Experimental optimization
Move the metrics that matter

David Sweet
Context
ML/AI in industry

- ML/AI models usually predictors, supervised learning
- Example predictions:
  - Probability a user will click on an ad
  - Probability a credit card charge is fraudulent
  - Expected return of a stock
  - Probability a user will “like” a post

Can you think of others?
**Prediction vs. control**

**RL in disguise**

- Predictor: Estimates target value
- Controller: acts on environment, receives reward
- In ML: Predictor: Supervised learning :: Controller: Reinforcement learning
- Predictor is usually embedded in a controller, ex:
  - Ad server
  - Credit card fraud detector
  - Stock trading strategy
  - Social media feed
## Predictors in controllers
Act on predictions to receive reward

<table>
<thead>
<tr>
<th>Controller</th>
<th>Prediction</th>
<th>Action</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad server</td>
<td>P{click}</td>
<td>Show ad with highest P{click}</td>
<td>CPC revenue</td>
</tr>
<tr>
<td>Fraud detector</td>
<td>P{fraudulent}</td>
<td>Hold charges with high P{fraudulent} until customer gives OK</td>
<td>Avoid losing money to fraud</td>
</tr>
<tr>
<td>Trading strategy</td>
<td>E[return]</td>
<td>Buy when E[return] &gt; 0, sell when E[return] &lt; 0</td>
<td>Revenue</td>
</tr>
<tr>
<td>Social media feed</td>
<td>P{like}</td>
<td>Show posts with highest P{like}</td>
<td>Users spend time on feed &amp; come back</td>
</tr>
</tbody>
</table>
Business metrics
The metrics that matter

• Business metrics == rewards

• Ex: dollars earned, dollars saved, MAU, time spent, risk taken

• Communicate in business metrics

• Compare these two self-assessments:
  • “I reduced RMSE by 23 basis points”
  • “I increased revenue by $90,000,000.”

• Translate prediction quality to business metrics with experiments
Questions?
Experiments
A/B tests in particular

• Translate “change in prediction quality” into “change in business metric”

• Example:
  • You design a new feature and add it to your model
  • Call the old model “A” and the new model “B”
  • Run A in the controller and measure business metric, $BM(A)$
  • Run B in the controller and measure business metric, $BM(B)$
Experiments
A/B tests in particular

• Goal is to answer:

\[ BM(B) > BM(A) ? \]

• If so, then

• Switch the controller to B

• Tell everyone you improved BM by \( BM(B) - BM(A) \)

“Revenue up by $90MM”
Problem: noise
Aka, variation, uncertainty, error

• BM will vary from measurement to measurement:
  • A user might not click on an ad now, but would have last week because, in the interim, they purchased the product.
  • A certain criminal might commit fraud next week, but won’t today while you’re taking your measurement
  • Stocks go up or down because of global news, industry news, stock-specific news, actions of specific traders, etc.
  • Maybe a user spends more time on social media on a Monday night than on a Friday night
I run a linear regression that minimizes SSE. You notice outliers in the data. You run a linear regression that minimizes least-absolute value (LAV).

How could tell which is a better model?
Problem: noise
Aka, variation, uncertainty, error

• How can you be certain $BM(B) > BM(A)$ will hold tomorrow or for a different user or at another time of day, etc.?

• You can’t.

• But you can limit your uncertainty.
Solution: Replication
Reduce noise by repeating measurements

- *Replication*: Take many measurements and average them
  
  $$
  \mu_A = \frac{\sum_i^N BM_i(A)}{N} \quad \text{and} \quad \mu_B = \frac{\sum_i^N BM_i(B)}{N}
  $$

- Repeat measurement for many users, many days, etc.

- Then ask:
  
  Is $\mu_B > \mu_A$?

- Put another way, set $\Delta = \mu_B - \mu_A$ and ask
  
  Is $\Delta > 0$?
Solution: Replication
Replicate to increase precision (reduce uncertainty)

- The uncertainty in $\mu_A$, called *standard error*, is

$$SE_A = \frac{\sigma_A}{\sqrt{N}}$$

- Reduce uncertainty (SE) by increasing N.

- How large should N be? It depends on how certain you want to be!
Probably not wrong
Limit the false positive rate

• Say you measure $\Delta = \mu_B - \mu_A > 0$.

• Maybe you just got lucky, and tomorrow you would measure $\Delta \leq 0$

• Called a “false positive” or Type I error

• Convention: Try for $P\{\text{just got lucky}\} < .05$

• More precisely: $P\{ \text{true } BM(B) \leq BM(A) | \mu_B > \mu_A \} < .05$

• Put another way: $P\{\text{“out-of-sample” will fail }|\text{ “in-sample” worked}\} < .05$
Probably not wrong
Limit the false positive rate

- Measured $\mu_A, \mu_B, SE_A, SE_B$
- How can you construct probabilities from these?
  - First: $\Delta = \mu_B - \mu_A$ and $SE_\Delta = \sqrt{SE_A^2 + SE_B^2}$
  - Then, central limit theorem says $\Delta \sim \mathcal{N}(\Delta_0, SE_\Delta^2)$

True for large $N$. Otherwise Student t distribution
Probably not wrong
Limit the false positive rate

- CLT: $\Delta \sim \mathcal{N}(\Delta_0, SE^2_\Delta)$

- Your entire experiment produces one draw from this distribution.
  - One experiment $\implies$ one $\Delta$
  - The smaller $SE_\Delta$ is, the closer $\Delta$ is to the true value, $\Delta_0$
Probably not wrong
Limit the false positive rate

