

# Ensembles

Combining continuous-valued outputs

---

**Javier Sáez**

January 18, 2023

**Universidad de Granada**

Generic Formulation

Equivalences

Generalized Mean Combiner

Theoretical Comparison of Simple combiners

Theoretical framework for the product combiner

## Notation

- $D_i : \mathbb{R}^n \rightarrow [0, 1]^c$  is a classifier.
- $d_{i,j}(\mathbf{x})$  represents the support (estimation of the posterior) that  $D_i$  gives to the hypothesis that  $\mathbf{x}$  comes from class  $\omega_j$ .
- We can build

$$DP(\mathbf{x}) = \begin{pmatrix} d_{1,1}(\mathbf{x}) & \cdots & d_{1,c}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ d_{L,1}(\mathbf{x}) & \cdots & d_{L,c}(\mathbf{x}) \end{pmatrix}.$$

- And calculate a degree of support for class  $\omega_j$ :

$$\mu_j(\mathbf{x}) = \mathcal{F}(d_{1,j}(\mathbf{x}), \cdots, d_{L,j}(\mathbf{x})).$$

We label  $\mathbf{x}$  as the class with the largest support.

# Generic Formulation

---

## Simple non-trainable combiners

The most common combiners are:

- Average (sum)

$$\mu_j(\mathbf{x}) = \frac{1}{L} \sum_{i=1}^L d_{i,j}(\mathbf{x})$$

- Max/min/median of the

$$\mu_j(\mathbf{x}) = \max_i \{d_{i,j}(\mathbf{x})\}.$$

- Trimmed mean: sort  $d_{i,j}$  and remove  $K/2\%$  from each side, and then compute  $\mu_j(\mathbf{x})$  as the average of the rest.
- Product combiner:

$$\mu_j(\mathbf{x}) = \prod_{i=1}^L d_{i,j}(\mathbf{x})$$

All these are non trainable.

## Equivalence between Min and Max for 2 classes

### Proposition

Let  $a_1, \dots, a_L$  be the outputs for class  $\omega_1$  and  $1 - a_1, \dots, 1 - a_L$  the outputs for class  $\omega_2$ ,  $a_i \in [0, 1]$ . Then, the class label assigned to  $\mathbf{x}$  by the MAX and MIN combination rules is the same.

# Equivalence between Min and Max for 2 classes

## Proposition

Let  $a_1, \dots, a_L$  be the outputs for class  $\omega_1$  and  $1 - a_1, \dots, 1 - a_L$  the outputs for class  $\omega_2$ ,  $a_i \in [0, 1]$ . Then, the class label assigned to  $\mathbf{x}$  by the MAX and MIN combination rules is the same.

Assume  $a_1 = \min_i a_i$  and  $a_L = \max_i a_i$ .

- Minimum will choose  $\mu_1(\mathbf{x}) = a_1$  and  $\mu_2(\mathbf{x}) = 1 - a_L$
- Maximum will choose  $\mu_1(\mathbf{x}) = a_L$  and  $\mu_2(\mathbf{x}) = 1 - a_1$

# Equivalence between Min and Max for 2 classes

## Proposition

Let  $a_1, \dots, a_L$  be the outputs for class  $\omega_1$  and  $1 - a_1, \dots, 1 - a_L$  the outputs for class  $\omega_2$ ,  $a_i \in [0, 1]$ . Then, the class label assigned to  $\mathbf{x}$  by the MAX and MIN combination rules is the same.

Assume  $a_1 = \min_i a_i$  and  $a_L = \max_i a_i$ .

- Minimum will choose  $\mu_1(\mathbf{x}) = a_1$  and  $\mu_2(\mathbf{x}) = 1 - a_L$
- Maximum will choose  $\mu_1(\mathbf{x}) = a_L$  and  $\mu_2(\mathbf{x}) = 1 - a_1$

Then,

- $a_1 > 1 - a_L \implies a_L < 1 - a_1$  both methods choose  $\omega_1$
- $a_1 < 1 - a_L \implies a_L < 1 - a_1$  both methods choose  $\omega_2$
- $a_1 = 1 - a_L$  both methods choose randomly



# Equivalence Between Majority Vote and Median for 2 classes

## Proposition

Let  $L$  be an odd natural number. Let  $a_1, \dots, a_L$  be the outputs for class  $\omega_1$  and  $1 - a_1, \dots, 1 - a_L$  the outputs for class  $\omega_2$ ,  $a_i \in [0, 1]$ . Then, the class label assigned to  $\mathbf{x}$  by the Majority Vote and Median combination rules is the same.

