Chapter 1 Human-Machine Ensemble Method

In this chapter, Section 1.1 describes the characteristics of forecasts by humans and a machine, and overview the proposed *human-machine ensemble method*. Then, I modeling the forecasts of humans and ones of machines in Section 1.2. Finally, Section 1.3 formulates the human-machine ensemble method as a problem to minimize the expected squared error of the ensemble, and solves it.

1.1 Overview

Traditional ensemble methods assumes that the error of each forecaster always follows a fixed probability distribution as mentioned in Section **??**. However, in fact, since each forecaster has good and weak points, the expected error of each changes depending on the situation. Let us consider economic forecasts. Since machines learn past patterns, they are good at a forecasting when the target index follows a pattern, but they are bad when given a pattern not in the past. On the other hand, humans can consider information such as economy and policies that machines can not lean well, so they are adaptable to patterns not in the past. Therefore, it is necessary to understand the characteristics of humans and machines for developing a human-machine ensemble method.

The characteristic of forecasts by machines is that their expected error can be quantified since they make forecasts statistically based on patterns they leaned. Let us explain by using an illustrative example. Figure 1.1 shows 200 data points belonging to one of two classes A and B. The classes A and B are sampled from the Gaussians with standard deviation $\sigma = 1$ centered at $(-2, 0)$ and $(2, 0)$, respectively. Assume that there is a prediction model that learned those samples. For instance, if this model is given an input $p = (-2, 0)$, which has the typical feature of class A, it returns that the probability of belonging to A is 0.95. Furthermore, given $q = (0.1, 2)$, which is located near the boundary of the two classes, it returns the probability is 0.45. Of course, humans also can answer the probability that p and q belong to class A based on the figure, but since they answer the probability intuitively, the grounds for the value is weak and the reliability is poor. Meanwhile, machines can return values with

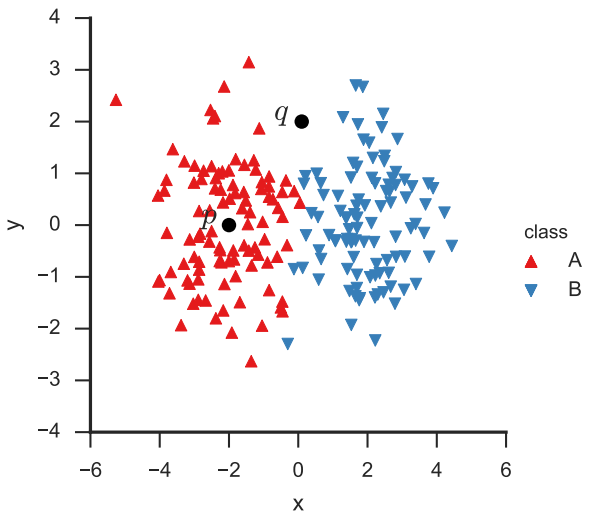

Figure 1.1: An illustrative example of a classification problem. There are 200 samples represented as points in a 2D feature space. Each of them belongs to a class A or B. Given a sample p as input to a prediction model that learned those samples as training set, the model outputs that the probability of belonging to A is high, while given a sample q, the probabilities of belonging to A and B are about the same.

grounds based on the concrete model.

On the other hand, the characteristic of forecasts by humans is that they can make forecasts considering information such as the current economy and new policies, which is difficult for machines to learn. In addition, forecasts by a group of humans can harness the *wisdom of the crowd* as mentioned in Section **??**. In economic forecasts, opinions are often different among experts, so it is easy to gather diverse forecasts by increasing the number of people. Conversely, it is difficult to construct diverse prediction models by machine learning. Even if the learning algorithm is different, the predicted values will be similar since the training set is the same. Algorithms such as *Random Forest* divide the training set and construct multiple weak learners, but they sacrifices the accuracy of each model.

The human-machine ensemble method combines forecasts of a machine and group of humans according to the expected error of the machine forecast. In other words, if the expected error of machine is small, the ensemble judge that the machine forecast alone is enough, otherwise it combines forecasts of humans. This procedure makes it possible to harness the characteristics of machines that can quantify the expected errors and the characteristics of humans that can compensate for the poor part of machines and increase diversity easily.

1.2 Model

In this section, I modeling forecasts of prediction models obtained by machine learning and forecasts of individual humans. They forecasts continuous values such as annual inflation rate.

