Probabilistic Multivariate Gaussian Distance for Uncertainty-Aware Learning

Mathis Reymond

April 2025

1 Introduction

The effectiveness KNN algorithm heavily relies on the distance metric chosen to quantify the similarity or dissimilarity between data points. While Euclidean distance is commonly used, it often falls short when dealing with data that inherently carries specific information that could be leveraged more efficiently. This document introduces a novel custom distance metric specifically designed for multivariate samples where each feature can be described by a Gaussian distribution, defined by its mean and variance. So, in our case, data carries information about its own uncertainty or spread.

Consider a scenario where each data point in our dataset is not a single vector of precise values, but rather a collection of observed means, each associated with its own standard deviation. In such cases, treating each data point as a Gaussian distribution—characterized by its mean and variance for each dimension—allows us to capture this inherent uncertainty.

The motivation for developing this custom distance stems from the need to move beyond traditional point-to-point distances, like Euclidean distance, which would not make a good use of the standard deviations. By considering the probabilistic nature of our samples, we aim to develop a distance for KNN algorithm that can make more informed decisions by understanding the overlap and separation of the underlying distributions.

2 Distance of a random variable to a fixed point

2.1 Random variable centered in 0

First, we are interested in the calculation for $\mathbb{E}[|X - d|]$ where X follows a Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma \in \mathbb{R}^+$. So, $X \sim N(0, \sigma^2)$.
The expectation is given by:

$$\mathbb{E}[|X - d|] = \int_{\mathbb{R}} |x - d| \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx$$

And by splitting the integral at $x = d$ we get

$$\begin{aligned}
\mathbb{E}[|X-d|] &= \int_d^{+\infty}(x-d)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}dx + \int_{-\infty}^{d}(d-x)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}dx \\
&= \frac{1}{\sigma\sqrt{2\pi}}\left[\int_d^{+\infty} xe^{-\frac{x^2}{2\sigma^2}}dx - d\int_d^{+\infty} e^{-\frac{x^2}{2\sigma^2}}dx + d\int_{-\infty}^{d} e^{-\frac{x^2}{2\sigma^2}}dx - \int_{-\infty}^{d} xe^{-\frac{x^2}{2\sigma^2}}dx\right] \\
&= \frac{1}{\sigma\sqrt{2\pi}}\left[\int_d^{+\infty} xe^{-\frac{x^2}{2\sigma^2}}dx - \int_{-\infty}^{d} xe^{-\frac{x^2}{2\sigma^2}}dx\right] + \frac{d}{\sigma\sqrt{2\pi}}\left[\int_{-\infty}^{d} e^{-\frac{x^2}{2\sigma^2}}dx - \int_d^{+\infty} e^{-\frac{x^2}{2\sigma^2}}dx\right]
\end{aligned}$$

Let's evaluate the integrals involving $xe^{-\frac{x^2}{2\sigma^2}}$: For $\int xe^{-\frac{x^2}{2\sigma^2}}dx$, let $u = -\frac{x^2}{2\sigma^2}$. Then $du = -\frac{2x}{2\sigma^2}dx = -\frac{x}{\sigma^2}dx$, so $xdx = -\sigma^2 du$. Thus, $\int xe^{-\frac{x^2}{2\sigma^2}}dx = \int e^u(-\sigma^2 du) = -\sigma^2 e^u = -\sigma^2 e^{-\frac{x^2}{2\sigma^2}}$.
Now, substitute this back:

$$\begin{aligned}
&\frac{1}{\sigma\sqrt{2\pi}}\left[\left[-\sigma^2 e^{-\frac{x^2}{2\sigma^2}}\right]_d^{+\infty} - \left[-\sigma^2 e^{-\frac{x^2}{2\sigma^2}}\right]_{-\infty}^{d}\right] \\
&= \frac{1}{\sigma\sqrt{2\pi}}\left[(0 - (-\sigma^2 e^{-\frac{d^2}{2\sigma^2}})) - (-\sigma^2 e^{-\frac{d^2}{2\sigma^2}} - 0)\right] \\
&= \frac{1}{\sigma\sqrt{2\pi}}\left[\sigma^2 e^{-\frac{d^2}{2\sigma^2}} + \sigma^2 e^{-\frac{d^2}{2\sigma^2}}\right] \\
&= \frac{1}{\sigma\sqrt{2\pi}}\left[2\sigma^2 e^{-\frac{d^2}{2\sigma^2}}\right] \\
&= \frac{2\sigma}{\sqrt{2\pi}}e^{-\frac{d^2}{2\sigma^2}} = \sigma\sqrt{\frac{2}{\pi}}e^{-\frac{d^2}{2\sigma^2}}
\end{aligned}$$

