# PH 240C Supervised Learning (3): Metric Learning

Jingshen Wang

September 22, 2021

## 1 Metric Learning

Although the origin of metric learning can be traced back to some earlier work, it really emerged in 2002 with the pioneering work of Xing et al. (2002) which formulates metric learning as a convex optimization problem. Before we (in)formally introduce metric learning, we first look in to two examples.

**Example 1** (Predicting CAD)**.** *Suppose we want to use the synthetic data presented in Table 1 for heart attack prediction. We can clearly see that attributes are scaled differently, and maybe correlated. This raises two questions:*

1. *If we go for the traditional approaches (well integrated in* `R` *packages), predictors like income might get down-weighted as it has with high variances. But income might be, in fact, a very important predictor for predicting hearth disease.*
2. *Say if we use SVM dual problem to do classification with a quadratic kernel function* $K(x,z) = ||x-z||_2^2$*, it ignores the correlation between variables and the genuinely significant variable might not play an important role.*

In practice, we might apply some preprocessing of the data (say, normalization or standardization) to resolve these issues:

$$\text{Original data: } \{(Y_i, X_i)\}_{i=1}^n, \quad \overset{\text{transformation}}{\Longrightarrow} \quad \text{Transformed data for prediction: } \widetilde{X}_i = \frac{X_i - \bar{X}_n}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}}.$$

Rather than working with some arbitrary transformation of the data, the intuition of metric learning is to learn a better distance metric that is more relevant for the prediction problem at hand.

| Patient ID | Gender | Age | Weight (kg) | Income ($) | Heart Attack |
|---|---|---|---|---|---|
| 1 | 1 | 63 | 82 | 56,000 | 1 |
| 2 | 0 | 43 | 67 | 105,000 | 0 |
| 3 | 0 | 55 | 70 | 75,000 | 0 |
| 4 | 1 | 76 | 68 | 60,000 | 1 |
| new | 0 | 50 | 78 | 90,000 | ? |

Table 1: (Synthetic) Data similar to Homework 1. How can we predict whether the new patient will get a heart attack?

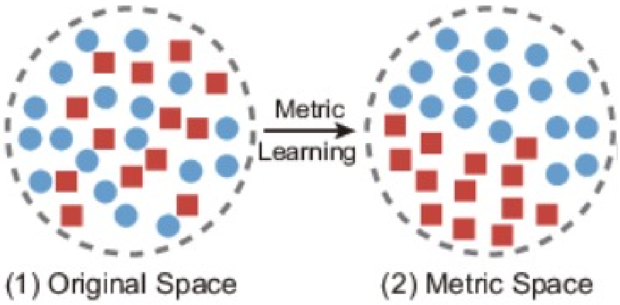

Figure 1: A illustrative example for Metric learning. Think about metric as a "ruler."

More rigorously, the goal of metric learning is to adapt some metric function to our data so that our supervised learning problem can provide more accurate prediction. For example, we want to find a metric $G \in \mathbb{R}^{d\times d}$ that transform the original data as the following:

$$\text{Original data: } \{(Y_i, X_i)\}_{i=1}^n, \quad \overset{\text{transformation}}{\Longrightarrow} \quad \text{Transformed data for prediction: } \widetilde{X}_i = GX_i.$$

Then, given $G$ is unknown, we must define certain criteria for an "ideal" $G$ for prediction purposes. The criteria of metric learning starts by defining two sets:

$$\mathcal{S} = \big\{(i,j): \ Y_i \text{ and } Y_j \text{ have the Same label}\big\},$$
$$\mathcal{D} = \big\{(i,j): \ Y_i \text{ and } Y_j \text{ do Not have the same label}\big\},$$

and it aims to find a matrix $G$ *such that* the sum of distances between similar individuals are minimized, and the sum of the distances between dissimilar individuals are maximized. Concretely, metric learning aims to find $G$ such that

$$\max_{(i,j)\in\mathcal{D}} \text{Distance}\big(GX_i, GX_j\big) \quad \text{and} \quad \min_{(i,j)\in\mathcal{S}} \text{Distance}\big(GX_i, GX_j\big).$$

See Figure 1 for a general illustration. We want to find a metric space to measure the distance between observations so that the similar objects move closer, and dissimilar objects can be separated. In other words, the distance metric provides a new data representation in the transformed space which is easily able to distinguish the items of different classes.