# Equivalence Between Majority Vote and Median for 2 classes

## Proposition

Let  $L$  be an odd natural number. Let  $a_1, \dots, a_L$  be the outputs for class  $\omega_1$  and  $1 - a_1, \dots, 1 - a_L$  the outputs for class  $\omega_2$ ,  $a_i \in [0, 1]$ . Then, the class label assigned to  $\mathbf{x}$  by the Majority Vote and Median combination rules is the same.

Assume  $a_1 = \min_i a_i$  and  $a_L = \max_i a_i$ . The median of the outputs for class  $\omega_1$  is  $a_{\frac{L+1}{2}} = m$ .

- If  $m > 0.5$ , the median of the outputs for  $\omega_2$  is  $1 - m < 0.5$ , so  $\omega_1$  is selected. This means that at least  $\frac{L+1}{2}$  posterior probabilities for  $\omega_1$  were greater than 0.5, so the majority vote also assigns  $\omega_1$  to  $\mathbf{x}$
- If  $m < 0.5$ , then  $1 - m > 0.5$  and the median assigns  $\omega_2$ , but also  $\frac{L+1}{2}$  posteriors for  $\omega_2$  are greater than 0.5, so Majority vote also assigns  $\omega_2$  to  $\mathbf{x}$

## Generalized Mean Combiner

---

## Definition

The **generalized mean combiner** assigns to each class  $\omega_j$  the support:

$$\mu_j(\mathbf{x}, \alpha) = \left( \frac{1}{L} \sum_{i=1}^L d_{i,j}(\mathbf{x})^\alpha \right)^{\frac{1}{\alpha}}$$

## Definition

The **generalized mean combiner** assigns to each class  $\omega_j$  the support:

$$\mu_j(\mathbf{x}, \alpha) = \left( \frac{1}{L} \sum_{i=1}^L d_{i,j}(\mathbf{x})^\alpha \right)^{\frac{1}{\alpha}}$$

Of course,  $\alpha$  can be trained!

## Special cases

- $\alpha \rightarrow \infty \implies \mu_j(\mathbf{x}, \alpha)$  is the maximum combiner.

## Special cases

- $\alpha \rightarrow \infty \implies \mu_j(\mathbf{x}, \alpha)$  is the maximum combiner.

Let  $d_*$  be the maximum of  $d_{i,j}$  for  $i = 1, \dots, L$ .

$$\begin{aligned}\lim_{\alpha \rightarrow \infty} \log \mu_j(\mathbf{x}, \alpha) &= \lim_{\alpha \rightarrow \infty} \frac{1}{\alpha} \log \frac{\sum_{i=1}^L d_{i,j}^\alpha}{L} \\ &= \log d_* + \lim_{\alpha \rightarrow \infty} \frac{1}{\alpha} \log \frac{\sum_{i=1}^L \left(\frac{d_{i,j}}{d_*}\right)^\alpha}{L} \\ &= \log d_*\end{aligned}$$

## Special cases

- $\alpha \rightarrow \infty \implies \mu_j(\mathbf{x}, \alpha)$  is the maximum combiner.
- $\alpha = 1 \implies \mu_j(\mathbf{x}, \alpha) = \frac{1}{L} \sum_{i=1}^L d_{i,j}$  is the arithmetic mean.
- $\alpha \rightarrow 0 \implies \mu_j(\mathbf{x}, \alpha) = \left( \prod_{i=1}^L d_{i,j} \right)^{\frac{1}{L}}$  is the geometric mean.
- $\alpha = -1 \implies \mu_j(\mathbf{x}, -1) = \left( \frac{1}{L} \sum_{i=1}^L \frac{1}{d_{i,j}(\mathbf{x})} \right)^{-1}$  is the harmonic mean.
- $\alpha \rightarrow -\infty \implies \mu_j(\mathbf{x}, \alpha)$  is the minimum combiner.



## Special cases

- $\alpha \rightarrow \infty \implies \mu_j(\mathbf{x}, \alpha)$  is the maximum combiner.
- $\alpha = 1 \implies \mu_j(\mathbf{x}, \alpha) = \frac{1}{L} \sum_{i=1}^L d_{i,j}$  is the arithmetic mean.
- $\alpha \rightarrow 0 \implies \mu_j(\mathbf{x}, \alpha) = \left( \prod_{i=1}^L d_{i,j} \right)^{\frac{1}{L}}$  is the geometric mean.
- $\alpha = -1 \implies \mu_j(\mathbf{x}, -1) = \left( \frac{1}{L} \sum_{i=1}^L \frac{1}{d_{i,j}(\mathbf{x})} \right)^{-1}$  is the harmonic mean.
- $\alpha \rightarrow -\infty \implies \mu_j(\mathbf{x}, \alpha)$  is the minimum combiner.

