Generalized Mahalanobis Distance in Reproducing Kernel Hilbert Space

Shaoheng Liang, Jinzhuang Dou, Ken Chen

The University of Texas MD Anderson Cancer Center

Rice University

1 BACKGROUND

In data science, Mahalanobis distance is designed for measuring data points in a space of widely varying variances on different directions. It is also used for deliberately ignoring differences on certain directions [Qi and Davidson SIGKDD 09]. However, only when the data distribution is nearly normal does the distance provide a concise dissimilarity of datapoints. Moreover, the linear transform nature of the distance also limits its application in complexly distributed data. Nevertheless, we review the definition of distance, which includes a $p \times p$ matrix M, for a dataset with n datapoints and p features, measures distance of two datapoints x and x' as

$$d_\Sigma(x, x') = \sqrt{(x - x')^T \Sigma^{-1} (x - x')}.$$

Note that when Σ is an identity matrix, Mahalanobis distance degrades to Euclidean distance. In most context Σ is chosen to be the (empirical) covariance matrix but also may be tailored to specific tasks. Remarkably, it can be proven that Mahalanobis distance also defines a norm $\|x\|_\Sigma = d_\Sigma(x, \mathbf{0})$ in the space. In some articles and software, the inverse is not taken, so care should be taken.

One way to improve its power is to transform the data into a higher dimensional feature space to introduce nonlinearity. However, this also means that the computational complexity may significantly increase. For example, introducing quadratic terms onto a n-dimensional feature will add $\left(\frac{n(n+1)}{2}\right)$ new dimensions. The burden will increase exponentially should one desires to move to cubic, quartic, etc. In data science, a mature way to handle this problem is the kernel trick, which uses a kernel function $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ in the original space $\mathcal{X}$ to calculate the inner product in the targeted space $\mathcal{H}$. Formally, if there exists a projection $\phi: \mathcal{X} \to \mathcal{H}$, then inner product in the targeted space

$$\langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = k(x, x').$$

This trick greatly facilitates the computation and is widely used in SVM, PCA, etc. to solve linearly inseparable problems. Off-the-shelf polynomial kernel $k(x, x') = (\langle x, x' \rangle + c)^2$ is for adding higher order terms.

Mathematicians has proven that there is a unique relationship between a well-formed kernel function and a space $\mathcal{H}$, namely, a jargon, *reproducing kernel Hilbert space*. In fact, the elements in $\mathcal{H}$ are functions whose domain are $\mathcal{X}$. Functions are capricious and are likely not

describable by using finite dimensions. Indeed, fancier than adding finite number of higher order terms, the radial basis function (RBF, also called squared exponential, SE) kernel $k(x,x') = \exp(-\gamma\|x-x'\|^2)$ defines an infinite-dimensional space $\mathcal{H}$, containing any order of terms one may desire.

The task now is to find distance in $\mathcal{H}$. To make the question more intuitive, the squared distance may be written as

$$d^2\big(\phi(x),\phi(x')\big) = \big(\phi(x)-\phi(x')\big)^T\big(\phi(x)-\phi(x')\big)$$
$$= \phi(x)^T\phi(x) + \phi(x)^T\phi(x) - 2\phi(x)^T\phi(x') = k(x,x) + k(x',x') - 2k(x,x').$$

However, the Mahalanobis distance is nontrivial. Although it can be formally expressed as

$$d_\Sigma^2\big(\phi(x),\phi(x')\big) = \big(\phi(x)-\phi(x')\big)^T\Sigma^{-1}\big(\phi(x)-\phi(x')\big),$$

it is impossible to calculate it in this way because the dimensionality of $\mathcal{H}$ is infinity, which means Σ is an $\infty\times\infty$ (or $\aleph_0\times\aleph_0$ if higher precision is preferred) matrix on which inverse may not be taken. Here, we follow [Hu et al. Stat Papers 2011] to show a tangible way to calculate Mahalanobis distance in the reproducing kernel Hilbert space. In section 2, we describe generalized Mahalanobis distance to make it suitable for any positive semidefinite matrix Σ. In section 3 we describe another important property of eigenvalues and eigenvectors of positive semidefinite matrices. In section 4, we show a way to represent a special case ([Qi and Davidson SIGKDD 09]) of Mahalanobis distance by using kernel functions.

