TU München
Fakultät für Informatik
PD Dr. Rudolph Triebel
John Chiotellis, Maximilian Denninger

Machine Learning for Computer Vision

June 2, 2018
Topic: Metric Learning

Exercise 1: Metric Learning

a) *Given a valid metric D_M, is D_M^2 also a metric? Why?*

It is easy to show that three properties of a metric hold:

- Non-negativity:
$$D_M(x,y) \geq 0 \Rightarrow D_M^2(x,y) \geq 0 \tag{1}$$
- Symmetry:
$$D_M(x,y) = D_M(y,x) \Rightarrow D_M^2(x,y) = D_M^2(y,x) \tag{2}$$
- Identity of indiscernibles:
$$D_M(x,y) = 0 \Leftrightarrow x = y \Rightarrow D_M^2(x,y) = 0 \Leftrightarrow x = y \tag{3}$$

But, D_M^2 is not a metric, as we cannot guarantee that the triangle inequality is preserved. We can easily find counterexamples. For instance, in the one-dimensional case with the euclidean metric we have:

$$D(0,2) \leq D(0,1) + D(1,2)$$

but

$$D^2(0,2) > D^2(0,1) + D^2(1,2).$$

We can also prove that the triangle inequality does not (always) hold. For arbitrary vectors $x, y, z \in \mathbb{R}^d$, we have:

$$\begin{aligned}
D_M^2(x,z) + D_M^2(z,y) &= (x-z)^T M(x-z) + (z-y)^T(z-y) \\
&= x^T Mx + z^T Mz - 2x^T Mz + z^T Mz + y^T My - 2z^T My \\
&= x^T Mx + y^T My + 2z^T Mz - 2x^T Mz - 2z^T My \\
&= D_M^2(x,y) + 2x^T My + 2z^T Mz - 2x^T Mz - 2z^T My \\
&= D_M^2(x,y) + 2(x^T M(y-z) + z^T M(z-y)) \\
&= D_M^2(x,y) + 2(x^T M(y-z) - z^T M(y-z)) \\
&= D_M^2(x,y) + 2(x-z)^T M(y-z).
\end{aligned}$$

The term $(x-z)^T M(y-z)$ is an inner product of the vectors $x-z$ and $y-z$ in the new metric space:

$$(x-z)^T M(y-z) = ||x-z||||M(y-z)|| \cos(\angle(x-z, M(y-z))) = \gamma \cos\theta \qquad , \gamma > 0$$

Since the vectors x, y, z were arbitrary, $|\theta| \in [0, \pi]$. Therefore we have:

$$D_M^2(x,z) + D_M^2(z,y) \begin{cases} \geq D_M^2(x,y) & \text{if } |\theta| \in [0, \frac{\pi}{2}] \\ < D_M^2(x,y) & \text{if } |\theta| \in (\frac{\pi}{2}, \pi] \end{cases}$$

For $|\theta| \in (\frac{\pi}{2}, \pi]$ the triangle inequality is violated.

b) *Given a matrix X of n data points $x_i \in \mathbb{R}^d$, show how computing the eigendecomposition of the covariance of X is equivalent to computing the singular value decomposition of X .*

The covariance of X is a real symmetric semi-positive definite matrix $C = \frac{1}{n} X^T X$. Therefore we know it has an eigendecomposition of the form $C = \frac{1}{n} V \Lambda V^T$, where V contains the eigenvectors of C as columns and Λ contains the corresponding non-negative eigenvalues in the diagonal.
Assuming X has a singular value decomposition $X = U \Sigma W^T$, we can infer a decomposition for C as

$$C = \frac{1}{n} X^T X = \frac{1}{n} W \Sigma U^T U \Sigma W^T = \frac{1}{n} W \Sigma^2 W^T$$

By inspection, we can see that the two decompositions are identical, with $W = V$ and $\Lambda = \Sigma^2$.

c) *What is the difference between metric learning and kernel learning? When would you prefer to use one over the other?*

Metric learning methods attempt to find a linear embedding of a given set of input vectors, such that a clustering or classification objective is optimized in the transformed space. This embedding has usually fewer or at worst equal number of dimensions as the original input space.

