Decision theory and supervised learning (continued)

M2 Stat SD 1

1 Framework and notations

- **Random dataset**: $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ where X_i and Y_i assume values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. The random pairs (X_i, Y_i) are independently and identically distributed according to an unknown probability measure p on $\mathcal{X} \times \mathcal{Y}$:
$$\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n \sim p^{\otimes n}.$$
- The **loss** function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ measures the cost of predicting z when the true value is y.
- The **expected risk** (or **generalization error**) of a **predictor** $f : \mathcal{X} \to \mathcal{Y}$ is defined as
$$R_p(f) = \mathbb{E}_p[\ell(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, f(x))dp(x, y).$$
- The **Bayes risk** is the minimal achievable expected risk:
$$R_p^\star = \inf_{f:\mathcal{X}\to\mathcal{Y}} R_p(f) = \mathbb{E}\left[\min_{z\in\mathcal{Y}} r_p(z, X)\right].$$
- A **machine learning algorithm** A maps a dataset $\mathcal{D}_n$ to a predictor $f : \mathcal{X} \to \mathcal{Y}$.

2 Two No-Free-Lunch Theorems

The following results, collectively known as the "no free lunch" theorems, elucidate the fundamental limitations inherent in learning algorithms when no structural assumptions are imposed upon the data-generating distribution.

The first theorem shall be established as an exercise in the subsequent section of this tutorial. The second theorem, being considerably more intricate in its proof, lies beyond the scope of our present treatment.

Theorem 1 (No Free Lunch, Fixed Sample Size)

Consider a binary classification problem with label space $\mathcal{Y} = \{0, 1\}$ and the 0-1 loss function $\ell(y, z) = \mathbf{1}\{y \neq z\}$. Let $\mathfrak{P}$ denote the collection of all probability distributions on $\mathcal{X} \times \mathcal{Y}$, and assume that the feature space $\mathcal{X}$ possesses infinite cardinality.
For any learning algorithm A and for any sample size $n > 0$,
$$\sup_{p\in\mathfrak{P}} \mathbb{E}\Big[R_p(A(\mathcal{D}_n))\Big] - R_p^\star \geq \frac{1}{2}.$$

- What does it mean that the excess risk is at least 1/2?
- This result is often used to argue that we must make assumptions about the data and its generating process in machine learning to select an algorithm and achieve meaningful learning. Can you comment on this interpretation?

Theorem 2 (No Free Lunch, Asymptotic Regime)

Consider a binary classification problem with label space $\mathcal{Y} = \{0, 1\}$ and the 0-1 loss function $\ell(y, z) = \mathbf{1}\{y \neq z\}$. Let $\mathfrak{P}$ denote the collection of all probability distributions on $\mathcal{X} \times \mathcal{Y}$, and assume that the feature space $\mathcal{X}$ possesses infinite cardinality.
For any monotonically decreasing sequence $\epsilon_n \downarrow 0$ satisfying $\epsilon_1 \leq 1/16$, and for any learning algorithm A, there exists a probability distribution $p \in \mathfrak{P}$ such that for all $n \geq 1$,

$$\mathbb{E}\Big[R_p(A(\mathcal{D}_n))\Big] - R_p^\star \geq \epsilon_n.$$

- Observe that Theorem 1 does not follow as a corollary of Theorem 2, nor conversely.
- In a certain sense, this result establishes an even more stringent impossibility than the first theorem. What is your interpretation of this enhanced impossibility?

3 Proof of Theorem 1

Since $\mathcal{X}$ possesses infinite cardinality, we may assume without loss of generality that $\mathbb{N} \subseteq \mathcal{X}$. (This amounts to a relabelling of elements in $\mathcal{X}$ by natural numbers.)