2 GENERALIZATION OF MAHALANOBIS DISTANCE

2.1 THEORY

Generalization of the definition of Mahalanobis is based on the orthogonality of the eigenvectors of positive semidefinite matrices. Assume that Σ has r nonzero eigenvalues σ_i^2, with eigenvectors v_i, the distance can be redefined as

$$d_\Sigma^2(x,x') = \sum_{i=1}^{r}\frac{((x-x')^Tv_i)^2}{\sigma_i^2},$$

which orthogonally extracts the component on each direction and divides it by the eigenvalue. The zero eigenvalues are naturally avoided in this manner.

2.2 IMPLEMENTATION

For simplicity, we denote

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \text{ and } D_\Sigma^2(X) = \begin{bmatrix} d_\Sigma^2(x_1,x_1) & \cdots & d_\Sigma^2(x_1,x_n) \\ \vdots & \ddots & \vdots \\ d_\Sigma^2(x_n,x_1) & \cdots & d_\Sigma^2(x_n,x_n) \end{bmatrix}.$$

Because $(x-x')^Tv_i = x^Tv_i - x'^Tv_i$. Apparently,

$$D_\Sigma^2(X) = \sum_{i=1}^{r}\frac{\big((Xv_i)\ominus(Xv_i)^T\big)^{\cdot 2}}{\sigma_i^2},$$

where $\ominus$ means outer subtract of vectors which subtracts entries pairwisely and yields a matrix

and $\cdot\ 2$ means pointwise square. In practice, we can use predefined top r eigenvalues to further expedite execution.

3 IMPORTANT PROPERTY OF EIGENVALUES AND EIGENVECTORS OF POSITIVE SEMIDEFINITE MATRICES

Two positive semidefinite matrices A^TA and AA^T, formed by two ways of matrix multiplication of an arbitrary rectangle matrix A, have related eigenvalues and eigenvectors. In fact, the nonzero eigenvalues σ_i^2 are the same (if sorting them into order). The eigenvectors v_i and u_i, respectively, are related by equations

$$v_i = \frac{A^T u_i}{\sigma_i} \text{ and } u_i = \frac{Av_i}{\sigma_i}.$$

Note that σ_i is the square root of the eigenvalue σ_i^2.

4 DERIVATION OF A SPECIAL CASE OF MAHALANOBIS DISTANCE

4.1 THEORY

This special case, inspired by [Qi and Davidson SIGKDD 09], aims to mitigate the batch effects between samples. The data are gathered among different times points, or even from different individuals, which may introduce discrepancies. We assume that we have c samples, among which batch effect may occur. Sample numbered a contains n_a datapoints, whose index are in C_a. There are $n = \sum_{a=1}^{c} n_a$ datapoints in total. To facilitate our derivation, we denote the datapoints by two systems. The first one is the most common, i.e., $\{x_i\}_{i=1}^{n}$. The second one reorders the datapoints by in or out a sample a, denoted as $\{x_{(a;i)}\}_{i=1}^{n_a}$ and $\{x_{(\neg a;i)}\}_{i=1}^{n-n_a}$. In words, a selects all datapoints from sample a (or $\neg a$ not from sample a), re-index the points from 1 to n_a (or $n - n_a$), and i further picks out a datapoint. The subscription uses a semicolon to help remember this fact.

4.1.1 Covariance matrix

The "covariance matrix" used in the Mahalanobis distance is defined as

$$\Sigma = \sum_{a=1}^{c} \sum_{i \notin C_a} \left(x_i - \frac{1}{n_a} \sum_{g \in C_a} x_g \right) \left(x_i - \frac{1}{n_a} \sum_{g \in C_a} x_g \right)^T$$

$$= \frac{1}{n} \sum_{a=1}^{c} \sum_{i=1}^{n-n_a} \left(x_{(\neg a;i)} - \frac{1}{n_a} \sum_{g=1}^{n_a} x_{(a;g)} \right) \left(x_{(\neg a;i)} - \frac{1}{n_a} \sum_{g=1}^{n_a} x_{(a;g)} \right)^T.$$

For $\mathcal{H}$, we can simply add the projection $\phi(\cdot)$ into it, as

$$\Sigma = \sum_{a=1}^{c} \sum_{i=1}^{n-n_a} \left(\phi\left(x_{(\neg a;i)}\right) - \frac{1}{n_a} \sum_{g=1}^{n_a} \phi\left(x_{(a;g)}\right) \right) \left(\phi\left(x_{(\neg a;i)}\right) - \frac{1}{n_a} \sum_{g=1}^{n_a} \phi\left(x_{(a;g)}\right) \right)^T.$$