$\alpha$  can then be understood as the level of *optimism* of the combiner:

- $\alpha \rightarrow -\infty$  is the most pessimistic, implying that **all** the classifiers must agree with the choice (minimum combiner).
- $\alpha \rightarrow \infty$  is the most optimistic (maximum combiner), where **at least one** of the classifiers supports  $\omega_j$ .

# **Theoretical Comparison of Simple combiners**

---

Let us consider the following scenario:

- There are only two classes  $\Omega = \{\omega_1, \omega_2\}$ .
- We assume that  $d_{j,i}(\mathbf{x})$  is an estimate of the posterior  $p(\omega_i | \mathbf{x})$  produced by the classifier  $D_j$  and that, for any  $\mathbf{x}$ ,

$$d_{j,1}(\mathbf{x}) + d_{j,2}(\mathbf{x}) = 1.$$

- We assume without loss of generality that the true posterior probability is

$$p(\omega_1 | \mathbf{x}) = p > 0.5,$$

so Bayes-optimal class label for  $\mathbf{x}$  is  $\omega_1$  (and assigning  $\omega_2$  is a classification error).

# Assumption

The classifiers commit i.i.d. errors in estimating  $p(\omega_1 \mid \mathbf{x})$  such that:

$$P_j \equiv d_{j,1}(\mathbf{x}) = p(\omega_1 \mid \mathbf{x}) + \eta(\mathbf{x}) = p + \eta(\mathbf{x})$$

and

$$d_{j,2}(\mathbf{x}) = 1 - p - \eta(\mathbf{x}),$$

where  $\eta(\mathbf{x})$  is often:

- a Gaussian distribution  $\eta(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2)$  or
- a continuous uniform distribution  $\eta(\mathbf{x}) \sim U([-b, b])$ .

## Assumption

The classifiers commit i.i.d. errors in estimating  $p(\omega_1 \mid \mathbf{x})$  such that:

$$P_j \equiv d_{j,1}(\mathbf{x}) = p(\omega_1 \mid \mathbf{x}) + \eta(\mathbf{x}) = p + \eta(\mathbf{x})$$

and

$$d_{j,2}(\mathbf{x}) = 1 - p - \eta(\mathbf{x}),$$

where  $\eta(\mathbf{x})$  is often:

- a Gaussian distribution  $\eta(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2)$  or
- a continuous uniform distribution  $\eta(\mathbf{x}) \sim U([-b, b])$ .

Thus,  $d_{j,i}(\mathbf{x})$  are random variables. Using the fusion method  $\mathcal{F}$ , the posterior estimates are:

$$\hat{P}_1 = \mathcal{F}(P_1, \dots, P_L), \quad \hat{P}_2 = \mathcal{F}(1 - P_1, \dots, 1 - P_L)$$

- For the single classifier, average and median fusion models  $\hat{P}_1 + \hat{P}_2 = 1$ . Thus, if  $\omega_1$  is the correct label, it is sufficient to have  $\hat{P}_1 > 0.5$  to label  $\mathbf{x}$  as  $\omega_1$ . The probability of error is then:

$$P_e = P(\hat{P}_1 \leq 0.5) = F_{\hat{P}_1}(0.5) = \int_0^{0.5} f_{\hat{P}_1}(y) dy$$

- For the single classifier, average and median fusion models  $\hat{P}_1 + \hat{P}_2 = 1$ . Thus, if  $\omega_1$  is the correct label, it is sufficient to have  $\hat{P}_1 > 0.5$  to label  $\mathbf{x}$  as  $\omega_1$ . The probability of error is then:

$$P_e = P(\hat{P}_1 \leq 0.5) = F_{\hat{P}_1}(0.5) = \int_0^{0.5} f_{\hat{P}_1}(y) dy$$

- For the minimum and maximum rules,  $\hat{P}_1 + \hat{P}_2 \neq 1$  necessarily. An error occurs if  $\hat{P}_1 \leq \hat{P}_2$ :

$$P_e = P(\hat{P}_1 \leq \hat{P}_2).$$

Using a Gaussian distribution,  $\hat{P}_1 \sim \mathcal{N}(p, \sigma^2)$ . Denoting by  $\Phi(z)$  to the C.D.F. of the  $\mathcal{N}(0, 1)$ , then

$$F(t) = \Phi\left(\frac{t - p}{\sigma}\right).$$



Using a Gaussian distribution,  $\hat{P}_1 \sim \mathcal{N}(p, \sigma^2)$ . Denoting by  $\Phi(z)$  to the C.D.F. of the  $\mathcal{N}(0, 1)$ , then

