9.2. Lack of inherent superiority of any classifier

Ana I. González Acuña

from *Pattern Recognition* book (Duda, Hart, Stork 2000)

17/02/2012

Contents

Introduction

Questions:

- If we are interested solely in the generalization performance, are there any reasons to prefer one classifier or learning algorithm over another?
- If we make no prior assumptions about the nature of the classification task, can we expect any classification method to be superior or inferior overall?
- Can we even find an algorithm that is overall superior to (or inferior to) random guessing?

No Free Lunch Theorem

No Free Lunch Theorem answers: ***NO***

No **pattern classification method** is inherently superior to any other, or even to random guessing

Nature of the problem, prior distribution, data distribution, amount of training data, cost or reward functions, ... are the aspects that determine which form of classifier should provide the best performance.

The ***off-training set error*** (expected value)— the error on points not in the training set — is a good measure for distinguishing algorithms.

No Free Lunch Theorem shows that in the absence of assumptions we should not prefer any learning or classification algorithm over another

Theorem 9.1 (No Free Lunch) *For any two learning algorithms $P_1(h|\mathcal{D})$ and $P_2(h|\mathcal{D})$, the following are true, independent of the sampling distribution $P(\mathbf{x})$ and the number n of training points:*

1. *Uniformly averaged over all target functions F, $\mathcal{E}_1(E|F,n) - \mathcal{E}_2(E|F,n) = 0$;*
2. *For any fixed training set $\mathcal{D}$, uniformly averaged over F, $\mathcal{E}_1(E|F,\mathcal{D}) - \mathcal{E}_2(E|F,\mathcal{D}) = 0$;*
3. *Uniformly averaged over all priors $P(F)$, $\mathcal{E}_1(E|n) - \mathcal{E}_2(E|n) = 0$;*
4. *For any fixed training set $\mathcal{D}$, uniformly averaged over $P(F)$, $\mathcal{E}_1(E|\mathcal{D}) - \mathcal{E}_2(E|\mathcal{D}) = 0$.**

*Ugly Ducking Theorem

In the absence of prior information, is there a principled reason to judge any two distinct patterns as more or less similar than two other distinct patterns?

The *Ugly Duckling Theorem* states that in the absence of assumptions there is no privileged or "best" **feature representation**, and that even the notion of similarity between patterns depends implicitly on assumptions which may or may not be correct.

Find a principled measure the similarity between two patterns, given some representation: the number of predicates (rather than the number of features) the patterns share.

The Theorem forces us to acknowledge that even the apparently simple notion of similarity between patterns is fundamentally based on implicit assumptions about the problem domain

Theorem 9.2 (Ugly Duckling) *Given that we use a finite set of predicates that enables us to distinguish any two patterns under consideration, the number of predicates shared by any two such patterns is constant and independent of the choice of those patterns. Furthermore, if pattern similarity is based on the total number of predicates shared by two patterns, then any two patterns are "equally similar."* *

Minimum Description Length

Algorithmic complexity — also known as Kolmogorov complexity, algorithmic entropy, ... — seeks to quantify an inherent complexity of a binary string (we shall assume both classifiers and patterns are described by such strings).

The *minimum description length* (MDL) principle states that we should minimize the sum of the model's algorithmic complexity and the description of the training data D with respect to that model, i.e.,

$$K(h, \mathcal{D}) = K(h) + K(\mathcal{D} \text{ using } h). \tag{8}$$

Thus we seek the model $h*$ that obeys $h^* = \arg\min_h K(h, D)$

Minimum Description Length

It can be shown theoretically that classifiers designed with a minimum description length principle are guaranteed to converge to the ideal or true model *in the limit of more and more data*.

However, such derivations cannot prove that the principle leads to superior performance in the *finite data case*; to do so would violate the No Free Lunch Theorems.

The minimum description length principle states that simple models (small $K(h)$) are to be preferred, and thus amounts to a bias toward "*simplicity*".

It is found empirically that classifiers designed using the minimum description length principle work well in many problems.

Overfitting avoidance and Occam's razor

To avoid overfitting can be applied regularization, pruning, inclusion of penalty terms, minimizing a description length, and so on.

The *No Free Lunch* results throw such techniques into question: If there are no problem-independent reasons to prefer one algorithm over another, why is overfitting avoidance nearly universally advocated? (but frequent empirical "successes")

Occam's razor: in pattern recognition, one should not use classifiers that are more complicated than are necessary, where "necessary" is determined by the quality of fit to the training data.

The frequent empirical "successes" of Occam's razor imply that the classes of problems addressed have certain properties:

What might be the reason we explore problems that tend to favor simpler classifiers?

Principle of satisficing:

- Human cognition: through evolution, we have had strong selection pressure on our pattern recognition apparatuses to be computationally simple (require fewer neurons, less time, ...).
- Pattern recognition: Design methodology itself imposes a bias toward "simple" classifiers; we generally stop searching for a design when the classifier is "good enough".