Using the property of block matrix, it can also be expressed as

$$\Sigma = A^T A,$$

where

$$A = \begin{bmatrix} A_{(1;1)}^T \\ \vdots \\ A_{(1;n-n_1)}^T \\ \vdots \\ \vdots \\ A_{(c;1)}^T \\ \vdots \\ A_{(c;n-n_1)}^T \end{bmatrix} = \begin{bmatrix} \left(\phi\left(x_{(\neg 1;1)}\right) - \frac{1}{n_1} \sum_{g=1}^{n_1} \phi\left(x_{(1;g)}\right) \right)^T \\ \vdots \\ \left(\phi\left(x_{(\neg 1;n-n_1)}\right) - \frac{1}{n_1} \sum_{g=1}^{n_1} \phi\left(x_{(1;g)}\right) \right)^T \\ \vdots \\ \vdots \\ \vdots \\ \left(\phi\left(x_{(\neg c;1)}\right) - \frac{1}{n_c} \sum_{g=1}^{n_c} \phi\left(x_{(c;g)}\right) \right)^T \\ \vdots \\ \left(\phi\left(x_{(\neg c;n-n_c)}\right) - \frac{1}{n_c} \sum_{g=1}^{n_c} \phi\left(x_{(c;g)}\right) \right) \end{bmatrix}$$

However, as stated before the Σ of infinite dimensions is in fact not attainable. Instead, we need to use AA^T to observe an "facet" of Σ. Thus, block matrix from which eigenvectors u_i are extracted

$$AA^T = \begin{bmatrix} [AA^T]_{(1,1)} & \cdots & [AA^T]_{(1,c)} \\ \vdots & \ddots & \vdots \\ [AA^T]_{(c,1)} & \cdots & [AA^T]_{(c,c)} \end{bmatrix}$$

where for $i \in \{1, \dots, (n - n_a)\}$ and $j \in \{1, \dots, (n - n_b)\}$, we have

$$
\begin{aligned}
[AA^T]_{(a,b;i,j)} &\triangleq \left[[AA^T]_{(a,b)}\right]_{(i,j)} \\
&= \left\langle \phi\left(x_{(\neg a;i)}\right) - \frac{1}{n_a}\sum_{g=1}^{n_a}\phi\left(x_{(a;g)}\right), \phi\left(x_{(\neg b;j)}\right) - \frac{1}{n_b}\sum_{h=1}^{n_b}\phi\left(x_{(b;h)}\right)\right\rangle \\
&= \left\langle \phi\left(x_{(\neg a;i)}\right), \phi\left(x_{(\neg b;j)}\right)\right\rangle + \left\langle \frac{1}{n_a}\sum_{g=1}^{n_a}\phi\left(x_{(a;g)}\right), \frac{1}{n_b}\sum_{h=1}^{n_b}\phi\left(x_{(b;h)}\right)\right\rangle \\
&- \left\langle \frac{1}{n_a}\sum_{g=1}^{n_a}\phi\left(x_{(a;g)}\right), \phi\left(x_{(\neg b;j)}\right)\right\rangle - \left\langle \phi\left(x_{(\neg a;i)}\right), \frac{1}{n_b}\sum_{h=1}^{n_b}\phi\left(x_{(b;h)}\right)\right\rangle \\
&= \left\langle \phi\left(x_{(\neg a;i)}\right), \phi\left(x_{(\neg b;j)}\right)\right\rangle + \frac{1}{n_a n_b}\sum_{g=1}^{n_a}\sum_{h=1}^{n_b}\left\langle \phi\left(x_{(a;g)}\right), \phi\left(x_{(b;h)}\right)\right\rangle \\
&- \frac{1}{n_a}\sum_{g=1}^{n_a}\left\langle \phi\left(x_{(a;g)}\right), \phi\left(x_{(\neg b;j)}\right)\right\rangle - \frac{1}{n_b}\sum_{h=1}^{n_b}\left\langle \phi\left(x_{(\neg a;i)}\right), \phi\left(x_{(b;h)}\right)\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
[AA^T]_{(a,b;i,j)} = k\left(x_{(\neg a;i)}, x_{(\neg b;j)}\right) &+ \frac{1}{n_a n_b}\sum_{g=1}^{n_a}\sum_{h=1}^{n_b}k(x_{(a;g)}, x_{(b;h)}) - \frac{1}{n_a}\sum_{g=1}^{n_a}k(x_{(a;g)}, x_{(\neg b;j)}) \\
&- \frac{1}{n_b}\sum_{h=1}^{n_b}k\left(x_{(\neg a;i)}, x_{(b;h)}\right)
\end{aligned}
$$