$$F(t) = \Phi\left(\frac{t - p}{\sigma}\right).$$

Since  $F_{\hat{P}_1}(t) = F(t)$ , **individual error** of a classifier is:

$$P_e = \Phi\left(\frac{0.5 - p}{\sigma}\right).$$

Given that  $\hat{P}_1 = \frac{1}{L} \sum_{j=1}^L P_j$ , since  $P_j$  are normally distributed and independent, then  $\hat{P}_1 \sim \mathcal{N}\left(p, \frac{\sigma^2}{L}\right)$ . Hence, the probability of error in this case is:

$$P_e = P(\hat{P}_1 < 0.5) = \Phi\left(\frac{\sqrt{L}(0.5 - p)}{\sigma}\right)$$

In the median fusion method:

$$\hat{P}_1 = \text{med}\{P_1, \dots, P_L\} = p + \text{med}\{\eta_1, \dots, \eta_L\} = p + \eta_m.$$

In the median fusion method:

$$\hat{P}_1 = \text{med}\{P_1, \dots, P_L\} = p + \text{med}\{\eta_1, \dots, \eta_L\} = p + \eta_m.$$

The probability of error is:

$$P_e = P(p + \eta_m < 0.5) = P(\eta_m < 0.5 - p) = F_{\eta_m}(0.5 - p)$$

In the median fusion method:

$$\hat{P}_1 = \text{med}\{P_1, \dots, P_L\} = p + \text{med}\{\eta_1, \dots, \eta_L\} = p + \eta_m.$$

The probability of error is:

$$P_e = P(p + \eta_m < 0.5) = P(\eta_m < 0.5 - p) = F_{\eta_m}(0.5 - p)$$

From order statistics theory [Mood et al., 1973]

$$F_{\eta_m}(t) = \sum_{j=\frac{L+1}{2}}^L \binom{L}{j} F_{\eta}(t)^j [1 - F_{\eta}(t)]^{L-j}$$

## Theorem (Theorem 11, page 252 Mood et al. [1973])

Let  $Y_1 \leq Y_2 \leq \dots \leq Y_n$  represent the order statistics from a c.d.f.  $F$ . The marginal c.d.f. of  $Y_\alpha$  is:

$$F_{Y_\alpha}(y) = \sum_{j=\alpha}^n \binom{n}{j} F_\eta(t)^j [1 - F_\eta(t)]^{n-j}$$

Proof: For a fixed  $y$ , let  $Z_i = I_{(-\infty, y)}(X_i)$ . Then  $\sum_{i=1}^n Z_i$  is the number of  $X_i$  that are lesser or equal than  $y$ . Note that  $\sum_i Z_i$  follows a binomial distribution with parameters  $n$  and  $F(y)$ . Thus,

$$F_{Y_\alpha}(y) = P[Y_\alpha \leq y] = P\left[\sum_i Z_i \geq \alpha\right] = \sum_{j=\alpha}^n \binom{n}{j} F_\eta(t)^j [1 - F_\eta(t)]^{n-j}$$

Using that  $\eta$  follows a Gaussian distribution, the probability of error is:

$$P_e = \sum_{j=\frac{L+1}{2}}^L \binom{L}{j} \Phi\left(\frac{0.5-p}{\sigma}\right)^j \left[1 - \Phi\left(\frac{0.5-p}{\sigma}\right)^j\right]^{L-j}$$

## **Theoretical framework for the product combiner**

---



# Assumptions

- We can consider each row of  $DP(\mathbf{x})$  as a prior on the set of class labels, and  $d_{i,j}$  as the estimate of  $p(\omega_j \mid \mathbf{x}, D_i)$ .
- Let  $P_{(i)} = (d_{i,1}(\mathbf{x}), \dots, d_{i,c}(\mathbf{x}))$ .

# Assumptions

- We can consider each row of  $DP(\mathbf{x})$  as a prior on the set of class labels, and  $d_{i,j}$  as the estimate of  $p(\omega_j \mid \mathbf{x}, D_i)$ .
- Let  $P_{(i)} = (d_{i,1}(\mathbf{x}), \dots, d_{i,c}(\mathbf{x}))$ .
- **Hypothesis:** The true values of  $p(\omega_i \mid \mathbf{x})$  are the ones most agreed upon the ensemble.
- We denote the agreed values by  $P_{ens} = (\mu_1(\mathbf{x}), \dots, \mu_c(\mathbf{x}))$ .

