Bayesian-Optimal Multi-Classification with Noisy Input Necessitates General Input Space Representation

Aman Bhargava

January 10, 2024

1 Introduction

Notation: lower case variables denote scalars (e.g., x), upper case variables denote random variables (e.g., X), and boldfaced variables denote vector quantities (e.g., $\mathbf{x}, \mathbf{X}$). We denote the $d \times d$ identity matrix as $\mathbf{I}_d$.

Here we analyze the latent representations in optimal Bayesian filter models trained to perform multi-classification on some ground truth input vector $\mathbf{x}^* \in \mathbb{R}^d$ based on noisy discrete-time measurement signals $\mathbf{X}(1), \ldots, \mathbf{X}(t)$ for i.i.d. $\mathbf{X}(i) = \mathbf{x}^* + \mathcal{N}(\mathbf{0}, \mathbf{I_d})$. Classification boundaries are denoted $\{(\mathbf{c}_i, b_i)\}_{i=0}^M$ where $\mathbf{c}_i \in \mathbb{R}^d$ is the normal vector to the hyperplane and $b_i \in \mathbb{R}$ is the offset from the origin. For boundary i, the separating plane defined as the affine set $\{\mathbf{x} | \mathbf{c}_i^\top \mathbf{x} = b_i\}$. The ground truth classification of a given point $\mathbf{x}$ is given by decision rule $y(\mathbf{x})$:

$$y_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{c}_i^\top \mathbf{x} > b_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

A restricted form of the problem is depicted in Figure 1. Note that we are interested in the general case where the hyperplanes do not necessarily pass through the origin.

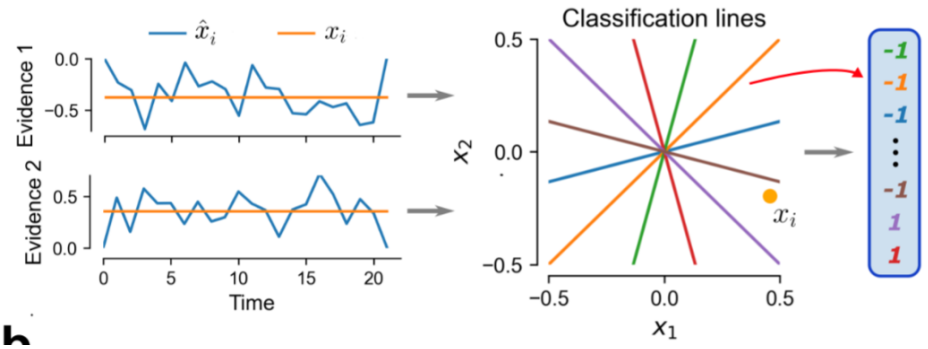

Figure 1: **Multitasking RNN learns abstract representations.** Data generating process. The task is to simultaneously report whether the true joint evidence (x_1, x_2) (yellow dot) lies above $(+1)$ or below (-1) a number of classification lines (here 6).

2 Bayesian Filtering Framework

We are interested in the properties of optimal Bayesian filter models that sequentially process inputs $\mathbf{X}(t)$ to estimate each classification $y_i(\mathbf{x}^*), i \in [M]$. Bayesian filters are a class of statistical models and algorithm that update a latent state based on noisy and uncertain observation signals. Rooted in principles of Bayesian inference, these filters combine aggregated "knowledge", represented by a latent state $\mathbf{Z}(t)$, with incoming observations $\mathbf{X}(t)$ to continually update the latent state to facilitate some prediction of some output $\mathbf{Y}(t) = f(\mathbf{Z}(t))$.

Definition 1 (Bayesian Filter Operation)**.** *A discrete-time Bayesian filter updates latent variable* $\mathbf{z}(t)$ *based on incoming data* $\mathbf{x}(t)$ *by applying Bayes' theorem:*

$$P\big(\mathbf{z}(t)|\mathbf{x}(t), \mathbf{z}(t-1)\big) = \frac{P\big(\mathbf{x}(t)|\mathbf{z}(t), \mathbf{z}(t-1)\big)P\big(\mathbf{z}(t)|\mathbf{z}(t-1)\big)}{P\big(\mathbf{x}(t)|\mathbf{z}(t-1)\big)} \tag{2}$$

$$\propto P\big(\mathbf{x}(t)|\mathbf{z}(t)\big)P\big(\mathbf{z}(t)|\mathbf{z}(t-1)\big) \tag{3}$$

Bayesian filters are commonly equipped with a "decoder" or "readout map" f *which maps latent* $\mathbf{Z}(t)$ *to readout estimation* $\hat{\mathbf{Y}}(t) = f(\mathbf{Z}(t))$.