Next, let's evaluate the integrals involving $e^{-\frac{x^2}{2\sigma^2}}$. Recall that $\int_{-\infty}^{d}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}dx = \Phi\left(\frac{d}{\sigma}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. And $\int_d^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}dx = 1 - \Phi\left(\frac{d}{\sigma}\right)$.
So, the second part of the original expression becomes:

$$\begin{aligned}
&\frac{d}{\sigma\sqrt{2\pi}}\left[\int_{-\infty}^{d} e^{-\frac{x^2}{2\sigma^2}}dx - \int_d^{+\infty} e^{-\frac{x^2}{2\sigma^2}}dx\right] \\
&= d\left[\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{d} e^{-\frac{x^2}{2\sigma^2}}dx - \frac{1}{\sigma\sqrt{2\pi}}\int_d^{+\infty} e^{-\frac{x^2}{2\sigma^2}}dx\right] \\
&= d\left[\Phi\left(\frac{d}{\sigma}\right) - \left(1 - \Phi\left(\frac{d}{\sigma}\right)\right)\right] \\
&= d\left[2\Phi\left(\frac{d}{\sigma}\right) - 1\right]
\end{aligned}$$

Combining both parts, we get the final result:

$$\mathbb{E}[|X-d|] = \sigma\sqrt{\frac{2}{\pi}}e^{-\frac{d^2}{2\sigma^2}} + d\left(2\Phi\left(\frac{d}{\sigma}\right) - 1\right)$$

2.2 Random variable centered in μ

Now, let $Y \sim N(\mu, \sigma^2)$. To compute $\mathbb{E}[|Y - d|]$ we can note that if we let $X := Y - \mu$ then $X \sim N(0, \sigma^2)$ and

$$\begin{aligned}\mathbb{E}[|Y - d|] &= \mathbb{E}[|(X + \mu) - d|] \\ &= \mathbb{E}[|X - (d - \mu)|]\end{aligned}$$

And by substituting $(d - \mu)$ for d in the previous fomrula, we find

$$\boxed{\mathbb{E}[|Y - d|] = \sigma\sqrt{\frac{2}{\pi}}e^{-\frac{(d-\mu)^2}{2\sigma^2}} + (d - \mu)\left(2\Phi\left(\frac{(d - \mu)}{\sigma}\right) - 1\right)}$$

This is the full expression for the expected absolute deviation from a fixed point d for a Gaussian random variable $Y \sim N(\mu, \sigma^2)$.

3 Distance of a random variable to another

Now we want to compute $\mathbb{E}[|A - B|]$ where $A \sim N(\mu_A, \sigma_A^2)$ and $B \sim N(\mu_B, \sigma_B^2)$. We'll assume A and B are independent random variables.
Let $Z = A - B$. Since A and B are independent Gaussian random variables, their difference Z is also a Gaussian random variable and follows $N(\mu_Z, \sigma_Z^2)$ with $\mu_Z = \mu_A - \mu_B$ and $\sigma_Z^2 = \sigma_A^2 + \sigma_B^2$.
We are therefore looking for

$$\begin{aligned}\mathbb{E}[|A - B|] &= \mathbb{E}[|Z|] \\ &= \mathbb{E}[|Z - 0|] \\ &= \sigma_Z\sqrt{\frac{2}{\pi}}e^{-\frac{(0-\mu_Z)^2}{2\sigma_Z^2}} + (0 - \mu_Z)\left(2\Phi\left(\frac{0 - \mu_Z}{\sigma_Z}\right) - 1\right) && \text{c.f. previous result} \\ &= \sigma_Z\sqrt{\frac{2}{\pi}}e^{-\frac{\mu_Z^2}{2\sigma_Z^2}} + \mu_Z\left(2\Phi\left(\frac{\mu_Z}{\sigma_Z}\right) - 1\right) && \text{remove zeros}\end{aligned}$$