• Proceed like this:
  • Hypothesize that $\Delta_0 = 0$, i.e., $BM(A) = BM(B)$

    • Then $\Delta \sim \mathcal{N}(0, SE^2_\Delta)$, i.e. $\frac{\Delta}{SE_\Delta} \sim \mathcal{N}(0,1)$

    • Define $z = \frac{\Delta}{SE_\Delta}$ and note that $P\{z > 1.64\} = .05$

• If you measure $z > 1.64$, then the probability you just got lucky is less than .05.
Design the experiment
Begin with the end in mind

• Goal: Determine the (minimum) number of replications needed to make $SE_{\Delta}$ small enough to get $z > .05$

• Choose $N$ such that:
  
  • If true value $\Delta_0 > 0$, then measured value will be
    
    $$z = \frac{\Delta}{SE_{\Delta}} = \frac{\Delta}{\sqrt{SE_A^2 + SE_B^2}} > 1.64$$

  • Define $\sigma^2_{\Delta} = \sigma^2_A + \sigma^2_B$, then write: $z = \sqrt{N} \frac{\Delta}{\sigma_{\Delta}} > 1.64$
Design the experiment
Find the minimum number of replications

- \[ z = \sqrt{N \frac{\Delta}{\sigma_{\Delta}}} > 1.64 \]
- Solve for \( N \): \( N > \left[ \frac{1.64 \sigma_{\Delta}}{\Delta} \right]^2 \)
- You must run at least \( N_{\text{min}} = \left[ \frac{1.64 \sigma_{\Delta}}{\Delta} \right]^2 \)
- But you don’t know \( \sigma_{\Delta} \) or \( \Delta \) before running the experiment!
Design the experiment
Find the minimum number of replications

• Replace $\Delta$ with the smallest difference between $BM(B)$ and $BM(A)$ that you care about, $\delta$

• Ex., Would an extra $1$/day matter? How about $10,000$/day?

• $\delta$ is the precision required of the measurement.

• Approximate $\sigma_{\Delta}$ by saying $\sigma_B \approx \sigma_A$: $\sigma_{\Delta} = \sqrt{2} \sigma_A$

• Estimate $\sigma_A$ from existing data: $\hat{\sigma}_A$
Design the experiment
Find the minimum number of replications

• Finally:

\[ N_{\text{min}} = \left[ 1.64 \frac{\sqrt{2} \hat{\sigma}_A}{\delta} \right]^2 \]
Design the experiment

One more thing: False negatives

- You’d also like to limit the probability that you’ll measure $\mu_B < \mu_A$ when, in fact, BM(B) > BM(A).
- That case is *unlucky*.
- That’s a *false negative*, or Type II error.
- We usually limit that to $P\{\text{false negative}\} > .20$
- Save that discussion for some other time.
Design the experiment

An example

- Ex: You build a new AI model for predicting whether a user will click on an ad. Your new model (B) has a lower cross entropy than the old model (A).

- If your model improved the ad revenue by anything less than $.001/pageview, likely no one would care. They wouldn’t even bother to deploy your model in production. Therefore, $\delta = $.0001$.

- From logged production data you measure the standard deviation of ad revenue/pageview of the old model as $\hat{\sigma}_A = $.10$

- Calculate: $N_{min} = \left[1.64 \frac{\sqrt{2\hat{\sigma}_A}}{\delta}\right]^2 = \left[1.64 \frac{\sqrt{2\$.10}}{$.001}\right]^2 \approx 54,000$ pageviews
Run the experiment
Measure $BM(A)$ and $BM(B)$

- Randomize: Each time you serve a page, “flip a coin”
  - Heads $\Rightarrow$ use the old model, $A$
  - Tails $\Rightarrow$ use the new model, $B$
- Record the revenue produced by that page
  - If the user clicks on the ad, revenue = $\$\text{CPC}$ for that ad
  - If not, revenue = $\$0$
- Until you have $N$ measurements of $BM(A)$ and $N$ of $BM(B)$
Run the experiment
Randomize to improve accuracy (lower bias)

- Consider non-randomizing approaches:
  - Use model A for US users and model B for non-US users, or
  - Use model A in the morning and model B in the evening, or
  - Use model A on Sunday and model B on Monday, etc.
- You’re not just measuring the BM difference between model A and B.
- You measuring the difference between US and non-US users, or morning and evening usage patterns, or Sunday and Monday usage patterns
- These other factors are called *confounders*. 
Analyze the experiment

\( z, \text{ again} \)

- Experiment is complete & you have your measurements

\[ z = \frac{\Delta}{SE_\Delta} \leq 1.64 \text{ from measurements now, not estimates} \]

- Is \( z > 1.64? \)
  - Yes ==\> Switch to model B
  - No ==\> Stay with model A
A/B tests are awesome
Because they’re simple

• Simple to design, run, and analyze
• Results are easy to communicate to experts and non-experts alike
• Applicable to arbitrary changes:
  • Changes to model features, architecture, loss function, …
  • Changes to controller
  • Changes to infrastructure
  • Changes to visual design
  • …
Optimization perspective
Monotonic improvement

• Accept B (new idea) ==> improvement
• Reject B ==> no improvement
Summary
Experimental optimization

• Measure and communicate business metrics (not loss functions)
• Run experiments to measure changes in business metrics
• Design to limit false positives and false negatives
• Replicate for precision and randomize for accuracy
• Switch from A (old) to B (new) if $z > 1.64$
• Keep experimenting to keep improving
Questions?