1 Background: Theory

We would like to search a set of models, $\mathcal{M}$, in order to predict observations from an unknown data-generating process q, when we only have a finite data set $\mathcal{D}$ of independent observations from q. A principle hazard of model searches is losing predictive power by choosing a model that describes the data better than the process that generated it. Our aim is to understand and avoid overfit in model searches. Therefore, we will build handles on predictiveness, overfit and the information in $\mathcal{D}$ and suggest an overfit preventing model search algorithm.

To clarify the discussion, observe that each parametrization of each predictive model induces a distribution over the event set, and that two distinct parametrized models inducing the same distribution make the same predictions. Overfit is a funtion of the parameters only through the distribution on the event set induced by the parametrized model. Therefore, we will use the term "model" for the equivalence class of parametrized models inducing the same distribution. A search of parametric forms followed by a parametrization will therefore be seen as a search across classes of models followed by a model search in the chosen class.

1.1 The predictiveness of m on data from p

The ability of a model m to predict observations generated from a model p is its asymptotic ability to count the observations drawn independently from p:

Definition 1. *The* predictiveness of model m on data from model p *is the average asymptotic expected likelihood of* m *given data from* p*, denoted:*

$$\mathcal{L}(m;p) \equiv \lim_{n\to\infty} \underset{X_n \sim p}{E} \left[\mathcal{L}(m;X_n)^{\frac{1}{n}} \right]$$

Where X_n is a random sample of n independent observations drawn from the distribution induced by model p. The average is taken geometrically because the likelihood almost surely declines geometrically with linear increases in the sample size.

The following theorem and corollary are computationally and theoretically important:

Theorem 1. *If p is a discrete probability vector with k categories, then:*

$$\lim_{n\to\infty}\binom{n}{np_1\ldots np_k}^{1/n}=\lim_{n\to\infty}\left(\frac{n!}{\prod_{i=1}^{k}(np_i)!}\right)^{1/n}=\prod_{i=1}^{k}p_i^{-p_i}=\exp\left\{-\sum_{i=1}^{k}p_i\ln p_i\right\}\tag{1}$$

Corollary 1. *For discrete probability vectors m and p:*

$$\begin{aligned}\mathcal{L}(m;p)&=\prod_{i=1}^{k}\left(\frac{m_i}{p_i}\right)^{p_i}=\exp\left\{-\sum_{i=1}^{k}p_i\ln\frac{p_i}{m_i}\right\}\\&=\exp\left\{-D_{KL}(p;m)\right\}\end{aligned}\tag{2}$$

Where $D_{KL}(p;m)$ is the Kullback-Leibler divergence of p given m. Notice that the predictiveness of m on p is 1 if and only if $m=p$, in the sense that they induce the same distribution, and that the predictiveness is zero if and only if m assigns a zero probability to an event of nonzero probability under p.

1.2 Overfit

Let $e(\mathcal{D})$ denote the distribution that gives precisely the observed relative frequencies of events in $\mathcal{D}$ as the probability of these events, and let us call this the *empirical model.* Then we can define the amount of overfit in a model choice as the extent to which the chosen model predicts data from the empirical model better than it predicts data from the generating process, q:

Definition 2. *The* amount of overfit *in a model choice, m, is given by*

$$V(m, \mathcal{D}, q) = \mathcal{L}(m; e(\mathcal{D})) - \mathcal{L}(m; q)$$

The goal of avoiding overfit is to increase predictiveness by refraining from learning information from the data, $\mathcal{D}$, that doesn't describe the data generating process. The amount of overfit for each model m on $\mathcal{D}$ is a function of the data. Therefore, any model search that uses an *a priori* compensation to avoid overfit does not use information in $\mathcal{D}$ that could be used to improve its predictiveness by estimating the overfit as a function of $\mathcal{D}$. If we are to use the notion of "simplicity" to think about avoiding overfit, then we ought to use an *a posteriori* definition of simplicity. Notice that both choosing a "simple enough" parametric form before fitting it to the data and including a punishment for model complexity in a Bayesian prior before updating use *a priori* methods to avoid overfit.