Fix an integer $k > n$ and a function $f : \mathcal{X} \to \mathcal{Y}$. Let $p_{k,f}$ denote the probability distribution in $\mathfrak{P}$ characterised by

$$\forall j \in \{1, \dots, k\}, \quad p_{k,f}\Big(X = j; Y = f(j)\Big) = \frac{1}{k}.$$

Consequently, when $(X, Y) \sim p_{k,f}$, the marginal distribution of X is uniform over $\{1, \dots, k\}$, and the relation $Y = f(X)$ holds almost surely.

The proof exemplifies the so-called **probabilistic method** characteristic of modern discrete mathematics and statistical theory. Rather than constructing a single adversarial distribution p for algorithm A, we shall introduce additional randomness by considering the distribution p as drawn over the parametric family $\{p_{k,f}\}_{f:\mathcal{X}\to\mathcal{Y}}$. We shall then demonstrate that the expected excess risk, averaged over this family, remains substantial. The proof concludes by taking the limit as $k \to \infty$.

1. Show that $R^\star_{p_{k,f}} = 0$.

Let $\hat{f}_n = A(\mathcal{D}_n)$ denote the classifier produced by algorithm A upon observing the training dataset $\mathcal{D}_n$. We define $S(f) = \mathbb{E}_{p_{k,f}}[R_{p_{k,f}}(\hat{f}_n)]$ as the expected risk of $\hat{f}_n$ under the data-generating distribution $p_{k,f}$.

2. Demonstrate that $S(f)$ depends upon f solely through its restriction to $\{1, \dots, k\}$ — that is, through the vector $\{f(1), \dots, f(k)\} \in \{0, 1\}^k$.

For any vector $b = (b_1, \dots, b_k) \in \{0, 1\}^k$, we define the function f_b by

$$f_b(x) = \begin{cases} b_x & \text{if } x \in \{1, \dots, k\}, \\ 0 & \text{otherwise.} \end{cases}$$

For notational convenience, we shall hereafter write $p_{k,b}$ in place of p_{k,f_b}.

We now fix a probability distribution q on $\{0,1\}^k$ and establish the following random construction:

$$\begin{aligned} B = (B_1, \dots, B_k) &\sim q \\ \{X_i\}_{i=1}^n &\sim \text{Uniform}(\{1, \dots, k\})^{\otimes n}, \quad \text{independent of } B \\ \forall i \in \{1, \dots, n\}, \quad Y_i &= f_B(X_i) \\ \hat{f}_n &= A(\mathcal{D}_n) \\ X &\sim \text{Uniform}(\{1, \dots, k\}), \quad \text{independent of } B \text{ and } \{X_i\}_{i=1}^n \\ Y &= f_B(X) = B_X. \end{aligned}$$

3. Prove that the conditional distributions satisfy $[\mathcal{D}_n \mid B = b] \sim p_{k,b}^{\otimes n}$ and $[(X, Y) \mid B = b] \sim p_{k,b}$ for any $b \in \{0,1\}^k$.
4. Establish that

$$\max_{f : \mathcal{X} \to \mathcal{Y}} S(f) \geq \sum_{b \in \{0,1\}^k} q(b) S(f_b) = \mathbb{E}_{b \sim q}[S(f_b)] = \mathbb{P}\left[\hat{f}_n(X) \neq f_B(X)\right].$$

We now specialise to the case where q is the uniform distribution on $\{0,1\}^k$.

5. Define the σ-algebra $\mathcal{F}_n = \sigma(X_1, \dots, X_n, B_{X_1}, \dots, B_{X_n})$. Prove that

$$\mathbb{P}\left[\hat{f}_n(X) \neq f_B(X) \,\middle|\, X \notin \{X_1, \dots, X_n\}, \mathcal{F}_n\right] = \frac{1}{2},$$

and consequently that

$$\mathbb{P}\left[\hat{f}_n(X) \neq f_B(X) \,\middle|\, \mathcal{F}_n\right] \geq \frac{1}{2} \mathbb{P}\left[X \notin \{X_1, \dots, X_n\} \,\middle|\, \mathcal{F}_n\right].$$

6. From this, derive the fundamental inequality

$$\mathbb{E}_q[S(f_B)] \geq \frac{1}{2}\left(1 - \frac{n}{k}\right).$$

7. Complete the proof of Theorem 1 by taking the appropriate limit.

4 Adaptive Learning and Model Selection

The no-free-lunch theorems illuminate the fundamental necessity of incorporating structural assumptions regarding the data-generating mechanism to achieve meaningful statistical learning. In practice, these assumptions typically manifest as regularities inherent in the underlying probability distribution — such as smoothness conditions, sparsity constraints, or low-dimensional geometric structure.