The readout $\hat{\mathbf{Y}}(t)$ is a vector of Bernoulli random variables with index i corresponds to the estimate of classification $(\mathbf{c}_i, b_i)$.

3 Main Result

In this section we show that an optimal Bayesian filter performing multi-classification as described in Section 1 must store a representation of x^* in the latent variable $\mathbf{Z}(t)$. More specifically, we show that the maximum likelihood estimate of x^* based on $\mathbf{X}(1), \ldots, \mathbf{X}(t)$ is recoverable from $\mathbf{Z}(t)$.

3.1 Single Decision Boundary

First, we will consider the Bayesian filtering of a single classification output $y(x^*)$ from noisy inputs $X(1), \ldots, X(n)$ with decision boundary parameters $(\mathbf{c}, b)$. We begin by scaling our coordinates such that the Gaussian noise has unit variance:

$$\mathbf{X}(t) = \mathbf{x}^* + \mathcal{N}(\mathbf{0}, \mathbf{I_d}) \tag{4}$$

The problem is further simplified by setting $b = 0$. We are interested in $P(\hat{Y}(t)|\mathbf{X}(1), \ldots, \mathbf{X}(t))$ where $\hat{Y}(t)$ is a Bernoulli random variable representing the probability that $y(\mathbf{x}^*) = 1$ given evidence $\mathbf{X}(1), \ldots, \mathbf{X}(t)$.
Since $y(\mathbf{x}^*)$ is a deterministic function of non-random variable x^*, we will first derive the probability distribution over $\mathbf{x}^*$ (denoted $\hat{\mathbf{X}}$) to determine $y(\hat{\mathbf{X}})$.

Lemma 1. *For $\mathbf{X}(t) = \mathbf{x}^* + \mathcal{N}(\mathbf{0}, \mathbf{I_d})$ and no prior on x^*, the conditional probability distribution $\hat{\mathbf{X}} = P(x^*|\mathbf{X}(t), \ldots, \mathbf{X}(t))$ is given by*

$$\hat{\mathbf{X}} = \mathcal{N}(\hat{\mu}, \hat{\Sigma}) \tag{5}$$

where $\hat{\mu}$ is the mean of $\mathbf{X}(1), \ldots, \mathbf{X}(t)$ and $\hat{\Sigma} = t^{-1/2}\mathbf{I}_d$.

Proof. Since $\mathbf{X}(1), \ldots, \mathbf{X}(t)$ are i.i.d. from a Gaussian distribution with mean $\mathbf{x}^*$ and identity covariance, the sample mean is known to be distributed normally centered at the ground truth $\mathbf{x}^*$. We apply the known standard deviation of the underlying distribution (identity covariance) to arrive at $\hat{\Sigma} = t^{-1/2}\mathbf{I}_d$ as the standard deviation on the sample mean (derived from the central limit theorem). □

We can use estimator $\hat{\mathbf{X}}$ to construct $\hat{\mathbf{Y}}$ by expanding $\hat{\mathbf{Y}} = y(\hat{\mathbf{X}})$ as

$$y(\hat{\mathbf{X}}) = \begin{cases} 1 & \text{if } \mathbf{c}^\top \hat{\mathbf{X}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

In essence, we we are interested in the amount of the probability density of $\hat{\mathbf{X}}$ that lies on each side of the decision boundary. Deriving this probability is simplified by the fact that $\hat{\mathbf{X}}$ is isotropic – i.e., it inherits the spherical covariance of the underlying data generation process.

Lemma 2. *Consider Gaussian-distributed d-dimensional random variable $\hat{\mathbf{X}}$ with isotropic covariance $\Sigma = t^{-1/2}\mathbf{I}_d$ and mean $\mu \in \mathbb{R}^d$. The probability density of $\hat{\mathbf{X}}$ on each side of the positive side of the decision boundary $\{\mathbf{x} : \mathbf{c}^\top \mathbf{x} > 0\}$ can be expressed as*

$$\Pr\{\mathbf{c}^\top \mathbf{x} > 0\} = 1 - \Phi(k\sqrt{t}) \tag{7}$$

where Φ is the CDF of the normal distribution and $k = (\mathbf{c}^\top \mu)/\|\mathbf{c}\|$ is the signed projection distance between the plane and the mean of $\hat{\mathbf{X}}$.

Proof. Since the $\hat{\mathbf{X}}$ is isotropic, the variance on every axis is equal and independent. We may rotate our coordinate system such that the projection line between the plane and the mean of $\hat{\mathbf{X}}$ aligns with an axis we denote as "axis 0". The rest of the axes must be orthogonal to the plane. Since each component of an isotropic Gaussian is independent, the marginal distribution of $\hat{\mathbf{X}}$ over axis 0 is a □