1.2.1 Humans model

Forecasts of individual humans are modeled as random variables like Lamberson and Page [**?**] introduced in Section **??**. Let random variables h_i and ε_{h_i} denote the forecast and error of a human i.

Assume that the error ε_{h_i} follows a distribution with mean $\mu = 0$ and variance $\sigma^2 = \mathrm{var}(\varepsilon_h)$. This means that forecasts by humans are unbiased. That is, although a forecast can be smaller or greater than the actual value, the positive errors and the negative errors are the same amount as a whole, and the errors

are not positively biased or negatively biased.

Moreover, let $H(n)$ denote the average of forecasts by n humans, and let $\mathrm{cov}(\varepsilon_h)$ denote the average covariance in the errors of two different humans as

$$\mathrm{cov}(\varepsilon_h) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{cov}(\varepsilon_{h_i}, \varepsilon_{h_j}). \tag{1.1}$$

According to [**?**], the expected squared error of $H(n)$ is

$$\mathrm{MSE}(H(n)) = \frac{1}{n} \mathrm{var}(\varepsilon_h) + (1 - \frac{1}{n}) \mathrm{cov}(\varepsilon_h). \tag{1.2}$$

1.2.2 Machine model

Forecasts of machines are modeled as that the expected error varies according to the input. The inputs are, for example, inflation fluctuations over the past 12 months. Assume that a prediction model θ obtained by machine learning outputs a probability distribution to an input vector $\boldsymbol{x}$. This is regarded as the posterior distribution $f_\theta(y|\boldsymbol{x})$ for the target value y when given the input $\boldsymbol{x}$.

The forecast value $y_\theta(\boldsymbol{x})$ is the mean of the posterior distribution outputted by model θ given input $\boldsymbol{x}$:

$$y_\theta(\boldsymbol{x}) = \int_{-\infty}^{\infty} y f_\theta(y|\boldsymbol{x}) dy. \tag{1.3}$$

Let $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x})$ denote the variance of the distribution. This represents the expected squared error of the forecast value $y_\theta(\boldsymbol{x})$ from the definition.

Finally, let $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ denote the average covariance in errors between the machine and an arbitrary human. That is,

$$\mathrm{cov}(\varepsilon_\theta, \varepsilon_h) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{cov}(\varepsilon_\theta, \varepsilon_{h_i}). \tag{1.4}$$

Section **??** verify whether these models are applicable to actual inflation forecasts.

Implementation A prediction model that outputs probability distributions can be implemented by using *artificial neural networks (ANN)* approximately. Separate the possible target values into K discrete values $y_1, \ldots, y_K$, and regard the problem as *classification* to the closest value. Then, if the *softmax function* is used for the output layer of ANN, the prediction model can output

approximate discrete probability distributions. For example, when forecasting a percentage of annual inflation change, an output of the ANN model is that the probability of 0% is 0.02, 0.5% is 0.05, 1% is 0.33, ..., and 4% is 0.03. In this case, $y_\theta(\boldsymbol{x})$ is obtained by

$$y_\theta(\boldsymbol{x}) = \sum_{i=0}^{K} y_i f(y_i|\boldsymbol{x}). \tag{1.5}$$

Section **??** describes more detailed implementation.

For regression problems such as economic forecasts, a model that outputs a single value is often used. However, Rothe, Timofte and Gool [**?**] have reported a model that outputs probability distributions made more accurate predictions than a model that outputs single value in age estimation.

Probability distributions are also easier to interpret for humans than single values. Even if a single forecast value is outputted, its grounds and how reliable it is are unknown. Meanwhile, humans can interpret the expected error of the forecast from the distribution, and estimate how reliable the forecast is even if the grounds of the distribution is unknown.

1.3 Optimal Composition of Humans and a Machine

This section formulates a human-machine ensemble method as a problem to find the number of humans n that minimizes the expected error of the ensemble, and analyse the solution.