And by substituting μ_Z and σ_Z we get:

$$\boxed{\mathbb{E}[|A - B|] = e^{-\frac{(\mu_A-\mu_B)^2}{2(\sigma_A^2+\sigma_B^2)}} + (\mu_A - \mu_B)\left(2\Phi\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right) - 1\right)}$$

4 The multidimensional case

Now, each sample is multidimensional, meaning that it is made of several means, each one coming with it's specific standard deviation. Thus, computing the distance between two samples requires a bit more effort. So let's consider that A and B are multidimensional Gaussian vectors where each pair of coordinates is independent, and the two vectors themselves are independent, then we can calculate $\mathbb{E}[|A - B|]$. We use the absolute value norm for $|.|$ (which allow us to consider each dimension separately). This choice

is motivated by the fact that using the Euclidean norm here would make the calculation significantly more complex as it involves the expectation of the square root of a sum of squared Gaussian variables, which leads to a non-central Chi distribution.

Let $A = (A_1, \ldots, A_D)$ and $B = (B_1, \ldots, B_D)$ to be independent Gaussian random variables. Assume $A_i \sim N(\mu_{A_i}, \sigma^2_{A_i})$ and $B_i \sim N(\mu_{B_i}, \sigma^2_{B_i})$. Let $Z_i = A_i - B_i$. Then $Z_i \sim N(\mu_{Z_i}, \sigma^2_{Z_i})$ with $\mu_{Z_i} = \mu_{A_i} - \mu_{B_i}$ and $\sigma^2_{Z_i} = \sigma^2_{A_i} + \sigma^2_{B_i}$.

We can use linearity of expectation:

$$\begin{aligned}
\mathbb{E}\left[|A - B|\right] &= \mathbb{E}\left[\sum_{i=1}^{D} |A_i - B_i|\right] \\
&= \sum_{i=1}^{D} \mathbb{E}\left[|A_i - B_i|\right] \\
&= \sum_{i=1}^{D} \left(\sqrt{\sigma^2_{A_i} + \sigma^2_{B_i}} \sqrt{\frac{2}{\pi}} e^{-\frac{\left(\mu_{A_i} - \mu_{B_i}\right)^2}{2\left(\sigma^2_{A_i} + \sigma^2_{B_i}\right)}} + (\mu_{A_i} - \mu_{B_i}) \left(2\Phi \left(\frac{\mu_{A_i} - \mu_{B_i}}{\sqrt{\sigma^2_{A_i} + \sigma^2_{B_i}}} \right) - 1 \right) \right)
\end{aligned}$$

Therefore, the final distance we were looking for is given by the expectation expectation:

$$\boxed{\mathbb{E}\left[|A - B|\right] = \sum_{i=1}^{D} \left[\sqrt{\sigma^2_{A_i} + \sigma^2_{B_i}} \sqrt{\frac{2}{\pi}} e^{-\frac{\left(\mu_{A_i} - \mu_{B_i}\right)^2}{2\left(\sigma^2_{A_i} + \sigma^2_{B_i}\right)}} + (\mu_{A_i} - \mu_{B_i}) \left(2\Phi \left(\frac{\mu_{A_i} - \mu_{B_i}}{\sqrt{\sigma^2_{A_i} + \sigma^2_{B_i}}} \right) - 1 \right) \right]}$$

5 Applications

This custom distance metric offers several advantages over traditional distance measures for KNN in specific contexts. First, it explicitly accounts for the variability within each feature of a sample. This is particularly valuable when data points are not fixed values but statistical estimates. By considering the spread of distributions, the metric can be more robust to noise or small variations in the mean, as it understands the inherent "fuzziness" of the data points. And for datasets where the variance itself carries important information (e.g., distinguishing between stable and volatile measurements), this metric can provide better discriminative power than methods that only compare means.