Predicting Adverse Events in Clinical Trials: A Nonstandard Problem in Statistical Learning?

M. Vidyasagar

Cecil & Ida Green Professor
The University of Texas at Dallas
M.Vidyasagar@utdallas.edu

University of Illinois at Urbana-Champaign
April 23, 2010

A Disclaimer

Unlike many of the speakers here, I have been working on these ideas for just about two or three months!

I have been discussing my ideas with some systems biology researchers involved in drug discovery, but basically *my thoughts are highly speculative and very preliminary*.

Based on further discussions and application to one or more specific problems, I hope to fine-tune the problem formulation.

All feedback is welcome!

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

General Motivation

An attempt to use statistical methods to predict "adverse events" in clinical trials.

A majority of drug candidates are rejected not for want of efficacy, but for toxicity (unwanted side effects).

Is it possible to predict that a particular drug candidate has a high likelihood of failure very early in the development cycle?

If so pharmaceutical companies could save billions of dollars by closing these programs very early.

"Fail early, fail cheaply" should be the motto.

What can *we* do to help?

Drug Discovery & Development Cycle (Simplified)

Basically two broad stages: Pre-clinical and clinical.

Pre-clinical stage: Experiments on target proteins and putative drug molecules, initially in microarrays, then in cells (*in vitro*) and finally in animals (*in vivo*).

Clinical stage: Experiments on humans in three phases:

• Phase I: 10 to 20 healthy volunteers are tested with drug candidate to establish no immediate harmful side effects
• Phase II: 100 to 300 afflicted patients are tested with drug candidate to establish efficacy
• Phase III: 1,000 to 3,000 patients are tested to establish long-term safety of usage

Late stage failures are *very* costly!

Two Kinds of Drug Candidates

Need to distinguish between two kinds of drug candidates: Small molecules and biologics (large protein molecules).

The former, being small, interact with many proteins in the body besides the target protein, leading to unwanted side effects; this is one form of toxicity.

These unwanted interactions are difficult to predict, so perhaps we have less to offer.

Biologics are very specific, but dosage is a critical factor. Too large a dosage can be toxic while while too small a dosage is ineffective.

How do we get the dosage 'just right' for a wide variety of people (the Goldilocks problem).

Why Do Drug Candidates Fail – 1?

Explanation No. 1: Cells just behave differently in a petri dish from the way they do in animals, and/or they behave differently in animals from the way they do in humans.

Why? Two main reasons: Absence of context, and absence of feedback (open-loop models).

As system theorists we can undertake to study and explain how interconnections of systems behave.

That is a topic for another talk.

Why Do Drug Candidates Fail – 2?

Explanation No. 2: Physiological parameters of people *vary across a very wide spectrum* – often an order of magnitude.

The "probability distribution" of physiological parameters is not known, and will probably *never be known* to any reasonable extent.

Adverse reactions in just 1% of patients can cause rejection!

Challenge: Predicting extreme events with very little data.

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

A Motivating Example (Out Dozens of Possibilities)

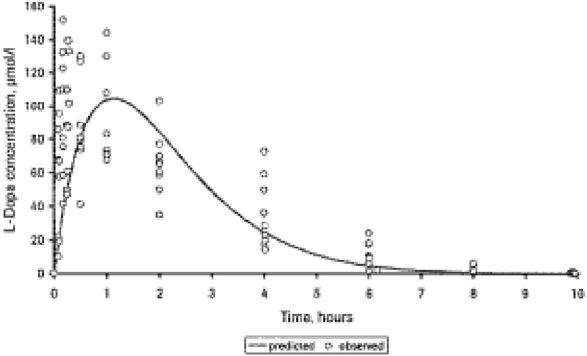
Reference: Susan Grange *et al.*, “A pharmacokinetic model to predict the PK interaction of L-Dopa and Benzerazide in rats,” *Pharmaceutical Research*, 18(8), 1174-1184, 2001.

21 male albino rats were administered L-D or B or both, and results observed. The pharmacokinetic interactions were modeled by a compartmental model consisting of 9 ODEs of the form

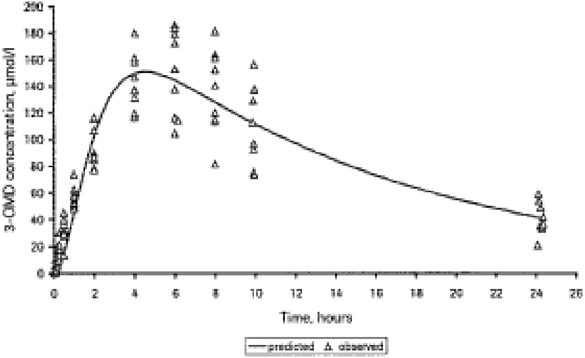
$$\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$$

where $\mathbf{x} \in \mathbb{R}^9$ is the ‘state’ of the system, $\mathbf{h}$ represents the dynamics, and $\boldsymbol{\alpha} \in \mathbb{R}^{30}$ represents the vector of physiological parameters.