1.3.1 Problem formulation

The forecast value of a human-machine ensemble, $Y_{\theta,h}(n|\boldsymbol{x})$ is the average of a forecast by a machine $y_\theta(\boldsymbol{x})$ and forecasts by n humans $\boldsymbol{h} = (h_1, \ldots, h_n)$:

$$Y_{\theta,h}(n|\boldsymbol{x}) = \begin{cases} \dfrac{y_\theta(\boldsymbol{x}) + \sum_{i=1}^{n} h_i}{n+1} & (n \geq 1) \\ y_\theta(\boldsymbol{x}) & (n = 0). \end{cases} \tag{1.6}$$

The problem is obtaining n that minimizes the expected squared error of the ensemble $\mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x}))$. This is formulated as an optimization problem as follows:

$$\begin{aligned} \text{minimize} \quad & \mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x})) \\ \text{subject to} \quad & n \geq 0. \end{aligned} \tag{1.7}$$

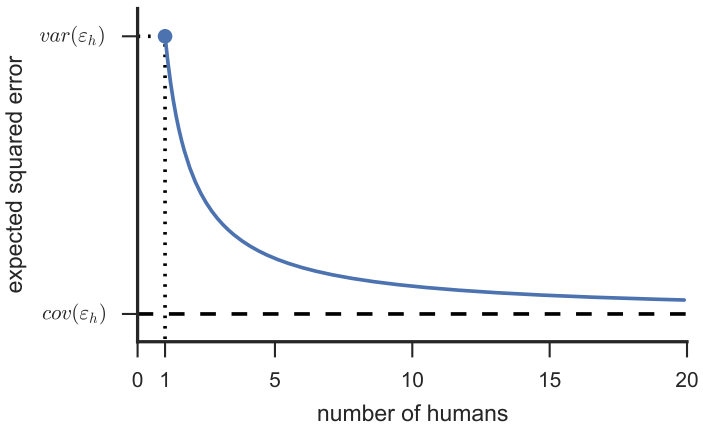

Figure 1.2: Graph of the expected squared error of the average of forecasts by n humans.

The expected squared error $\mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x}))$ can be derived from the results of Lamberson and Page [**?**] introduced in Section **??**. In equation (**??**), assumed that type a is human and type b is machine, $\mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x}))$ is

$$\mathrm{MSE}(Y_{\theta,h}(n|\boldsymbol{x})) = \frac{n\,\mathrm{var}(\varepsilon_h) + \mathrm{var}(\varepsilon_\theta|\boldsymbol{x}) + n(n-1)\,\mathrm{cov}(\varepsilon_h) + 2n\,\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)}{(n+1)^2}. \tag{1.8}$$

It is same as Lamberson and Page to obtain the optimal ratio of forecasters that have types. However, in this study, the expected error of a machine $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x})$ varies due to the input $\boldsymbol{x}$. The next section analyses how the optimal number of humans n is determined according to each parameter.

1.3.2 Theoretical results

First, let us examine the relationship between the expected squared errors of forecasts by a group of humans $H(n)$ and the number of humans n. From (1.2), the expected squared error takes $\mathrm{var}(\varepsilon_h)$ when $n = 1$ and approaches $\mathrm{cov}(\varepsilon_h)$ when n increases. From the definition, $\mathrm{var}(\varepsilon_h) \geq \mathrm{cov}(\varepsilon_h)$ holds, so (1.2) decreases monotonically with $n \geq 1$. Figure 1.2 shows the graph of $\mathrm{MSE}(H(n))$ when n is a variable.

The expected squared error of a human-machine ensemble also approaches $\mathrm{cov}(\varepsilon_h)$ when $n \to \infty$ from (1.8). And it takes $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x})$ when $n = 0$. However, whether $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x})$ or $\mathrm{cov}(\varepsilon_h)$ is greater is different depending on $\boldsymbol{x}$, and

sometimes (1.8) has a local minimum.

If (1.8) has a local minimum, the optimal number of humans n^* always exists. The condition to have a local minimum is

$$\mathrm{cov}(\varepsilon_\theta, \varepsilon_h) < \frac{3\,\mathrm{cov}(\varepsilon_h) - \mathrm{var}(\varepsilon_h)}{2}. \tag{1.9}$$

When taking the local minimum, n is

$$N^*(\boldsymbol{x}) = \frac{2var(\varepsilon_\theta|\boldsymbol{x}) - var(\varepsilon_h) + cov(\varepsilon_h) - 2cov(\varepsilon_\theta, \varepsilon_h)}{3cov(\varepsilon_h) - var(\varepsilon_h) - 2cov(\varepsilon_\theta, \varepsilon_h)}. \tag{1.10}$$

