University of Edinburgh

INFR11156: Algorithmic Foundations of Data Science (2019)

Lecture 7: Best-fit Subspaces and Singular Value Decomposition (3)

Partitioning a set of data points into different clusters according to some criteria is a fundamental task in Data Science, and over the years entirely different techniques have been developed for the clustering problem. In today's lecture, we will discuss how singular vector decomposition is applied for clustering data points generated from mixture model. In particular, we will informally view singular vector decomposition as *dimensionality reduction*, which connection allows us to improve the condition from which different Gaussians can be easily separated, i.e., we'll improve the condition of $\Delta = \omega\left(d^{1/4}\right)$ from Lecture 4.

We first recall the mixture model. Let $p_1, \ldots, p_k$ be Gaussian density functions, and $w_1, \ldots, w_k$ are positive real numbers called mixture weighs such that $\sum_{i=1}^{k} w_i = 1$. We define the mixture as the distribution with the probability density function

$$p = w_1 p_1 + w_2 p_2 + \cdots + w_k p_k.$$

The *model fitting problem* is to fit a mixture of k basic densities from n independent, identically distributed samples, each sample drawn according to the same mixture distribution p. For simplicity, we assume that all the basic probability densities are spherical Gaussians. One approach to the model fitting problem is to decompose the problem into two subproblems:

1. Cluster the set of samples into k clusters $C_1, C_2, \ldots, C_k$, where C_i is the set of samples generated according to p_i;

2. Fit a single Gaussian distribution to each cluster of sample points by taking the empirical mean and the empirical standard derivation of the sample points, see the notes of Lecture 4.

Hence, it suffices to study the first problem. Our starting point is the following lemma which shows that the projection of a spherical Gaussian with standard deviation σ remains a spherical Gaussian with variance σ^2.

Lemma 1. *Suppose p is a spherical Gaussian in $\mathbb{R}^d$ with centre μ and variance σ^2. The density of p projected onto a k-dimensional subspace V is a spherical Gaussian with the same standard deviation.*

Proof. Rotate the coordinate system so that V is spanned by the first k coordinate vectors. The Gaussian remains spherical with standard deviation σ although the coordinates of its centre has changed. For a point $x = (x_1, \ldots, x_d)$, we will use the notation $x' = (x_1, x_2, \ldots, x_k)$ and $x'' = (x_{k+1}, x_{k+1}, \ldots, x_d)$. The density of the projected Gaussian at point $(x_1, x_2, \ldots, x_k)$ is

$$c\mathrm{e}^{-\frac{|x'-\mu'|^2}{2\sigma^2}} \int_{x''} \mathrm{e}^{-\frac{|x''-\mu''|^2}{2\sigma^2}} \mathrm{d}x'' = c'\mathrm{e}^{-\frac{|x'-\mu'|^2}{2\sigma^2}}$$

for some parameter c'. This proves the lemma. □

Based on Lemma 1, we would like to project this spherical Gaussian into $\mathbb{R}^k$, and the inter-centre separation remains the same. If this is the case, instead of assuming the inter-centre separation distance is $\Omega(d^{1/4})$, the distance of $\Omega(k^{1/4})$ is sufficient to separate these Gaussians from each other. We will see that the top k singular vectors produced by the SVD span the space of the k centres.

Recall that for a set of points, the best-fit line is the line passing through the origin that maximises the sum of squared lengths of the projections of the points. However, as our data points are drawn from some probability distribution, we'll slightly adjust the definition of the best-fit line as follows.

