University of Edinburgh

INFR11156: Algorithmic Foundations of Data Science (2019)

Lecture 4: High-Dimensional Spaces (3)

We study the Gaussian mixture model, which is formulated by

$$p(x) = \sum_{i=1}^{k} w_i p_i(x).$$

Here every p_i is a Gaussian density function, and w_i is the mixture weight representing the proportion relative to the other Gaussians. The *parameter estimation problem* for a mixture model is the problem that, given access to samples from the overall density p, reconstruct the parameters for the distribution. In today's lecture, we will focus on the case of $k = 2$: we will study the condition under which data points drawn from two Gaussians can be separated, and the method for estimating the mean and variance of each Gaussian.

1 Separating Gaussians

The following Gaussian Annulus Theorem will be used in our analysis:

Theorem 1 (Gaussian Annulus Theorem)**.** *For a d-dimensional spherical Gaussian with unit variance in each direction, for any $\beta \leq \sqrt{d}$, all but at most $3e^{-c\beta^2}$ of the probability mass is within the annulus $\sqrt{d} - \beta \leq \|x\| \leq \sqrt{d} + \beta$, where c is a fixed constant.*

Notice that

$$\mathbf{E}\left[\, \|x\|^2 \,\right] = \sum_{i=1}^{d} \mathbf{E}\left[\, x_i^2 \,\right] = d \cdot \mathbf{E}\left[\, x_i^2 \,\right] = d,$$

so the expected ℓ_2^2-distance of a random point from the centre is d. The Gaussian Annulus Theorem states that the points are highly concentrated. We call $\sqrt{d}$ the radius of the Gaussian.

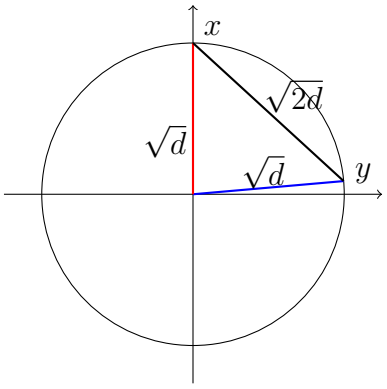

Figure 1: Point y's component along x's direction is $O(1)$ with high probability, which implies that y is nearly perpendicular to x and $\|x - y\| \approx \sqrt{\|x\|^2 + \|y\|^2}$.

Now we are ready to derive the condition for which two Gaussians can be easily separated by some simple algorithm. We first study the locations of two random points drawn from the same Gaussian. We set the parameter β from Theorem 1 as $\beta = O(1)$, and the analysis consists of the following steps:

1. Pick a point x from a unit-variance Gaussian centred at the origin, i.e., point x is chosen according to the density function
$$p(x) = \frac{1}{(2\pi)^{d/2}} \cdot \exp\left(-\frac{\|x\|^2}{2}\right).$$

2. Rotate the coordinate system to make the first axis align with x, i.e.,
$$x = \left(\sqrt{d} \pm \beta, 0, \ldots, 0\right) = \left(\sqrt{d} \pm O(1), 0, \ldots, 0\right) \tag{1}$$

3. Independently pick a second point y from this Gaussian. Since y is almost on the equator from our previous discussion, we can further rotate the coordinate system so that the component of y that is perpendicular to the axis of the North Pole is in the second coordinate. That is,
$$y = \left(O(1), \sqrt{d} \pm O(1), 0, \ldots, 0\right). \tag{2}$$
Combining (1) and (2), we have
$$\|x - y\|^2 = d \pm O\left(\sqrt{d}\right) + d \pm O\left(\sqrt{d}\right)$$
and $\|x - y\| \approx \sqrt{2d}$ with high probability, i.e., x and y are nearly perpendicular with high probability. See Figure 1 for illustration.

Now we study the case where x and y are random points drawn from two different spherical unit-variance Gaussians with centres p and q satisfying $\|p - q\| = \Delta$. We are interested in the condition on Δ under which data points drawn from different Gaussians can be easily separated. Our analysis follows the same technique used before, and is illustrated in Figure 2.

1. Let x be a randomly chosen point from the Gaussian centred at p, and we rotate the coordinate system so that x is at the North Pole.

