Master's Thesis Proposal

Marcel Gietzmann-Sanders

Contents

1 Objectives

Objective 1. *To provide tooling that allows for using machine learning methods to fit models of the form*

$$\psi_k = G(\eta_k)$$

that maximize the following objective:

$$\mathcal{L} = \prod_i P'(v_i|\eta_i)$$

where:

$$P'(v_i|\eta_i) = \frac{\psi_i}{\sum_k \psi_k}$$

These models will be known as odds models as they predict the "odds for" each outcome v_k given the information contained in η_k.

1.1 Rationale for Objective 1

1.1.1 Value of Machine Learning in Model Building

One of the functions of science is to build models of the world around us. Specifically to look for functions of the form:

$$v = F(\theta)$$

where θ is some vector of information, v is a particular outcome given that information, and F is the model in question. This we might term a *deterministic model* because for any specific θ there is one single outcome v.

However, life is rarely so kind to us. From quantum physics to fisheries science the world is replete with examples where the same information does not always lead to the same conclusion. Sometimes that is because we do not possess all of the relevant information but sometimes things are just truly random. As such, the model above will in some sense be lying to us in its deterministic certainty - while we will always predict v, v is not always what we will receive.

Instead suppose that for a specific θ there exist a set of possible outcomes $V(\theta) = \{v\}$. We can define a *stochastic model* as:

$$P(v|\theta) = F(v, \theta)$$

where now the model is predicting the probability of v given θ instead of just predicting v itself.

Now what is especially interesting about models of this form is the fact that in some sense for any combination of V and θ they always exist. Whereas to get a deterministic model we need to know precisely what kind of information (θ) we need to make a deterministic prediction, we can always make a stochastic one, even in the presence of no information. The model can always give us a correct answer, it's just a question of how useful that answer is. If you base it off of better data it will become more useful.

Unfortunately, unless we ourselves have built the part of the world we're studying, F is not known to us. Instead our models are really just proposals $\hat{F}$ on what F could be. This presents us with an immediate problem, how do we evaluate any specific $\hat{F}$ if we don't know F itself? Well suppose we've captured a whole series of pairs of θ_i and v_i where i allows us to index the pair in question. Given we know $\hat{F}$ we can directly ask what the likelihood of the data we have collected is, given $\hat{F}$:

$$\mathcal{L} = \prod_i \hat{F}(v_i, \theta_i)$$

or equivalently (and more easily computed)

$$\ln \mathcal{L} = \sum_i \ln \hat{F}(v_i, \theta_i)$$

$\hat{F}$'s with higher $\ln \mathcal{L}$ (log likelihood) are better fits to the given data and as we have more and more comprehensive data those data better and better represent F (see Appendix 1).

Therefore finding the "true" F can be summarized with two steps:

1. Collect large amounts of comprehensive data.
2. Find the $\hat{F}$ that maximizes the likelihood of that data.

All of this, however, has been conditional on the form of θ having been chosen. Obviously, the information we choose to collect and build our model with has an impact on how good the model is from a purely predictive perspective. So how can we compare the predictive power of different θ? Well as we find θ that are more predictive the corresponding F's will have greater and greater likelihoods $\mathcal{L}$, given the data. But this presents us with a chicken and the egg style problem - to know the value of θ we must know F, but to know F we must have fit $\hat{F}$ which requires a reasonable amount of data to have been collected. I.e. I cannot know in advance if the data I am collecting is going to be useful or not. Instead anyone doing modeling finds themselves in the cycle illustrated by Fig. 1.