**What is a metric? A simple example** To better understand the role of $G$ in metric learning, let's look at an example in Figure 2. If we set $G = I$, the identity matrix, separating low risk and high risk patients is not straightforward. Because the distances between similar/dissimilar objectives are very close. We would measure the distance between, for example, $X_1$ and $X_3$ via the simple Euclidean distance:

$$d(X_1, X_3) = ||X_1 - X_3||_2^2 = (X_{11} - X_{31})^2 + (X_{12} - X_{32})^2.$$

But if we do not stick to Euclidean distance, instead switch to a different distance measure:

$$d^*(X_1, X_3) = ||GX_1 - GX_3||_2^2 = (X_{12} - X_{32})^2,$$

where

$$G = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Such a distance measure can more effectively put similar/dissimilar individuals into two groups, as we can calculate

$$\sum_{(i,j)\in\mathcal{D}} d^*(X_i, X_j) = \binom{2}{5} \times 10 = 100, \quad \sum_{(i,j)\in\mathcal{S}} d^*(X_i, X_j) = 0.$$

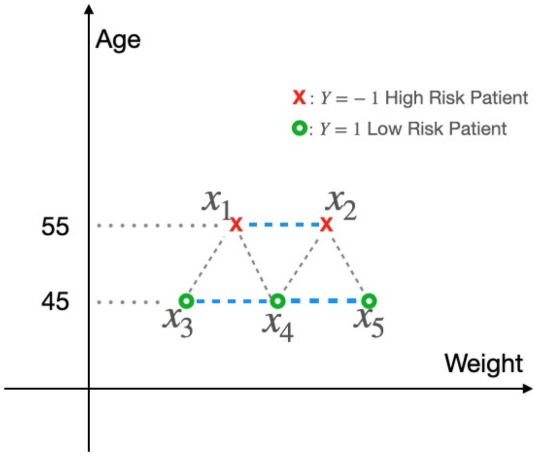

Figure 2: An example for Metric learning.

By doing this exercise, we have changed the "ruler" that measures the distance between data points from $d(\cdot,\cdot)$ to $d^*(\cdot,\cdot)$, metric learning can then very efficiently separate the data into different groups.

**Metric learning with Mahalanobis distance** To introduce some statistical terms (jargon), the distance measure used in the seminal work Xing et al. (2002) is based on the Mahalanobis distance, which measures the distance between two data points via

$$d_M(X_i, X_j) = \sqrt{(X_i - X_j)' M (X_i - X_j)}.$$

And they aim to find a matrix $M$ so that:

$$\begin{aligned} \max_M \quad & \sum_{(i,j)\in\mathcal{D}} d_M(X_i, X_j) \\ \text{s.t.} \quad & \sum_{(i,j)\in\mathcal{S}} d_M(X_i, X_j) \leq 1. \end{aligned} \tag{1}$$

The optimization problem in (1) is a convex optimization problem and can be solved efficiently via projected gradient descent algorithm as discussed in Xing et al. (2002). Furthermore, the distance measure we adopt in (1) is a non-isotropic distance that reflects some intrinsic structure of the data.

After solving for $M$, for a new patient with attribute x, we can calculate the average distances from x to the data labelled as 1, that is $\frac{1}{\sum_i \mathbf{1}(Y_i=1)} \sum_{i:Y_i=1} d_M(X_i, \mathrm{x})$, and to the data labelled as 0, that is $\frac{1}{\sum_i \mathbf{1}(Y_i=0)} \sum_{i:Y_i=0} d_M(X_i, \mathrm{x})$. We predict y = 1 whenever

$$\frac{1}{\sum_i \mathbf{1}(Y_i = 1)} \sum_{i:Y_i=1} d_M(X_i, \mathrm{x}) \geq \frac{1}{\sum_i \mathbf{1}(Y_i = 0)} \sum_{i:Y_i=0} d_M(X_i, \mathrm{x}),$$

and vice versa.

Can you think about a different formulation of the optimization problem?

**Difference with Kernel methods?** Metric learning is natural to adapt for semi-supervised learning. Think about the example in Figure 3, where in total we have $n = 23$ pictures of 4 bald Hollywood action