Experimental Results and Fit to the Model Predictions – I



Experimental Results and Fit to the Model Predictions – II



Why are Predictions So Bad?

Model predicts *average behavior* well but does not even come close to predicting the *range of behavior* as physiological parameters vary.

Why? Because we have only 21 data points in $\mathbb{R}^{30}$!

So what is the remedy?

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

Abstract Problem Formulation

The physiological process under study is modeled by

$$\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, \boldsymbol{\alpha}),$$

where $\mathbf{x}$ is the state of the system, and $\boldsymbol{\alpha}$ is the vector of physiological parameters. Typically $\mathbf{h}$ contains terms that are linear, bilinear, or 'saturating' (Michaelis-Menten kinetics) in $\mathbf{x}$, and linear in $\boldsymbol{\alpha}$.

Problem: The vector $\boldsymbol{\alpha}$ is random and has an unknown probability distribution. What can we say about the probability distribution of the solution $\mathbf{x}(\cdot)$?

Some Simplifications – 1

The probability distribution of $\boldsymbol{\alpha}$ is *not entirely unknown*!

There are strong correlations between components of $\boldsymbol{\alpha}$. If $\boldsymbol{\alpha}$ has k components and we just discretize to H and L (high and low), then not all 2^k possible combinations are physiologically meaningful!

This can be captured by postulating a *family* of probability distributions $\mathcal{P}$, and saying that the true probability distribution P of $\boldsymbol{\alpha}$ belongs to $\mathcal{P}$ but is otherwise unknown.

Examples: Mixture models (on which more later).

Some Simplifications – 2

Often statements about steady state values are often good enough (we can ignore dynamics).

Let $f(\boldsymbol{\alpha})$ denote some function of the steady-state solution of the equation $\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$.

So f is the quantity of interest, e.g. the peak value over time of some response; if $f(\boldsymbol{\alpha})$ is higher than some threshold then drug gets rejected. We can have multiple quantities of interest also.

We can 'compute' f as a function of $\boldsymbol{\alpha}$ by solving system equations for many combinations of $\boldsymbol{\alpha}$.

Less Abstract Problem Formulation

There is a *known function* f of $\boldsymbol{\alpha}$, and a *known family of probability distributions* $\mathcal{P}$ to which the distribution of $\boldsymbol{\alpha}$ belongs. A threshold ϵ is specified.

Problem: Estimate $\Pr\{f(\boldsymbol{\alpha}) > \epsilon\}$.

Again, we can have multiple functions and multiple thresholds, and we can seek to estimate the probability of any Boolean function of the events $\{f_i(\boldsymbol{\alpha}) > \epsilon_i\}$, such as and, or, not, and so on.

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

A Conventional Learning Problem

Given a function $f : A \rightarrow \mathbb{R}$, and an unknown probability distribution P on A, estimate $E[f, P]$.

Standard solution: Generate i.i.d. samples $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_l$ from A with distribution P. Compute the 'empirical mean'

$$\hat{E}(f; \boldsymbol{\alpha}_1^l) := \frac{1}{l} \sum_{j=1}^{l} f(\boldsymbol{\alpha}_j).$$

Then $\hat{E}(f; \boldsymbol{\alpha}_1^l)$ is a decent approximation to $E[f, P]$.

Depending on stopping criterion, known as 'Monte Carlo' or 'Las Vegas' algorithm.

Sample Complexity Estimates

Hoeffding's inequality states that, if f is bounded between $[a, b]$, then

$$P^l \{ \boldsymbol{\alpha}_1^l \in A^l : | \hat{E}(f; \boldsymbol{\alpha}_1^l) - E[f, P] | > \epsilon \} \leq 2 \exp(-2l\epsilon^2/(b-a)^2).$$

A 'universal' bound, valid for *every* probability measure P. If

$$l \geq \frac{(b-a)^2}{2\epsilon^2} \ln \frac{2}{\delta},$$

then we can say that $| \hat{E}(f; \boldsymbol{\alpha}_1^l) - E[f, P] | \leq \epsilon$ with confidence $1 - \delta$.

Recent work by Abdallah, Dorato, Tempo, Alamo et al. makes $l \sim O(1/\epsilon)$, not $O(1/\epsilon^2)$.

Vapnik-Chervonenkis Theory

Hoeffding's inequality can be used to estimate the means of *finitely many* functions simultaneously.