4.1.2 Distance of two points

Besides the covariance matrix, we also need the

$$
\frac{\left(\left(\phi(x) - \phi(x')\right)^T v_i\right)^2}{\sigma_i^2} = \frac{\left(\left(\phi(x) - \phi(x')\right)^T A^T u_i\right)^2}{\sigma_i^4}
$$

where

$$
\left(\phi(x) - \phi(x')\right)^T A^T = \left(\phi(x) - \phi(x')\right)^T \begin{bmatrix} A_{(1;1)}^T \\ \vdots \\ A_{(c;n-n_c)}^T \end{bmatrix}^T,
$$

where

$$\delta_{(a;i)}(x,x') \triangleq \left(\phi(x)-\phi(x')\right)^T A_{(a;i)}^T = \left\langle \phi(x)-\phi(x'), \phi\left(x_{(\neg a;i)}\right) - \frac{1}{n_a}\sum_{g=1}^{n_a}\phi\left(x_{(a;g)}\right)\right\rangle$$

$$= \left\langle \phi(x), \phi\left(x_{(\neg a;i)}\right)\right\rangle + \frac{1}{n_a}\sum_{g=1}^{n_a}\left\langle \phi(x'), \phi\left(x_{(a;g)}\right)\right\rangle - \left\langle \phi(x'), \phi\left(x_{(\neg a;i)}\right)\right\rangle$$

$$-\frac{1}{n_a}\sum_{g=1}^{n_a}\left\langle \phi(x), \phi\left(x_{(a;g)}\right)\right\rangle$$

$$= k\left(x, x_{(\neg a;i)}\right) - k\left(x', x_{(\neg a;i)}\right) + \frac{1}{n_a}\sum_{g=1}^{n_a} k\left(x', x_{(a;g)}\right) - \frac{1}{n_a}\sum_{g=1}^{n_a} k\left(x, x_{(a;g)}\right).$$

Conclusively, in

$$d_\Sigma^2(x,x') = \sum_{l=1}^{r}\frac{((\phi(x)-\phi(x'))^T v_l)^2}{\sigma_l^2} = \sum_{l=1}^{r}\frac{\left(\left(\phi(x)-\phi(x')\right)^T A^T u_l\right)^2}{\sigma_l^4}$$

$$= \sum_{i=l}^{r}\frac{\left(\begin{bmatrix}\delta_{(1;1)}(x,x')^T\\ \vdots \\ \delta_{(c;c-n_a)}(x,x')^T\end{bmatrix}^T u_l\right)^2}{\sigma_l^4},$$

where $\delta_{(a;i)}(x,x')$ can be easily calculated and u_i can be extracted from a finite-dimensional matrix AA^T. In fact, AA^T is an $\left((c-1)n\right)\times\left((c-1)n\right)$ matrix. This yields the Mahalanobis distance in the reproducing kernel Hilbert space $\mathcal{H}$.

4.2 IMPLEMENTATION

It is noticeable that in

$$[AA^T]_{(a,b;i,j)} = k\left(x_{(\neg a;i)}, x_{(\neg b;j)}\right) + \sum_{g=1}^{n_a}\sum_{h=1}^{n_b} k(x_{(a;g)}, x_{(b;h)}) - \sum_{g=1}^{n_a} k\left(x_{(a;g)}, x_{(\neg b;j)}\right)$$

$$- \sum_{h=1}^{n_b} k\left(x_{(\neg a;i)}, x_{(b;h)}\right),$$

only kernel values are included. Thus, we may first calculate the matrix $K_{ij} = k\left(x_i, x_j\right)$ to avoid duplicate calculations. It is also observable that the last three terms are shared among many entries.

5 REFERENCES

[Hu et al. Stat Papers 2011] Hu, Yonggang, et al. "Generalized Mahalanobis depth in the reproducing kernel Hilbert space." Statistical Papers 52.3 (2011): 511-522.

[Qi and Davidson KDD 09] Qi, ZiJie, and Ian Davidson. "A principled and flexible framework for finding alternative clusterings." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.