Mahalanobis Distance

Uri Shaham

March 4, 2024

1 Issues with Euclidean Distance

Euclidean distance is by far the most popular measure for (dis)similarity of numerical data points. Yet, one should be aware of some of the pitfalls in using it. Here we describe a few

1.1 In a high-dimensional noisy setting, Euclidean distance may represent noise rather than signal

Let $x, y \in \mathbb{R}^d$ be data points, and suppose instead of measuring x, y we measure $\tilde{x} = x+\epsilon$, and $\tilde{y} = y+\delta$, where $\epsilon, \delta \sim \mathcal{N}(0, \sigma^2 I)$. Then

$$\begin{aligned}
\mathbb{E}\left\|\tilde{x}-\tilde{y}\right\|^2 &= \mathbb{E}\left[(\tilde{x}-\tilde{y})^T(\tilde{x}-\tilde{y})\right] \\
&= \mathbb{E}\left[(x+\epsilon-y-\delta)^T(x+\epsilon-y-\delta)\right] \\
&= x^Tx + y^Ty - 2x^Ty + \mathbb{E}\left[2(x^T\epsilon + x^T\delta + y^T\epsilon + y^T\delta) + \epsilon^T\epsilon + \delta^T\delta\right] \\
&= \|x-y\|^2 + \mathbb{E}\left[\epsilon^T\epsilon + \delta^T\delta\right] \\
&= \|x-y\|^2 + 2d\sigma^2.
\end{aligned}$$

From this, we can see that when d is large, measured Euclidean distance is affected by noise (σ^2), rather than signal ($\|x-y\|^2$). It therefore turns out that Euclidean distance is a problematic measure of (dis)similarity in high dimensions, with the exception of immediate neighbors. Diffusion maps, presented in the next lecture is one way to overcome this difficulty.

1.2 Euclidean distance does not consider co(variance)

To answer questions like "are two points 1cm apart close?", we have to consider the layout of the other data points (or equivalently, the covariance structure of the data generating mechanism). See, for example, figure 1. In a similar spirit, questions like "are two points 1cm apart significantly closer than two points 2cm apart?".

Mahalanobis distance, discussed next, adjusts Euclidean distance to take this covariance structure into consideration.

2 Mahalanobis Distance

Proposition 2.1. *Let $x_1, \ldots, x_n$ be datapoints in $\mathbb{R}^m$, and let C be the $m \times m$ data covariance matrix $C = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$. Then C is semi positive-definite. If, in addition, $n \geq m$ and $x_1, \ldots, x_n$ span $\mathbb{R}^m$, then C is strictly positive-definite.*

Proof. Pick $y \in \mathbb{R}^m$, and let $z_i = x_i - \bar{x}$. Then $y^T C y = \frac{1}{n}\sum_{i=1}^n \left(z_i^T y\right)^T \left(z_i^T y\right) \geq 0$, so C is positive semi-definite. Since $x_1, \ldots, x_n$ span $\mathbb{R}^m$, so do $z_1, \ldots z_n$. If y is nonzero, then $y^T C y = 0$ implies that $z_i^T y = 0$ for all i, which is a contradiction, as there is a linear combination $y = \sum_i a_i z_i$, so $y^T y = \sum_i a_i y^T z_i = 0$. □

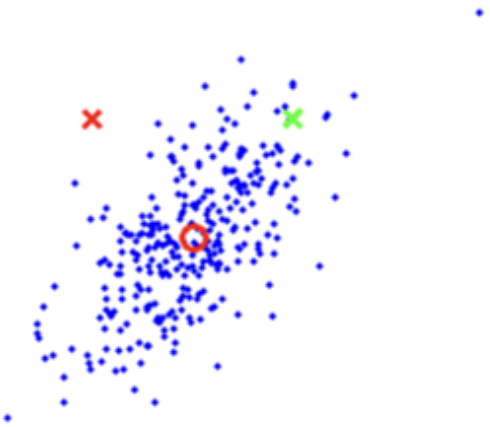

Figure 1: Which of the points marked in x is closer to the cluster centroid? in what sense?

The main takeout from the above example is that distance should be data-driven, and take the distribution of the data into account. Mahalanobis distance considers the covariance of the data, by multiplying the Euclidean distance with the inverse covariance. Specifically, for two points $x_i, x_j \in \mathbb{R}^m$, the distance is defined as

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T C^{-1} (x_i - x_j)}.$$

Since C is typically positive definite (for $n \geq m$), it can be inverted, so the distance is well-defined. To understand the Mahalanobis distance, consider the eigendecomposition $C = V \Lambda V^T$, and let $W := \Lambda^{-\frac{1}{2}} V^T$ be the PCA-whitening matrix. Then $C^{-1} = W W^T$. So the Mahalanobis distance is

$$d_M^2(x_i, x_j) = (x_i - x_j)^T W^T W (x_i - x_j) = \|W x_i - W x_j\|^2,$$

i.e., the Euclidean distance between the whitened points $W^T x_i$ and $W^T x_j$. Thus Mahalanobis distance is in fact the standard Euclidean distance on whitened data.