Definition 2. *If p is a probability density in $\mathbb{R}^d$, then the best-fit line for p is the line in the v_1 direction, where*

$$v_1 = \arg\max_{\|v\|=1} \mathbf{E}\left[(v^\intercal x)^2\right].$$

Lemma 3 (Best-fit line)**.** *Let p be the probability density defined as above such that $\mu \neq 0$. The unique best-fit 1-dimensional subspace is the line passing through μ and the origin. If $\mu = 0$, then any line through the origin is a best-fit line.*

Proof. For an randomly chosen x according to p and a fixed unit length vector v, it holds that

$$\begin{aligned}
\mathbf{E}_{x\sim p}\left[(v^\intercal x)^2\right] &= \mathbf{E}_{x\sim p}\left[(v^\intercal(x-\mu) + v^\intercal\mu)^2\right] \\
&= \mathbf{E}_{x\sim p}\left[(v^\intercal(x-\mu))^2 + 2\left(v^\intercal\mu\right)\left(v^\intercal(x-\mu)\right) + \left(v^\intercal\mu\right)^2\right] \\
&= \mathbf{E}_{x\sim p}\left[(v^\intercal(x-\mu))^2\right] + 2(v^\intercal\mu)\mathbf{E}\left[v^\intercal(x-\mu)\right] + (v^\intercal\mu)^2 \\
&= \mathbf{E}_{x\sim p}\left[(v^\intercal(x-\mu))^2\right] + (v^\intercal\mu)^2 \qquad (1) \\
&= \sigma^2 + \left(v^\intercal\mu\right)^2, \qquad (2)
\end{aligned}$$

where (1) follows from the fact that $\mathbf{E}\left[v^\intercal(x-\mu)\right] = 0$, and (2) follows from the fact that $\mathbf{E}_{x\sim p}\left[(v^\intercal(x-\mu))^2\right]$ is the variance of x in the direction of v.

From (2), we see that the line v maximising $\mathbf{E}_{x\sim p}\left[(v^\intercal x)^2\right]$ is the line v that maximises $(v^\intercal\mu)^2$, which is the case if v is aligned with μ. From (2), we know that, when $\mu = 0$, any line through the origin is a best-fit line. □

Generalising the result above, we study the best-fit k-dimensional subspace for p.

Definition 4. *If p is a probability density function in $\mathbb{R}^d$, then the best-fit k-dimensional subspace V_k is*

$$V_k = \arg\max_{V:\dim(V)=k} \mathbf{E}_{x\sim p}\left[|\mathrm{proj}(x,V)|^2\right],$$

where $\mathrm{proj}(x,V)$ *is the orthogonal projection of x onto V.*

Theorem 5. *If p is a mixture of k spherical Gaussians, then the best-fit k-dimensional subspace contains the centres. In particular, if the means of the Gaussians are linearly independent, then space spanned by them is the unique best-fit k dimensional subspace.*

Proof. Let p be the mixture $w_1p_1 + w_2p_2 + \cdots + w_kp_k$. Let V be any subspace of dimension k or less. Then, it holds that

$$\mathbf{E}_{x\sim p}\left[|\mathrm{proj}(x,V)|^2\right] = \sum_{i=1}^{k} w_i \mathbf{E}_{x\sim p_i}\left[|\mathrm{proj}(x,V)|^2\right].$$

Hence, in order to maximise $\mathbf{E}_{x\sim p}\left[|\mathrm{proj}(x,V)|^2\right]$, it suffices to maximise every $\mathbf{E}_{x\sim p_i}\left[|\mathrm{proj}(x,V)|^2\right]$. Then, by defining V as the span of k centres, the first statement follows directly from Lemma 3. The second statement follows from the fact that the k centres are linearly independent. □

When an infinite set of points drawn from probability density p is available, then the k-dimensional subspace from SVD gives us exactly the space of the centres. In reality, although only a finite number of data points are available, it is intuitively clear that, as the number of samples increases, the set of the sampled points will approximate the probability density better and therefore the SVD subspace of the sampled points will be close to the space spanned by the centres.

Another interesting fact about our analysis above is that, similar to the Johnson-Lindenstrauss Lemma, we use SVD for dimension reduction: every data point in $\mathbb{R}^d$ is embedded in $\mathbb{R}^k$ through SVD. However, for this specific setting the choice of k depends on the inherent structure of data points and can be chosen as a constant for many applications, while for the Johnson-Lindenstrauss Lemma, the target dimension $O(\log n/\varepsilon^2)$ is known to be tight.