What happens if we want to estimate, simultaneously, *infinitely many means?*

One computes the so-called VC-dimension, or its generalization the so-called Pollard dimension. If it is finite, then again the maximum error between the empirical means and true means goes to zero as $l \to \infty$, where l is the number of samples.

Extensions to the case where successive samples are correlated, etc.

These are conventional problems in statistical learning theory.

Key Assumptions Underlying the Theory

1. We have access to samples $\boldsymbol{\alpha}_j$ generated according to the 'true but unknown' probability measure P.
2. For each sample, we can compute $f(\boldsymbol{\alpha}_j)$.

What if these assumptions do not hold?

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

Non-Standard Learning Problem

Given known functions $f, g_1, \ldots, g_k$ of a random parameter vector $\boldsymbol{\alpha}$ with unknown probability distribution $P \in \mathcal{P}$.

Given the values

$$g_i(\boldsymbol{\alpha}_j), i = 1, \ldots, k, j = 1, \ldots, l$$

corresponding to randomly generated samples $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l$ distributed according to P.

Compute upper and lower bounds for $E[f, P]$.

Distinguishing Features of Non-Standard Problem

1. We are not allowed to see the samples $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l$ directly. (Otherwise we could ignore the functions g_i and just use Monte Carlo simulation.)
2. $k \ll l$, so we cannot 'invert' the functions g_i to deduce samples.
3. We are happy to get just upper and lower bounds for $E[f, P]$.

In clinical analysis, the function g_i can be thought of as 'bio-markers' – they give an indication of the unknown and unmeasurable physiological parameters $\boldsymbol{\alpha}$.

Usefulness of Bounds on Expected Value

How can we use bounds on f_u, f_l on $E[f, P]$ to estimate tail probabilities?

Markov's inequality: Suppose $f \geq 0$. For every $\epsilon > 0$, we have

$$P\{\boldsymbol{\alpha} \in A : f(\boldsymbol{\alpha}) > \epsilon\} \leq \frac{E[f, P]}{\epsilon} \leq \frac{f_u}{\epsilon}.$$

Refined Markov's inequality: For every $\epsilon > 0$ and every λ, we have

$$P\{\boldsymbol{\alpha} \in A : f(\boldsymbol{\alpha}) > \epsilon\} \leq \exp(-\lambda\epsilon) E[\exp(\lambda f), P].$$

Proof: Note that $\{f(\boldsymbol{\alpha}) > \epsilon\} \Leftrightarrow \{e^{\lambda f(\boldsymbol{\alpha})} > e^{\lambda\epsilon}\}$, and apply standard Markov inequality.

A Further Refinement

If there are too few samples, we can modify the problem:

Given known functions $f, g_1, \ldots, g_k$ of a random parameter vector $\boldsymbol{\alpha}$ with unknown probability distribution $P \in \mathcal{P}$, and given that $E[g_i, P] = c_i, i = 1, \ldots, k$, compute upper and lower bounds for $E[f, P]$.

Given the random measurements of $g_i(\boldsymbol{\alpha}_j)$, the estimated means $\hat{E}[g_i, P], i = 1, \ldots, k$ are more reliable than individual samples.

This is especially true when some measurements are 'missing', i.e. $g_i(\boldsymbol{\alpha}_j)$ is not available for some pairs (i, j).

Outline

1 Motivation
- General Motivation
- A Motivating Example

2 Abstract Problem Formulation

3 Relationship to Conventional Learning Problem

4 Non-Standard Learning Problem
- Problem Formulation and Significance
- A Classical Statistical Mechanics Approach
- A Linear Programming Formulation

5 Next Steps

Statistical Mechanics Approach (Jaynes 1957)

Specific reference: E. T. Jaynes, “Information theory and statistical mechanics,” *Physical Review*, 106(4), 620-630, May 15, 1957.

Suppose $\boldsymbol{\alpha} \in A$, a finite set. Find the probability distribution P on A that *has maximum entropy* while satisfying the k equality constraints

$$E[g_i, P] = c_i, i = 1, \ldots, k.$$

Recall that if P has the distribution $\mathbf{p} = [p_i]$, then the **entropy** of P is given by

$$H(\mathbf{p}) = \sum_{i=1}^{m} p_i \log(1/p_i),$$

where m is the size of the set A where $\boldsymbol{\alpha}$ lives.

Statistical Mechanics Solution

It turns out that P is unique because the above is a convex optimization problem. It leads naturally to the so-called ‘partition function’ of statistical mechanics.

Perfectly fine for ‘equilibrium’ situations, i.e. when we can assume that the world tends towards maximum entropy while respecting physical measurements.