In contrast, kernel methods attempt to implicitly find a higher-dimensional representation of the data, such that the classes can be linearly separated. This representation does not need to be found explicitly but is exploited by using only inner products of the high-dimensional features (kernel trick).

Therefore when the dataset size is larger than the dimensionality of the given inputs ($n > d$), it will be more efficient to work with a covariance-like matrix. Metric learning will try to find inherent object relationships in the essential dimensions and weight them appropriately. When the size of the dataset is smaller than the inputs dimensionality ($n < d$), kernel methods are a more appropriate choice, as it will be more efficient to work with the kernel matrix and more complex - non-linear - relationships can be expressed.

d) *In Neighborhood Component Analysis, we define a stochastic neighbor selection rule. The probability that a data point j is selected as neighbor of point i is given by:*

$$p_{ij} = \frac{\exp\{-||Lx_i - Lx_j||^2\}}{\sum\limits_{k \neq i} \exp\{-||Lx_i - Lx_k||^2\}} \tag{4}$$

namely a softmax over the squared distances to all points in the transformed space. The goal is to maximize

$$f(L) = \sum_i \sum_{j \in C_i} p_{ij} \tag{5}$$

namely the probability that the neighbors that will be selected for each point i will belong to the same class C_i. Can you derive the gradient of $f(L)$?

Let us notate the denominator of p_{ij} as $Z(L) = \sum\limits_{k \neq i} \exp\{-||Lx_i - Lx_k||^2\}$.
We will make use of the fact that we can write squared distances as traces:

$$\begin{aligned} ||Lx_i - Lx_j||^2 &= (L(x_i - x_j))^T(L(x_i - x_j)) = (x_i - x_j)^T L^T L(x_i - x_j) \\ &= tr(L^T L(x_i - x_j)(x_i - x_j)^T) = tr(L^T L C_{ij}) \end{aligned}$$

where for convenience we defined $C_{ij} = (x_i - x_j)(x_i - x_j)^T$. And we will use the gradient w.r.t. L: $\nabla_L tr(L^T L C_{ij}) = 2LC_{ij}$.

First we can see that the gradient can go inside of the sums: $\nabla_L f(L) = \sum\limits_i \sum\limits_{j \in C_i} \nabla_L p_{ij}$.

Then we see that we have to apply the quotient rule, since L appears in both the nominator and denominator of p_{ij}:

$$\begin{aligned} \nabla_L p_{ij} &= \nabla_L \left(\frac{\exp\{-||Lx_i - Lx_j||^2\}}{Z(L)} \right) \\ &= \frac{(\nabla_L \exp\{-||Lx_i - Lx_j||^2\})\, Z(L) - (\nabla_L Z(L)) \exp\{-||Lx_i - Lx_j||^2\}}{Z(L)^2} \end{aligned}$$

The gradient of the nominator is

$$\begin{aligned} \nabla_L \exp\{-||Lx_i - Lx_j||^2\} &= \exp\{-||Lx_i - Lx_j||^2\} \nabla_L (-||Lx_i - Lx_j||^2) \\ &= -\exp\{-||Lx_i - Lx_j||^2\} \nabla_L tr(L^T L C_{ij}) \\ &= -2 \exp\{-||Lx_i - Lx_j||^2\} L C_{ij} \end{aligned}$$

The gradient of the denominator is similar:

$$\begin{aligned}\nabla_L Z(L) &= \nabla_L \sum_{k\neq i} \exp\{-||Lx_i - Lx_k||^2\} = \sum_{k\neq i} \nabla_L \exp\{-||Lx_i - Lx_k||^2\} \\ &= \sum_{k\neq i} \nabla_L \exp\{-||Lx_i - Lx_k||^2\} = -2L \sum_{k\neq i} \exp\{-||Lx_i - Lx_k||^2\} C_{ik}\end{aligned}$$