# Assumptions

- We can consider each row of  $DP(\mathbf{x})$  as a prior on the set of class labels, and  $d_{i,j}$  as the estimate of  $p(\omega_j \mid \mathbf{x}, D_i)$ .
- Let  $P_{(i)} = (d_{i,1}(\mathbf{x}), \dots, d_{i,c}(\mathbf{x}))$ .
- **Hypothesis:** The true values of  $p(\omega_i \mid \mathbf{x})$  are the ones most agreed upon the ensemble.
- We denote the agreed values by  $P_{ens} = (\mu_1(\mathbf{x}), \dots, \mu_c(\mathbf{x}))$ .

The average KL divergence across the  $L$  members is:

$$KL_{av} = \frac{1}{L} \sum_{i=1}^L KL(P_{ens} \parallel P_{(i)})$$

## Problem

$$\min_{P_{ens}} KL_{av} \quad s.t. \quad \sum_{k=1}^c \mu_k = 1$$

## Problem

$$\min_{P_{ens}} KL_{av} \quad s.t. \quad \sum_{k=1}^c \mu_k = 1$$

We solve it using its Lagrangian:

$$\mathcal{L}(\{\mu_i\}_{i=1}^c, \lambda) = KL_{av} + \lambda \left( 1 + \sum_{k=1}^c \mu_k \right)$$

Recalling that

$$KL(p \parallel q) = \sum_x p(x) \log_2 \left( \frac{p(x)}{q(x)} \right),$$

# Minimization

Recalling that

$$KL(p \parallel q) = \sum_x p(x) \log_2 \left( \frac{p(x)}{q(x)} \right),$$

We derivate  $\mathcal{L}$  w.r.t. each of its parameters to find its minimum:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu_j} &= \frac{\partial}{\partial \mu_j} \left[ KL_{av} + \lambda \left( 1 - \sum_{k=1}^c \mu_k \right) \right] \\ &= \frac{1}{L} \sum_{i=1}^L \frac{\partial}{\partial \mu_j} \left[ \sum_{k=1}^c \mu_k \log_2 \left( \frac{\mu_k}{d_{i,k}} \right) \right] - \lambda \\ &= \frac{1}{L} \sum_{i=1}^L \left( \log_2 \left( \frac{\mu_j}{d_{i,j}} \right) + C \right) - \lambda = 0, \end{aligned}$$

with  $C = \frac{1}{\ln(2)}$ .

Solving for  $\mu_j$ , we obtain:

$$\mu_j = 2^{(\lambda-C)} \prod_{i=1}^L (d_{i,j})^{1/L}.$$

Using in that expression our problem constraint, we obtain:

$$\lambda = C - \log_2 \left( \sum_{k=1}^c \prod_{i=1}^L (d_{i,k})^{1/L} \right),$$



## Minimization

Solving for  $\mu_j$ , we obtain:

$$\mu_j = 2^{(\lambda-C)} \prod_{i=1}^L (d_{i,j})^{1/L}.$$

Using in that expression our problem constraint, we obtain:

$$\lambda = C - \log_2 \left( \sum_{k=1}^c \prod_{i=1}^L (d_{i,k})^{1/L} \right),$$

which leads to the final expression:

$$\mu_j = \frac{\prod_{i=1}^L (d_{i,j})^{1/L}}{\sum_{k=1}^c \prod_{i=1}^L (d_{i,k})^{1/L}}.$$

which is **essentially** the **product rule**.

- Remark: the previous hypothesis was that  $P_{ens}$  (vector of posterior probabilities) are the ones most agreed upon by the ensemble.

- Remark: the previous hypothesis was that  $P_{ens}$  (vector of posterior probabilities) are the ones most agreed upon by the ensemble.
- If instead we wrote

$$KL(P_{(i)} \parallel P_{ens}),$$

and we repeat the process we obtain that  $\mu_j$  is the average combiner.

- Remark: the previous hypothesis was that  $P_{ens}$  (vector of posterior probabilities) are the ones most agreed upon by the ensemble.
- If instead we wrote

$$KL(P_{(i)} \parallel P_{ens}),$$

and we repeat the process we obtain that  $\mu_j$  is the average combiner.

- The hypothesis in this case is different:  $P_{ens}$  is an unknown prior which needs to be transformed to  $L$  posteriors (one for each of the classifiers in the ensemble).

Thank you for your attention

## References

---

Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley Publishing, 2nd edition, 2014. ISBN 1118315235.

A.M. Mood, F.A. Graybill, and D.C. Boes. *Introduction to the Theory of Statistics*. International Student edition. McGraw-Hill, 1973. ISBN 9780070428645. URL <https://books.google.es/books?id=Viu2AAAAIAAJ>.