1.3 Information in $\mathcal{D}$

The logic of statistical hypothesis testing can be used to avoid rejecting model choices that might predict the data-generating process even when the current observations are relatively unlikely according to the models: If a model m induces the same distribution over the event set as the data generating process q, then the p-value of m given data $\mathcal{D}$ from q is uniformly distributed on the unit interval. If the p-value of a model m on the data is less than some α, then either m does not induce the same distribution as q, or the p-value was drawn to be less than α from a uniform distribution on the unit interval. Using a smaller significance level α makes it less likely that you will reject model choice m due to unpredictive information in $\mathcal{D}$. So we propose the following definition:

Definition 3. *The* information in $\mathcal{D}$ at the α-level *is:*

$$\mathcal{I}_\alpha(\mathcal{D}, \mathcal{M}) \equiv \{m \in \mathcal{M} \mid \textit{P-value}(\mathcal{D}; m) < \alpha\}$$

That is, the α-level information in $\mathcal{D}$ is defined as the set of models in $\mathcal{M}$ that are rejected by a hypothesis test with significance level α. This definition places the information in $\mathcal{D}$ in the context of a model search; learning nothing means not being able to rule out any additional models from your search.

1.4 The α-level underfit score

Let us treat $\mathcal{A}_\alpha \equiv \mathcal{M} \setminus \mathcal{I}_\alpha(\mathcal{D}, \mathcal{M})$ as the set of candidate models; the set of models that might induce the same distribution over the event set as the data generating process. Then we can write the following definition:

Definition 4. *The* predictiveness profile of model m in candidate set $\mathcal{A}_\alpha$ *is:*

$$\mathcal{L}(m; \mathcal{A}_\alpha) \equiv \{\mathcal{L}(m; p) \mid p \in \mathcal{A}_\alpha\}$$

Since we don't necessarily have a measure on $\mathcal{A}_\alpha$, the most natural function of the predictiveness profile to use as a score for a model search is the "worst case predictiveness", which we will define as:

Definition 5. *The* α-level underfit score *of m given candidate set $\mathcal{A}_\alpha \subset \mathcal{M}$ is:*

$$UF_\alpha(m, \mathcal{M}) \equiv \inf_{p \in \mathcal{A}_\alpha} \mathcal{L}(m; p)$$

We call this the "α-level underfit score" because it assigns the worst predictiveness that is possible for each model as its score, if it is given or assumed that there is a model in $\mathcal{M}$ which induces the same distribution as the data generating process and that the α level we have chosen is "the right significance level". That is, it aims to give models the worst score that is might eventually achieve, given the data. In principle, models with better worst case scenarios are more robust to overfitting than models with more severe worst cases.

1.5 Aggression and Avoiding Overfit

In the context of a model search there is no reason, in principle, why we should prefer one significance level over another. Lower significance levels read less information from the data (they rule out fewer models) than higher ones. Each will imply different worst case predictivenesses for each model. Using a very low significance level will give a very conservative α-level underfit scoring of models, while using a high significance level will give a more aggressive choice. Therefore, we will define an "information preferences" that give weight or importance to each significance level:

Definition 6. *An* aggression profile *is a distribution on the levels of statistical significance, i.e. a probability measure on the unit interval.*

We can now integrate the α out of α-underfit scores using a given aggression profile as the integrating measure, to suggest a scoring:

Definition 7. *The* underfit score of m relative to the aggression profile μ *is:*

$$UF(m;\mu) = \int_0^1 UF_\alpha(m, \mathcal{M}) d\mu$$

This model search score compensates for overfit as a function of the data, or more specifically, it compensates for overfit as a function of the information in the data at every significance level. The performance of this model search method will be compared to many Bayesian choice rules for discrete multinomial data, in the next section.