Therefore we have

$$\begin{aligned}\nabla_L p_{ij} &= -2L \frac{\exp\{-||Lx_i - Lx_j||^2\} C_{ij} Z(L) - \sum_{k\neq i} \exp\{-||Lx_i - Lx_k||^2\} C_{ik} \exp\{-||Lx_i - Lx_j||^2\}}{Z(L)^2} \\ &= -2L \left(\frac{\exp\{-||Lx_i - Lx_j||^2\}}{Z(L)} C_{ij} - \sum_{k\neq i} \frac{\exp\{-||Lx_i - Lx_k||^2\}}{Z(L)} C_{ik} \frac{\exp\{-||Lx_i - Lx_j||^2\}}{Z(L)} \right) \\ &= -2L \left(p_{ij} C_{ij} - \sum_{k\neq i} p_{ik} C_{ik} p_{ij} \right) = 2L p_{ij} \left(\sum_{k\neq i} p_{ik} C_{ik} - C_{ij} \right)\end{aligned}$$

And summing over all inputs again we get the total gradient

$$\begin{aligned}\nabla_L f(L) &= \sum_i \sum_{j\in C_i} 2L p_{ij} \left(\sum_{k\neq i} p_{ik} C_{ik} - C_{ij} \right) & &= 2L \sum_i \sum_{j\in C_i} p_{ij} \left(\sum_{k\neq i} p_{ik} C_{ik} - C_{ij} \right) \\ &= 2L \sum_i \left(\sum_{j\in C_i} p_{ij} \sum_{k\neq i} p_{ik} C_{ik} - \sum_{j\in C_i} p_{ij} C_{ij} \right) & &= 2L \sum_i \left(p_i \sum_{k\neq i} p_{ik} C_{ik} - \sum_{j\in C_i} p_{ij} C_{ij} \right)\end{aligned}$$

e) *What is the difference between LDA and NCA?* Linear Discriminant Analysis (LDA) maximizes the variance between clusters with respect to the variance within clusters. This can be formulated in one nice short equation that has a closed form solution, namely the optimal transformation is the eigenvectors of the matrix $C_w^{-1} C_b$. The main drawback of LDA is that it assumes that each cluster follows a normal distribution.

Neighborhood Component Analysis (NCA) on the other hand is a neighbors based method (as the name suggests) and therefore does not make any assumptions about the clusters distributions. The objective it tries to maximize is derived from a probabilistic point of view. It wants to maximize the probability that the neighbors that will be selected, according to the learned metric, will belong to the same class. The drawback of NCA is that it needs to estimate $n \times n$ probabilities at each iteration of the optimization.

f) *The KL-divergence measures the (dis-)similarity of two probability distributions. It is defined as:*

$$D_{KL}(p||q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

Is the KL-divergence a metric? Why?

No, the KL-divergence is not a metric, because it does not satisfy the symmetry property. It is non-negative, and the identity of indiscernibles holds for distributions but symmetry cannot be guaranteed. If it was symmetric, it would mean that $D_{KL}(p||q) = D_{KL}(q||p)$.

$$\begin{aligned}
D_{KL}(p||q) - D_{KL}(q||p) &= \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx - \int q(x) \log\left(\frac{q(x)}{p(x)}\right) dx \\
&= \int p(x) \log p(x) - p(x) \log q(x) - q(x) \log q(x) + q(x) \log p(x) dx \\
&= \int (p(x) + q(x))(\log p(x) - \log q(x)) \\
&= \int (p(x) + q(x)) \log\left(\frac{p(x)}{q(x)}\right) \\
&= 2\mathbb{E}_{\frac{p+q}{2}}[\log\left(\frac{p(x)}{q(x)}\right)]
\end{aligned}$$

We can see this as a weighted average of $\log\left(\frac{p(x)}{q(x)}\right)$ under the distribution $\frac{1}{2}(p(x) + q(x))$. There is no reason to assume that this quantity is zero.