



Why

Since our predictions are often uncertain, we can use the language of probability distributions to characterize them.¹

Definition

Denote the set of probability distributions on a set X by $\Delta(X)$.

A *probabilistic classifier* $G : A \rightarrow \Delta(B)$ is a function from inputs A to probability distributions over the classes B .

Given an input a , the *prediction* of G on a is a probability distribution $\hat{p}_a = G(a)$ on B .

Point classifier as probabilistic classifier

Given a point classifier $f : A \rightarrow B$, we can define a probabilistic classifier $G : A \rightarrow \Delta(B)$ corresponding to f by

$$\hat{p}_a(b) = \begin{cases} 1 & \text{if } f(a) = b \\ 0 & \text{otherwise.} \end{cases}$$

where $\hat{p}_a = G(a)$.

Probabilistic classifier from point classifier

On the other hand, given probabilistic classifier $G : A \rightarrow \Delta(B)$, we can define a point classifier $f : A \rightarrow B$ by

$$f(a) = \operatorname{argmax}_{b \in B} \hat{p}_a(b)$$

We call f the *maximum likelihood classifier* corresponding to G . If there are ties, we can order the (finite) set B arbitrarily, and break ties accordingly.

We can extend this idea, and define a list classifier by sorting the outputs by their probability, from largest to smallest.

¹Future editions will improve this.