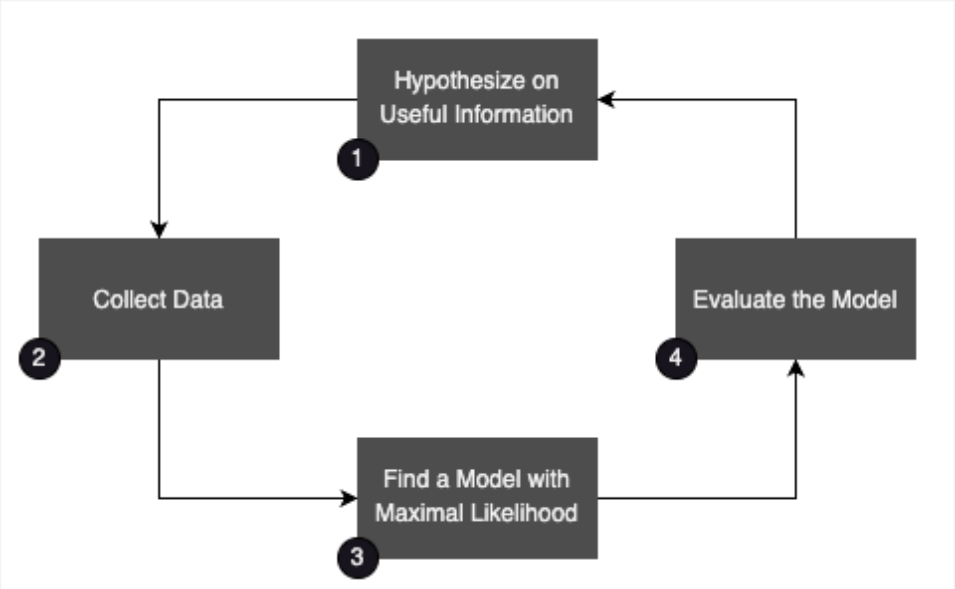

Figure 1: Modeling Cycle

Each step in this cycle is highly non-trivial. However special interest should be paid to step 3 (finding an $\hat{F}$ that best approximates F).

The standard method here has been to propose a family of possible F which are defined by a set of parameters μ. Then using some choice of likelihood maximization method the set of μ are searched for that maximize the likelihood of the data $\mathcal{L}$. While this method has borne considerable fruit it still requires a great deal of effort on the side of the scientist to envision various hypotheses on what families of functions would make sense, coding those up, fitting them, and then evaluating them post-fit. Furthermore if the true form of F is not contained in those set of hypotheses tested then F is never found. Lots of effort without any kind of guarantee that F will be discovered.

The field of Machine Learning (ML) (specifically probabilistic machine learning) on the other hand has been specifically occupied with finding ways to maximize objectives without having to assume much, if anything, about the functional form of the prediction function. Therefore it provides an opportunity to lessen the toil and uncertainty in step 3 and allow scientists to focus more of the efforts on the other steps (and thereby accelerate the whole cycle).

1.1.2 Issues with Probabilistic Machine Learning when Applied to Behavior

Probabilistic machine classification of the kind we were pointing to in the last section has two pitfalls when it comes to modeling behavior.

The first comes from the fact that such models predict on the same set of outcomes each time, regardless of θ. For example, if we were trying to predict which drink a person will buy at their local cafe we would not only have to include everything currently on the menu but also everything that could be on the menu across all time points we are interested in predicting. If suddenly a new menu item appears that we have not trained on, our probabilistic classifier will be at a loss.

The second issue comes from the "curse of dimensionality". Basically this is an issue where as the dimensionality of θ increases, the amount of data required to find real relationships in that data increases exponentially. This is a particular problem for behavioral modeling because we in all likelihood need to include data on each of the possible decisions before us. So if I can choose between 5 different drinks I need to provide features on each of those five drinks. This leads to very high dimensional spaces very quickly.

We can get around both of these problems by introducing what we will term an "odds model". Specifically if we imagine θ now as a variable length vector composed of vectors η_k for each of our possible outcomes v_k then the odds model is given by the function:

$$\psi_k = G(\eta_k)$$

where we now say that:

$$P'(v_k|\theta) = \frac{\psi_k}{\sum_j \psi_j}$$

We can see why we chose the name "odds model" as the ψ_k are simply the "odds for" outcome v_k.

Now note it is not necessarily true that:

$$P'(v_k|\eta_k) = P(v_k|\theta)$$

because it is possible that information in the other η_j informs the probability of v_k. However it should be possible to include such information in the η_k if necessary.

By making this sacrifice, however, we have achieved two things:

1. So long as we can provide a η_k for a newly seen outcome we can predict on it. Therefore we need not have a fixed set of possible outcomes (or a fixed number of such outcomes per θ).

2. Compared to the corresponding probabilistic machine learning classifier, if our total number of outcomes were $|V|$ we've now reduced the dimensionality of our feature space by $|V|$.

And this means that unless we have to pass large amounts of information between the η_k the amount of data we need to collect to fit our model has decreased tremendously. And in cases where data is hard to come by (like animal behavior), any reduction in data requirements is of the utmost importance.

So finally to model behavior we'd like to fit an odds model of the form:

$$\psi_k = G(\eta_k)$$

that maximizes:

$$\mathcal{L} = \prod_i P'(v_i|\eta_i)$$

where

$$P'(v_i|\eta_i) = \frac{\psi_i}{\sum_k \psi_k}$$

To the author's knowledge there is presently no tooling to fit and diagnose such models.