If $N^*(\boldsymbol{x})$ is greater than 0, the optimal n is $N^*(\boldsymbol{x})$, and if $N^*(\boldsymbol{x})$ is less than 0, the optimal n is 0. That is, n^* that minimizes the expected squared error when (1.8) has a local minimum is

$$n^* = \begin{cases} 0 & (N^*(\boldsymbol{x}) <= 0) \\ N^*(\boldsymbol{x}) & (N^*(\boldsymbol{x}) > 0). \end{cases} \tag{1.11}$$

If (1.9) does not hold, the optimal n does not always exist. In this case, the optimal n is 0 when $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x}) \leq \mathrm{cov}(\varepsilon_h)$, and the optimal n does not exist when $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x}) > \mathrm{cov}(\varepsilon_h)$. When n^* does not exist, you can make the expected squared error close to $\mathrm{cov}(\varepsilon_h)$ by increasing n.

Figure 1.3 shows four graphs that have different optimal n. These are grasped by adding a machine forecast to Figure 1.2 as the case of $n = 0$. The parameters related to machine forecasts $\mathrm{var}(\varepsilon_\theta|\boldsymbol{x})$ and $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ influence the values near $n = 0$. In particular, due to (1.9), if $\mathrm{cov}(\varepsilon_\theta, \varepsilon_h)$ is sufficient smaller than the parameters related only to humans, that is, if the forecasts between a machine and humans are sufficiently different, (1.8) has local minimum.

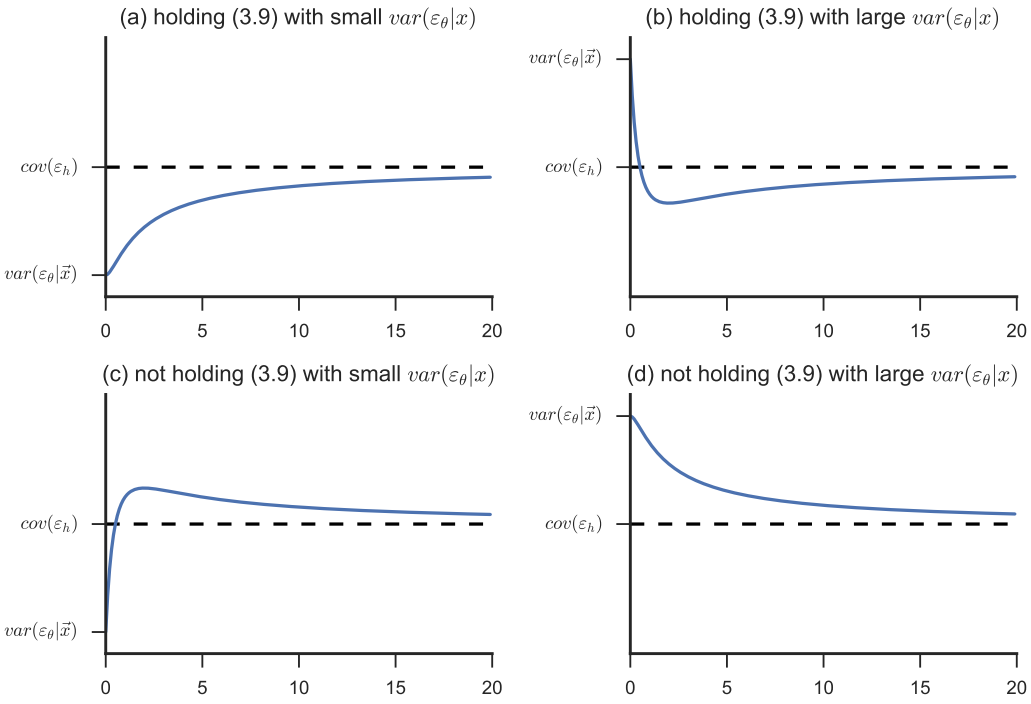

Figure 1.3: Approximate form of graphs about the expected squared errors of ensembles with n as a variable. The upper two are cases where the condition (1.9) holds, and $n^* = 0$ in (a) and $n^* = N^*(\boldsymbol{x})$ in (b). The lower two are cases where the condition (1.9) does not hold, and $n^* = 0$ in (c) and n^* does not exist in (d).