A central concept in circumventing the limitations imposed by these theorems is that of **adaptivity**. An algorithm is characterized as adaptive if it possesses the capacity to modulate its performance in response to latent properties of the data-generating distribution, thereby attaining near-optimal statistical efficiency across diverse scenarios without *a priori* knowledge of these distributional characteristics.

Consider, for illustration, the high-dimensional linear regression paradigm, wherein we may reasonably postulate sparsity — that is, the response variable depends upon merely a small subset of the covariates in X. Whilst this constitutes a substantive structural assumption, there remains the crucial challenge of adapting to the unknown sparsity level, e.g., the cardinality of the active covariate set.

Discussion question Provide exemplars of adaptive algorithms from your knowledge of machine learning or statistical methodology, elucidating the mechanism by which it achieves adaptivity.

Adaptivity is frequently realized through the systematic evaluation of empirical risk on a dedicated **validation** dataset, or alternatively via **cross-validation** procedures for hyperparameter calibration. The complete machine learning algorithm A that yields predictor $\hat{f}_n$ thus constitutes a sophisticated composition: a primary learning algorithm operating on the training data, coupled with a principled model selection or calibration procedure executed on the validation set, or through cross-validation applied to the training corpus itself.

The Fundamental Limitations of Adaptivity

It is crucial to recognize that even the most sophisticated adaptive algorithms cannot transcend the fundamental constraints established by the no-free-lunch theorems. Whilst adaptivity enables algorithms to achieve superior performance within restricted classes of distributions, it cannot eliminate the necessity for *some* underlying structural assumption. The theorems remain inviolate: without any assumptions whatsoever about the data-generating mechanism, no learning algorithm can achieve meaningful generalization.

What adaptive algorithms accomplish is the weakening of required assumptions — they permit us to operate under broader, less restrictive conditions whilst maintaining statistical efficiency. However, the foundational requirement for assumptions of some form persists. For instance, adaptive sparse regression methods may not require knowledge of the precise sparsity level, yet they fundamentally rely upon the assumption that sparsity exists.

This observation leads to a profound epistemological question regarding the nature of statistical machine learning: to what extent can our methodological choices be deemed objective, and where does subjectivity inevitably enter our inferential framework?

Discussion Questions

1. **The Objectivity Paradox**: Statistical machine learning is often presented as an objective, data-driven methodology. However, the no-free-lunch theorems demonstrate that assumptions are unavoidable. What constitutes “objective” versus “subjective” elements in our statistical practice? Consider the roles of, e.g, algorithm selection, model specification, hyperparameter tuning, performance metrics, data preprocessing choices.
2. **Assumption Hierarchy**: Different learning paradigms embody different philosophical stances about what constitutes reasonable assumptions. Compare and contrast the implicit assumptions underlying: Linear models, Lasso methods, k-nearest neighbors and Nadaraya-Watson regression, deep neural networks and other machine learning models.
3. **The Role of Domain Knowledge**: How should subject-matter expertise influence our choice of assumptions? When is the incorporation of domain knowledge scientifically justified versus potentially biasing? Discuss the tension between “letting the data speak” and leveraging prior understanding.
4. **Empirical Validation of Assumptions**: Can we objectively assess the validity of our assumptions through purely empirical means, or does such assessment inevitably require additional assumptions? What are the philosophical implications for the scientific method in data-driven disciplines?