Paranoid, huh?

November 25, 2013

**PRISM: and real paranoia**

- NSA is watching us
- But how many of us?
- I talked to an aid to the only SNL write / senator: Can you name him?
  - Hint: Minn
  - Democrat (obviously, since republicans aren’t funny)
  - Al Franken (Alvaro Bedoya is who I talked to)
- NSA says sampling is impossible
- Statisticians beg to differ
• So how’s that for real paranoia?

• We’ll look at a less evil form: namely nature

**Models: A review**

Recall our usual model;

\[(\forall i)Y_i = \alpha + \sum_j \beta_j X_{ij} + \epsilon_i\]

where

\[\epsilon_i \sim_{iid} N(0, \sigma^2)\]

• Strong stuff, lots of data aren’t plausibly normal

• Wonderful conclusions, MLE, regression, efficiency, RIC, etc, all revolve around this model

• Which of these statements are still true if the model is wrong?

• This is the question of model free statistics

• Championed by modern machine learning

**An always true model**

The equation:

\[(\forall i)Y_i = \alpha + \sum_j \beta_j X_{ij} + \epsilon_i\]
always holds for some $\epsilon_i$’s. So if we drop our conditions on $\epsilon$’s we have an always true model.

- What theorems are still true? (None as previously stated)
- Can we replace the old theorems with new ones? (amazingly, this is pretty much always yes)
- IID normal can mostly be replaced by bounded

**Typical statements**

- MR. Afterdefact looks at all the data after it was collected and says, “You know, a regression of $Y$ on $X$ fits pretty darn well.”
- Ms. Statistician, looks at the data sequentially and predicts the next observation.
- Competitive comparison: Can Ms. Statistician have almost as good as fit as Mr. Afterdefact?

**Model based on money**

One of your close friends goes to Stanford business school and the other goes to Chicago. Each claims they are going to kick the markets butt over the next 50 years. They have very different approaches. Your goal is
Choosing between two investments

- **Problem:** Suppose I have two friends who are hot-shot financial wizards. They come from different schools of thought and both believe the other to be totally clueless. So, in fact, I have one friend who is a financial wizard, and one friend who is an impostor. But, I don’t know which is which!

- **Goal:** I want to get as rich as my financial wizard friend—whichever that empirically turns out to be.

- **No assumptions:** I will not make any probabilistic assumptions.

Setup for finance

- **Notation:**
  
  - $A_t$ is the wealth of my first friend at time $t$
  
  - $B_t$ is the wealth of my second friend
  
  - $C_t$ is my wealth
\( w_t \) is fraction of my wealth \( A \) invests for me at time \( t \)

- \( A_0 = B_0 = C_0 = 1 \)

- **Returns:**
  - \( R_t^A = A_t / A_{t-1} \) is \( A \)'s return
  - \( R_t^B = B_t / B_{t-1} \) is \( B \)'s return
  - \( R_t^C = w_{t-1} R_t^A + (1 - w_{t-1}) R_t^B \) is my return

- **Goal:**
  \[
  C_t \equiv \max(A_t, B_t)
  \]
  All three are growing “exponentially,” so use \( \log(C_t) \) instead.
  Now growing “linearilly.” So use
  \[
  \frac{\log(C_t)}{t} \approx \max\left( \frac{\log(A_t)}{t}, \frac{\log(B_t)}{t} \right)
  \]
  as our goal.

**Invest with my BFF**

- **Scheme:** Whichever friend is currently wealthier is “more likely” to be the financial wizard. So have her invest all my wealth:
  \[
  w_t = \begin{cases} 
  1 & \text{if } A_{t-1} \geq B_{t-1} \\
  0 & \text{if } A_{t-1} < B_{t-1} 
  \end{cases}
  \]
• **Evil data:** Will this scheme always work? No, the following example shows this:

<table>
<thead>
<tr>
<th>time</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>⋯</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_t$</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>⋯</td>
</tr>
<tr>
<td>$B_t$</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>⋯</td>
</tr>
<tr>
<td>$w_t$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>⋯</td>
</tr>
<tr>
<td>$C_t$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>⋯</td>
</tr>
</tbody>
</table>

• **growth rates:**
  
  - A’s growth rate: $\ln(2)/2$
  - B’s growth rate: $\ln(2)/2$
  - C’s growth rate: 0

**Equal weight:** “some are more equal than others.”

• **Scheme:** Always have each friend invest $1/2$ of my wealth:

$$w_t = 1/2$$

• **Evil data:** This scheme isn’t sensitive enough. If one friend is doing better, it won’t notice. This motivates the following bad data:
growth rates:

- A’s growth rate: \( \frac{\ln(2)}{2} \)
- B’s growth rate: 0
- C’s growth rate: \( \frac{\ln(1.5)}{2} \)

Value weighted

- **Scheme:** Have each invest in proportion to how well the have done so far.

\[
w_t = \frac{A_{t-1}}{A_{t-1} + B_{t-1}}
\]

- No evil data exist:
- Growth rate of C:

\[
\frac{\ln(C_t)}{t} = \frac{\ln(A_t/2 + B_t/2)}{t} \geq \max\left\{\frac{\ln(A_t)}{t}, \frac{\ln(B_t)}{t}\right\} - \frac{\ln(2)}{t}
\]

In particular:

\[
\lim_{t \to \infty} \frac{\ln(C_t)}{t} - \max\left\{\frac{\ln(A_t)}{t}, \frac{\ln(B_t)}{t}\right\} = 0
\]

**Theorem**

- Without any assumptions on the returns
- We can statistically combine two investments
- such that our growth rate is as good as the better of these two growth rates
What is the trick?

- It is all in the “log”
- We will have 1/2 the wealth of our richer friend
- $\log(2 \text{ Billion}) \approx \log(1 \text{ billion})$

But this trick works in statistics too

- It works for regression, and other estimation problems
- It works for creating a good definition of probability (we’ll look at this next time)
- It turns out to be equivalent to our usual approach—so we inherently already are using the trick