2 Appendices

2.1 Proof of Log Likelihood Maximization

Theorem 1. *Suppose we have a set of possible outcomes $V = \{v_k\}$ with probabilities P_k. Such that:*

$$\sum_k P_k = P$$

Next suppose we concoct a new series of probabilities $U_k = P_k\alpha_k$ s.t.

$$\sum_k U_k = P - \epsilon$$

where $\epsilon \geq 0$. There does not exist a set of α_k s.t.

$$\prod_k \left(\frac{P_k\alpha_k}{P_k}\right)^{P_k N} > 1$$

Proof. We proceed by induction.

First note that from the equation above we get:

$$\sum_k P_k \ln \alpha_k > 0$$

and subtracting $\sum_k U_k$ we arrive at:

$$\sum_k P_k \left(\ln \alpha_k - \alpha_k\right) > -P + \epsilon$$

Now suppose for k outcomes we know no such α_k exist as to satisfy the above. Suppose we incorporate a new $k+1$ term s.t:

$$P_{k+1} + \sum_k P_k = P_{k+1} + P$$

and:

$$P_{k+1}\alpha_{k+1} + \sum_k P_k\alpha_k = P_{k+1} + P - \epsilon = P_{k+1}\alpha_{k+1} + (P - P_{k+1}(\alpha_{k+1} - 1)) - \epsilon$$

Now we need:

$$P_{k+1}\left(\ln \alpha_{k+1} - \alpha_{k+1}\right) + \sum_k P_k \left(\ln \alpha_k - \alpha_k\right) > -P_{k+1} - P + \epsilon$$

However given our inductive assumption we know that at best,

$$\sum_k P_k \left(\ln \alpha_k - \alpha_k\right) = -P$$

therefore at best:

$$P_{k+1}\left(\ln \alpha_{k+1} - \alpha_{k+1}\right) - P > -P - P_{k+1} + \epsilon$$

Furthermore the easiest case for us is if $\epsilon = 0$. If it cannot be satisfied than this certainly doesn't work for $\epsilon > 0$. So let's consider:

$$P_{k+1}\left(\ln \alpha_{k+1} - \alpha_{k+1}\right) - P > -P - P_{k+1}$$

or equivalently:

$$\ln \alpha_{k+1} > \alpha_{k+1} - 1$$

Now if $\alpha_{k+1} = 1$ we have:

$$\ln 1 = 1 - 1$$

Further consider the derivatives of each side:

$$\partial_\alpha \ln \alpha_{k+1} = \frac{1}{\alpha_{k+1}}$$

$$\partial_\alpha(\alpha_{k+1} - 1) = 1$$

If $\alpha_{k+1} > 1$ then the log component rises more slowly than the constant component. I.e. our left side will be larger than the right. Likewise if $\alpha_{k+1} < 1$ the log component will shrink faster that the constant component which means it will also be less than the constant component. Therefore given our assumptions:

$$\ln \alpha_{k+1} \ngeq \alpha_{k+1} - 1$$

and so:

$$P_{k+1} (\ln \alpha_{k+1} - \alpha_{k+1}) - P \ngeq -P - P_{k+1} + \epsilon$$

So much for the $k+1$th case. What about $k = 1$. This is trivial because any α_1 that satisfies:

$$\sum_k U_k = P - \epsilon$$

must be less than or equal to 1 and therefore our product of quotients:

$$\prod_k \left(\frac{P_k \alpha_k}{P_k} \right)^{P_k N} > 1$$

must be less than or equal to one.

□

With this proof in hand we can now show how in the limit as the number of samples taken N goes to infinity our likelihood is only maximized if $\hat{F} \to F$.

For some given information θ and outcomes $V = \{v_k\}$ we have, in gathering our data, observed v_k N_k times. Therefore our overall likelihood is:

$$\mathcal{L} = \prod_k \hat{F}(v_k, \theta)^{N_k}$$

Now suppose that we represent:

$$\hat{F}(v_k, \theta) = F(v_k, \theta) \alpha_k$$

That is the ratio of our likelihoods between $\hat{F}$ and F is given by:

$$\prod_k \left(\frac{F(v_k, \theta) \alpha_k}{F(v_k, \theta)} \right)^{N_k}$$

but we also know that:

$$\sum_k F(v_k, \theta) = \sum_k F(v_k, \theta)\alpha_k = 1$$

and that $\lim_{N \to \inf} N_k = F(v_k, \theta)N$ so that we have:

$$\prod_k \left(\frac{F(v_k, \theta)\alpha_k}{F(v_k, \theta)} \right)^{F(v_k, \theta)N}$$

which is now in the exact same form as our theorem above. And we now know that this ratio is maximized when $\alpha_k \equiv 1$. So as $N \to \inf$ a higher $\mathcal{L}$ means we are closer to approximating F.