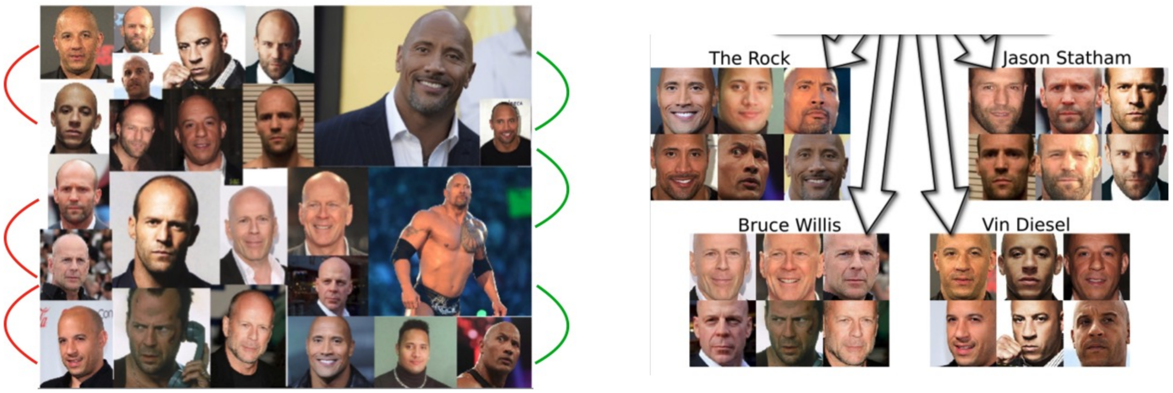

Figure 3: Who's who? Illustration of metric learning applied to a face recognition task.

heroes. The goal is to use metric learning to identify how many different people there are and which faces belong to each person. Because metric learning needs us to specify a set that contains similar individuals $\mathcal{S}$, and a set that contains dissimilar individuals $\mathcal{D}$. Pairwise similarities are measured based on whether two images representing the same person (similar-link, shown in green) or different persons (dissimilar-link, shown in red). Based on measured similarities (note that we do not need to go for all pairs), we wish to adapt the metric so that we can separate these persons. The (deep) metric learning algorithms can automatically identified face into fours clusters. You can read more about this example in this post.

While metric learning is parametric (one learns the parameters of a given form of metric, such as a Mahalanobis distance), kernel learning is usually nonparametric: one learns the kernel matrix without any assumption on the form of the kernel that implicitly generated it. These approaches are thus very powerful but limited to the transductive setting and can hardly be applied to new data.

## 2 Brainstorm time for the final project

In the past few lectures, we have learnt some supervised machine learning methods. Other than disease risk prediction, I want to open up the discussion regarding what machine learning methods can do for us. Researchers have known about the link between blood pressure and Alzheimer's disease (AD) for years. In Skoog and Gustafson (2006), investigators have showed that older people with high blood pressure, or hypertension, were more likely to have biomarkers of Alzheimer's in their spinal fluid.

Let's think about the following questions:

- While research investigates the brain benefits of blood pressure medications, one can play it smart by taking healthy lifestyle steps to keep blood pressure in a healthy range, like having a mediterranean diet (a diet is high in fruits, vegetables, whole grains, low-fat dairy, poultry, fish and nuts), shed extra weight by doing more exercise (but what kind of exercises?). Given there are many options to modify one's lifestyle, what lifestyle modifications are helpful for AD patients to lower their blood pressure? Note AD patients may behave very differently from other patients (Nasrallah et al., 2021), and intensive systolic blood pressure control might be harmful for AD patients.

- There is increasing evidence linking cholesterol metabolism with the pathology of Alzheimer's disease (AD). Statins are perhaps the most commonly prescribed drug due to its clear benefits in reducing the level of low–density lipoprotein (LDL)–the"bad" cholesterol. Studies have examined the role of statins in the prevention of dementia and treatment of established AD. Let alone the broad use of statins, there have been increasing concerns that the statin usage is potentially associated with the increased risk of new-onset type II diabetes (T2D) (Waters et al., 2013). Moreover, the FDA warns on statin labels that some people have developed memory loss or confusion while taking statins. Given people respond differently to different treatments, can we find out what kind of patients benefit from taking statins? And what kind of patients suffer from the adverse effect of the statin usage?

- If you want to answer the above questions, what kind of data would you need? Randomized control trials? Observational studies? What are the benefits and the cavities?

# References

Ilya M Nasrallah, Sarah A Gaussoin, Raymond Pomponio, Sudipto Dolui, Guray Erus, Clinton B Wright, Lenore J Launer, John A Detre, David A Wolk, Christos Davatzikos, et al. Association of intensive vs standard blood pressure control with magnetic resonance imaging biomarkers of alzheimer disease: Secondary analysis of the sprint mind randomized trial. *JAMA neurology*, 78(5):568–577, 2021.

Ingmar Skoog and Deborah Gustafson. Update on hypertension and alzheimer's disease. *Neurological research*, 28(6):605–611, 2006.

David D Waters, Jennifer E Ho, S Matthijs Boekholdt, David A DeMicco, John JP Kastelein, Michael Messig, Andrei Breazna, and Terje R Pedersen. Cardiovascular event reduction versus new-onset diabetes during atorvastatin therapy: effect of baseline risk factors for diabetes. *Journal of the American College of Cardiology*, 61(2):148–152, 2013.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15:521–528, 2002.