HILBERT SCHMIDT INDEPENDENCE CRITERION IN MACHINE LEARNING

Yurong, Chen*
Hunan University
chenyurong1998@outlook.com

Hui Zhang
Hunan University

August 9, 2021

1 Introduction

Independence. In statistics, the observations in a sample are commonly assumed as independent and identically distributed (i.i.d) random variables. For example, an region in an image $X = \{x_1, x_2, ..., x_n\}$ belongs to the cat class, where x_n represents a pixel (i.e. a random variable). Each pixel has the same probability distribution as the others and all are mutually independent. The formal definition can be written as:

$$x_n \sim P(X);$$

$$p(x_i, x_j) = p(x_i) \times p(x_j).$$

The first equation denotes that all pixels are identically distributed (i.e. each pixel follows the same distribution), and the second equation means the two are independent. Generally, two random features are independent if the occurrence of one does not affect the probability of occurrence of the other (e.g. the occurrence pixel of cat's ear does not affect the pixel of cat's tail). From the probabilistic theory point of view, the joint distribution can be factorized as the product of margin distribution. This assumption (or requirement) is the base of many statistical methods (e.g. the Naive Bayes Classifier) [1]. In practical applications of statistical modeling, however, the assumption may or may not be realistic. The measure of independence is needed but ignored.

Independence Measure. Random variables independence measures have been widely applied in recent machine learning literature, which plays an important role in independent component analysis (ICA), and feature selection [2]. Different from test of homogeneity (e.g. z-test), statistical test of independence, such as Spearman's ρ and Kendall's τ, determine whether one variable is associated with another. Spearman's rank correlation coefficient is non-parametric measure, which assesses how well the relationship between two variables can be described using a monotonic function [3]. It equal to the Pearson correlation between the rank values of two variables. A perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other. However, they are not guaranteed to detect all modes of independence. Moreover, the characteristic function-based tests [4] have been proposed, which are more general but limited to comparing univariate random variables.

In [2], Dr. Gretton et al. propose Hilbert Schmidt Independence Criterion (HSIC), which is a particular kernel independence measure. Similar with other statistical tests, given two random variables X and Y, the HSIC test is used to distinguish the null hypothesis $\mathcal{H}_0 : P_{XY} = P_X P_Y$, and the alternative hypothesis $\mathcal{H}_1 : P_{XY} \neq P_X P_Y$. This is achieved by comparing the HSIC with a particular threshold. If the threshold is exceeded, the test rejects the null hypothesis (bearing in mind that a zero population HSIC indicates $P_{XY} = P_X P_Y$).

*https://yurongchen1998.github.io

2 Hilbert Schmidt Independence Criterion

With the above background knowledge, we will discuss how the HSIC works in detail. HSIC is the Hilbert Schmidt norm of the cross-covariance operator between the distributions in Reproducing Kernel Hilbert Space (RKHS). As a beginner, we would like to understand this by dividing it into three parts: cross-covariance, Reproducing Kernel Hilbert Space, Hilbert Schmidt norm.

Cross-covariance. For two random variables X and Y, the cross-covariance matrix of X and Y is defined by [5]:

$$\mathbf{K}_{XY} = \text{cov}(\mathbf{X}, \mathbf{Y}) := \mathbb{E}[(\mathbf{X} - \mu_X)(\mathbf{Y} - \mu_Y)^T].$$

The $\mathbf{K}_{X_iY_j}$ entry of the cross-covariance matrix is the covariance:

$$\mathbf{K}_{X_iY_j} = \text{cov}[\mathbf{X}_i, \mathbf{Y}_j] = \mathbb{E}[(\mathbf{X}_i - \mu_{X_i})(\mathbf{Y}_j - \mu_{Y_j})].$$

The python implement can be seen [2]. Covariance values range from -inf to +inf where a positive value denotes that the two variables move in the same direction, vice versa.