2. Let z be the North Pole of the ball approximating the Gaussian centred at q.

3. Let y be a randomly chosen point from the Gaussian centred at q. Since $x - p, p - q$ and $q - y$ are nearly mutually perpendicular, it holds that
$$\|x - y\|^2 \approx \Delta^2 + \|x - p\|^2 + \|q - y\|^2 = \Delta^2 + 2d.$$

Hence, in order to separate two Gaussians, we need to ensure that $2d + O(\sqrt{d}) \leq 2d + \Delta^2$, which holds if $\Delta = \omega(d^{1/4})$. We can generalise this argument, and derive the condition on Δ under which n points can be separated from each other with high probability. Similar with our analysis for the Johnson-Lindenstrauss Lemma, we need to apply the union bound for $\Theta(n^2)$ pair of points to ensure that the n points will be separated correctly. Therefore, we need to ensure that the error probability from Theorem 1 satisfies $3 \cdot \mathrm{e}^{-c\beta^2} = O(1/\mathrm{poly}(n))$, which holds if $\beta = \Theta(\sqrt{\log n})$. Therefore we need to include an extra $\Theta(\sqrt{\log n})$ term in the separation distance.

Our analysis shows that under such condition on Δ, we have a simple algorithm from separating points from two Gaussians: (1) Calculate the distances between all pairs of points; (2) Points whose distances are small come from the same Gaussian, while the points whose distances are large come from different Gaussians.

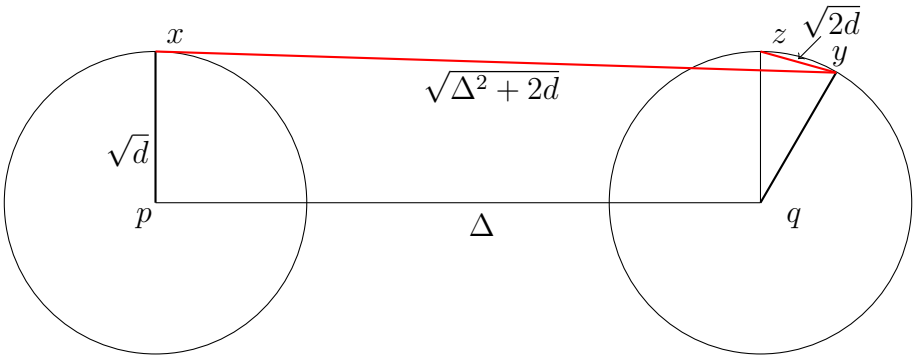

Figure 2: The distance between a pair of random points from two different unit balls approximating the annuli of two Gaussians.

2 Fitting a Spherical Gaussian to Data

After separating the data points from different Gaussians, we need to find the correct parameters for each Gaussian. Formally, let $x_1, \ldots, x_n$ be random samples in d-dimensional space, and we would like to find the spherical Gaussian that best fits the points.

We assume that f is an unknown Gaussian with mean μ and variance σ^2 in each direction. The probability density for picking these points when sampling according to f is given by

$$c \cdot \exp\left(-\frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_n-\mu)^2}{2\sigma^2}\right)$$

for some normalisation factor c.

Lemma 2. *Let $\{x_1, \ldots, x_n\}$ be a set of n d-dimensional points. Then $(x_1-\mu)^2 + \cdots (x_n-\mu)^2$ is minimised when μ is the centroid of the points $x_1, \ldots, x_n$, i.e., $\mu = \frac{1}{n} \cdot (x_1 + \ldots + x_n)$.*

Proof. Let $h(\mu) = (x_1-\mu)^2 + \cdots (x_n-\mu)^2$. Then, it holds that

$$\frac{\mathrm{d}h}{\mathrm{d}\mu} = \sum_{i=1}^{n} (2\mu - 2x_i),$$

which equals to 0 if $\mu = \frac{1}{n} \cdot (x_1 + \ldots + x_n)$. Since the second derivative of h is positive at point $\frac{1}{n} \cdot (x_1 + \ldots + x_n)$, the statement holds. □

Assuming that we know the true mean μ, we use the similar but slightly complicated analysis and obtain the following lemma.

Lemma 3. *The maximum likelihood spherical Gaussian for a set of samples is the Gaussian with centre equal to the sample mean and standard deviation equal to the standard deviation of the sample from the true mean.*

However, although $\tilde{\mu} = \frac{1}{n} \cdot (x_1 + \ldots + x_n)$ is an unbiased estimator of the expected value of the mean, applying $\tilde{\mu}$ directly won't give us an unbiased estimate of the variance, and one should use $\mu^* = \frac{1}{n-1}(x_1 + \ldots + x_n)$ instead when estimating the variance.