Not so fine for *our* situation – *Why* should unknown probability distribution P have maximum entropy?

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

A Linear Programming Formulation

Problem: Given functions $f, g_1, \ldots, g_k : A \to \mathbb{R}$, and constants $c_1, \ldots, c_k$, find

$$\min E[f, P] \text{ s.t. } E[g_i, P] = c_i, i = 1, \ldots, k,$$

$$\max E[f, P] \text{ s.t. } E[g_i, P] = c_i, i = 1, \ldots, k.$$

If universe A where $\boldsymbol{\alpha}$ lives is a finite set, then both are *linear programming* problems!

Some Observations

Note: If f is a linear combination of the functions $g_1, \ldots, g_k$, then the constraints *automatically specify* $E[f, P]$!

In general, *project* f onto subspace spanned by the functions $g_1, \ldots, g_k$. Write

$$f(\boldsymbol{\alpha}) = f_r(\boldsymbol{\alpha}) + \sum_{i=1}^{k} b_i g_i(\boldsymbol{\alpha}),$$

where f_r is the 'residual' or unpredictable part.

Choose a 'nominal' probability measure $P_0 \in \mathcal{P}$, and choose the constants b_i to minimize the ℓ_2-norm of f_r, i.e. $E[f_r^2, P_0]$. This guarantees that f_r is orthogonal to each g_i, i.e.

$$E[f_r g_i, P_0] = 0, i = 1, \ldots, k.$$

Reformulation of Problem

As stated problem is infeasible!

Suppose $\boldsymbol{\alpha}$ has 30 components (as in Susan Grange's paper), and we discretize each component to just two values (high and low). Then $|A| = \{H, L\}^{30}$ has $2^{30} \approx 10^9$ elements!

And what do we do if A is an infinite set (continuously varying parameters $\boldsymbol{\alpha}$)?

Source of difficulty: Failure to use *prior information* about the possible probability distribution P. We have already seen that *arbitrary* probability distributions of physiological parameters make no sense!

Formulation of Mixture Models

Remedy: Assume a 'mixture model' – works even if A is infinite.

Assume that the unknown probability distribution $P \in \mathcal{P}$, where

$$\mathcal{P} = \left\{ P = \sum_{i=1}^{s} \lambda_i P_i \right\},$$

where $P_1, \ldots, P_s$ are **known probability distributions** that reflect physiological realism.

Prior information (or *a priori* belief) is incorporated into the choice of the 'extremal' distributions $P_1, \ldots, P_s$.

Use of Mixture Models

In the standard PK/PD world, one uses (say) a mixture of three Gaussian measures:

$$P = \lambda_1 P(N(\mu_1, \sigma_1)) + \lambda_2 P(N(\mu - 2, \sigma_2)) + \lambda_3 P(N(\mu_3, \sigma_3)),$$

and uses the observed values of the physiological parameters $\alpha_1, \ldots, \alpha_l$ to estimate the means μ_i, variances σ_i, and weights λ_i.

Highly nonlinear problem, and answers are not very reliable.

Our approach: Instead of a mixture of three Gaussians, take a mixture of *fifty or a hundred* Gaussians, with unknown weightages λ_i!

Converts a nonlinear estimation problem into a linear programming problem!

Linear Problem Formulation With Mixture Models

In our setting, the problem becomes

$$\max E[f, P] \text{ s.t. } E[g_i, P] = c_i, i = 1, \ldots, k, \textit{ and } P = \sum_{i=1}^{s} \lambda_i P_i.$$

Features:

- This is also an LP, but in $\lambda_1, \ldots, \lambda_s$, the weights used in the mixture model.
- Size of problem is now s, the number of 'corner' probability distributions.
- Therefore we can have a very large number of mixture elements (several hundred if we wish) and the problem is still tractable.
- *We don't try to estimate the mixture model itself.*

Outline

1. Motivation
   - General Motivation
   - A Motivating Example
2. Abstract Problem Formulation
3. Relationship to Conventional Learning Problem
4. Non-Standard Learning Problem
   - Problem Formulation and Significance
   - A Classical Statistical Mechanics Approach
   - A Linear Programming Formulation
5. Next Steps

Choosing the ‘Right’ Disease to Apply Theory

Need to find a disease in which

- The mechanisms of disease onset and drug action (in terms of the cascade of pathways) are fairly well-understood. This leads to a ‘known functions of unknown parameters’.
- A few moderately reliable biomarkers are available.
- A medical researcher is interested.

Current Status

Have identified some practical problems in clinical studies that fit this framework.

Working with various clinicians to obtain ‘real data’.

Getting ‘real data’ is as hard as pulling ‘real teeth’!

Some hope on the horizon – will report when some success is realized.

Thank You!