Reproducing Kernel Hilbert Space. Firstly, a Hilbert Space can be defined as an inner product space that is complete and separable with respect to the norm defined by the inner product [6]. One of the most familiar examples of a Hilbert space is the Euclidean vector space consisting of three-dimensional vectors, denoted by $\mathbb{R}^3$, and equipped with the inner product:

$$< f, g > = \sum_i f_i \, g_i.$$

Now, defines a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as a kernel if there exists an $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$ [7]. A reproducing kernel Hilbert space (RKHS) is a **Hilbert space** of functions in which point evaluation is a continuous linear functional:

$$f(x) = f(\cdot)^T \phi(x) :=< f(\cdot), \phi(x) >_{\mathcal{H}} .$$

Hilbert Schmidt norm. The Hilbert-Schmidt norm of a matrix **A** is a matrix norm:

$$||A||_{HS} = \sqrt{\sum_i \sum_j a_{ij}^2}.$$

Hilbert Schmidt Independence Criterion. Let $\mathcal{D} := \{(x_1, y_1), \ ... \ , (x_m, y_m)\}$ contain m samples drawn from P_{xy}, where $x_i \in \mathbb{R}^{d1}$ and $y_i \in \mathbb{R}^{d2}$. Given two separable (having a complete orthonormal basis) RKHSs $\mathcal{F}$ and $\mathcal{G}$ with continuous feature mapping $\phi(x) \in \mathcal{F}$. $\psi(y) \in \mathcal{G}$ from each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, respectively. The cross-covariance $C_{xy} : \mathcal{F} \to \mathcal{G}$ can be written:

$$C_{xy} := \mathbb{E}_{xy}[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)],$$

where μ_x and μ_y is the mean of mapped features, i.e., $\phi(x)$ and $\psi(y)$, and $\otimes$ is the tensor product [8]:

$$C_{xy} := \mathbb{E}_{xy}[\phi(x) \otimes \psi(y)] - \mu_x \otimes \mu_y.$$

It is a generalization of the cross-covariance matrix between random vectors. More information of the HSIC can be seen in [2][8].

3 HSIC in Machine Learning

In [9], the HSIC is introduced as a loss function for learning robust representations. They consider the unsupervised covariate shift learning problem, in which (X, Y) be a pair of random variables such that X are the instances and Y are labels. The objective of [9] is to learn a model predicting Y from X that can work well on a different, a-prior unknown target distribution $P_{target}(X, Y)$, given a training distribution $P_{source}(X, Y)$. According to the probability product rule $P(X, Y) = P(Y|X)P(X)$, the domain-shift scenario can be assumed as $P_{source}(Y|X) = P_{target}(Y|X)$ while $P_{source}(X) \neq P_{target}(X)$. The core of [9] is to learn a model f such that $Y - f(X)$ is independent of the distribution of X. With the HSIC, the loss function can be formulated as:

$$\min \text{HSIC}(X, Y - f(X); \mathcal{F}, \mathcal{G}).$$

In processing.

[2] Covariance and Correlation

References

[1] https://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[2] Gretton A, Fukumizu K, Teo C H, et al. "A kernel statistical test of independence," in *NIPS*, 2007, 20: 585-592.

[3] https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.

[4] Kankainen, Annaliisa. "Consistent testing of total independence based on the empirical characteristic function". PhD thesis, 1995.

[5] G. John, *Probability and Random Processes for Electrical and Computer Engineers.* Cambridge University Press, pp. 336, 2006.

[6] https://people.eecs.berkeley.edu/ bartlett/courses/281b-sp08/7.pdf

[7] https://www.gatsby.ucl.ac.uk/ gretton/coursefiles/lecture4_introToRKHS.pdf

[8] A. Gretton, O. Bousquet, A.J. Smola, and B. Scholkop, " Measuring statistical dependence with Hilbert-Schmidt norm," in *ALT*, pp. 63–77, 2005.

[9] Greenfeld D, Shalit U, "Robust learning with the hilbert-schmidt independence criterion," in *International Conference on Machine Learning,* 2020: 3759-3768.