are trying to minimize θ, or the angle between the projection of $x^{(i)}$ onto the unit sphere and a centroid lying on the sphere.

3 ZCA Whitening

Suppose X is a random vector with mean 0 and the covariance matrix $\mathbb{E}[XX^T]$ is positive definite (and hence invertible); since V is also symmetric, we may decompose it, using the spectral decomposition, into $V = LDL^T$, where L is orthogonal and D represents a diagonal matrix containing the eigenvalues. Let $K = \sqrt{(D+\epsilon)^{-1}} \approx \sqrt{D^{-1}}$. Recall that $K^T = D$ as D is diagonal. Also $K^2 D = I$ by definition.

Now consider the ZCA transformation denoted by the transformation $T(X) = LKL^{-1}X$. Let the transformed variable be $\tilde{X}$ and consider the variance of this term. $\mathbb{E}$ is of course in front of each term.

$$\begin{aligned}
\tilde{X}\tilde{X}^T &= LKL^{-1}LDL^T L^{-T} K L^T \\
&= LL^T \\
&= I
\end{aligned}$$

Thus, ZCA whitening approximately results in spherical variance.