3 Local Mahalanobis

Real world data often lies on low dimensional manifold (e.g., a spiral). Many times, we are interested in finding neighboring points (for example, in applications of kNN. In such cases, it is benefitial to take into account local covariance matrices, rather than the global covariance matrix. For example, data on a cross (isotropic covariance). This yields the local Mahalanobis distance, where for each point we compute neighbors using its local metric, defined using the local covariance matrix. This can be used to design an iterated kNN algorithm as follows. In each iterations we find the k nearest neighbors using the current local metric (starting with the Euclidean metric at the first iteration). Then we compute the local covariance, and use the local metric to find new neighbors, The process is repeated until convergence (i.e., neighbors don't change).

4 Quadratic Discriminant Analysis

Suppose we have data samples from two classes, and the samples from each class as normally distributed with known mean vectors μ_0, μ_1 and covariance matrices Σ_0, Σ_1. Given a test point x, we form the likelihood ratio

$$\text{likelihood ratio} = \frac{\Pr(x|\mu_1, \Sigma_1)}{\Pr(x|\mu_0, \Sigma_0)} = \frac{\sqrt{2\pi|\Sigma_1|}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}{\sqrt{2\pi|\Sigma_0|}\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)}.$$

According to Bayes Optimal decision rule, we predict the class of x to be 1 if the likelihood ratio is greater than 1, and 0 otherwise. Taking the logarithm of the likelihood ratio, and ignoring terms not depending on x we get a decision rule according to which we predict 1 iff

$$(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) > t,$$

where t is some threshold depending on Σ_1, Σ_0. It can thus be seen that QDA essentially compares the Mahalanobis distances of x to the class means.

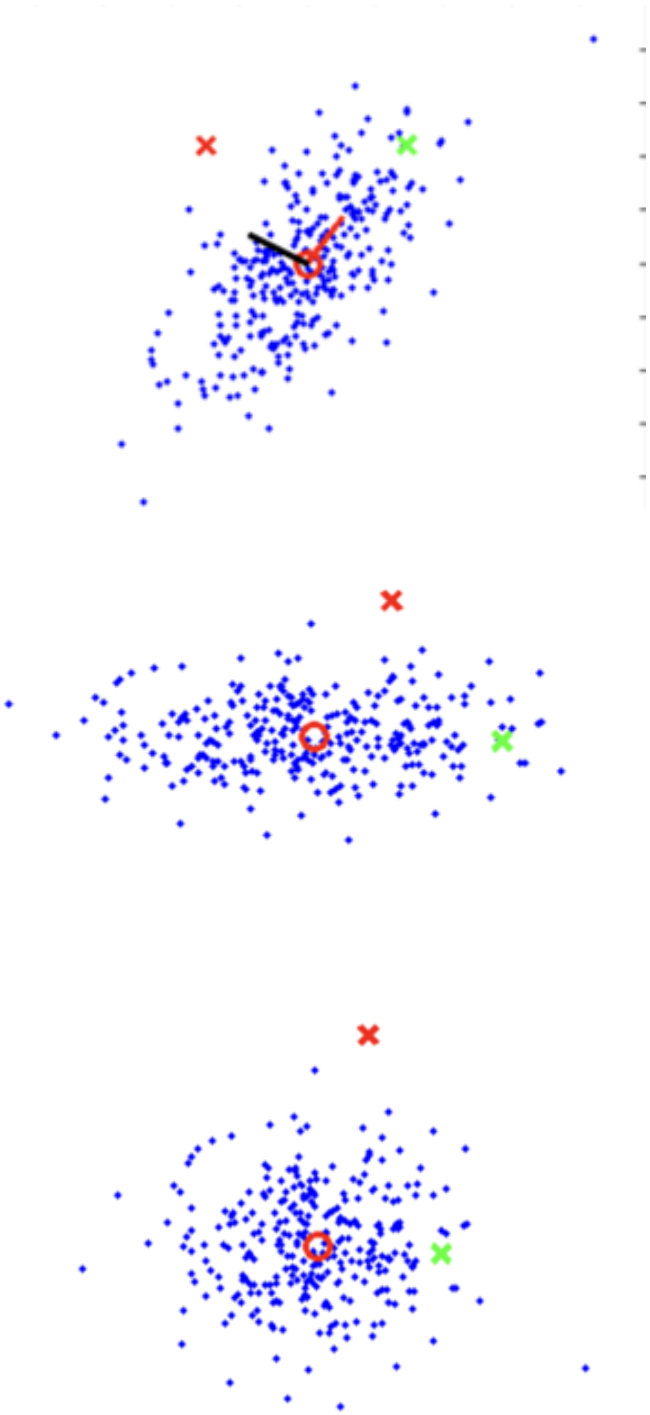

Figure 2: Mahalanobis distance is Euclidean distance after whitening.; Top: principal directions. Center: projection onto the principal directions. Bottom: PCA-whitening (identity covariance)