Experimental optimization

Lecture 1: Definition and motivation

David Sweet

Introductions

Welcome

- Why are you getting a DAV masters?
- What has been your favorite topic (not class) that you've studied so far in DAV?
- What do you plan to do in your career after you receive your masters degree?

Review

Supervised learning

- Data: (y, X)
  - y : target, regressand
  - X: features, predictors, regressors
- Prediction: $\hat{y} = f(X)$
- Minimize loss / error function, $E(y, \hat{y})$, over fit set
- Check $E(y, \hat{y})$ on test set

Question

You fit two models with different sets of features to the same fitting data.

How do you decide which model is better?

Question

You fit two linear regression models using the same features and fitting data.

One fit minimizes squared error, SSE.

The other fit minimizes least absolute value (LAV).

How do you decide which model is better?

Industrial engineered systems

Prediction vs. control

- Predictor: Estimates target value
- Controller: acts on environment, receives reward
- 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

Industrial engineered systems

Predictors in controllers

Controller	Prediction	Action	Reward
Ad server	P{click}	Show ad with highest P{click}	CPC revenue
Fraud detector	P{fraudulent}	Hold charges with high P{fraudulent} until customer gives OK	Avoid losing money to fraud
Trading strategy	E[return]	Buy when E[return] > 0, sell when E[return] < 0	Revenue (“PnL”)
Social media feed	P{like}	Show posts with highest P{like}	Users spend more time on feed

Business metrics
The metrics that matter

- Business metrics == rewards
- Ex: dollars earned, dollars saved, MAU, time spent, risk taken
- Communicate in business metrics, not losses
- Compare these two self-assessments:
  - “I reduced RMSE by 23 basis points”
  - “I increased revenue by $20,000,000.”

Question

You fit two predictors of advertisement click rate with different sets of features.

How do you determine which generates more revenue?

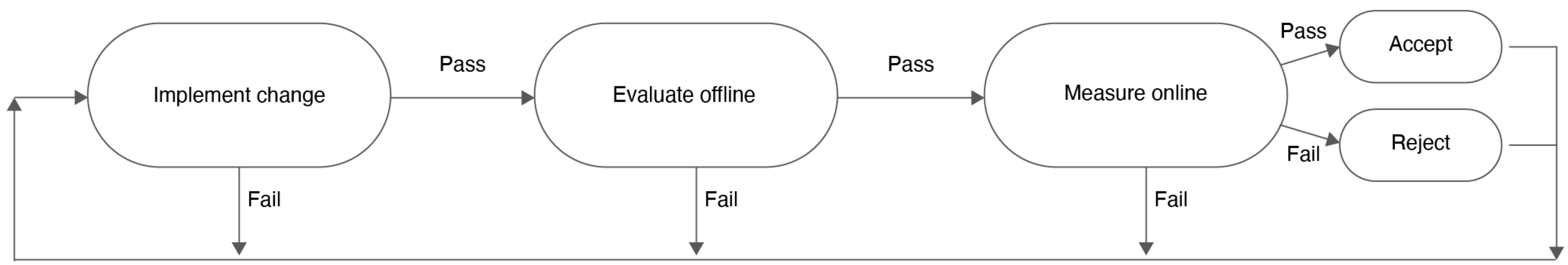
Experiment

Measure and compare

- Measure: Run your new model in production and measure business metric directly
- Compare: If better, switch to new model, else don’t
- Measure and compare changes to model, code, hardware, configuration, etc.

Workflow / pipeline

Monotonic improvement



- Only system-improving changes make it all the way through

Experimentation costs

- Time: days - weeks typical
- Metric reduction: might lose money, reduce clicks, drive users away, etc.
- Damage: new code might have bugs, release process might fail
- Engineer needs to measure and compare
- Any alternatives to experimentation?

Experimentation alternatives

Domain knowledge

- Can't a knowledgable, experienced person tell what's going to work?
- Consider: Engineers only experiment on system changes that they think will be accepted, yet changes are usually rejected by experiments.
- Amazon: 50% rejected
- Microsoft: 2/3 rejected
- Netflix: 90% rejected [1]
- Why? Complexity.

Experimentation alternatives
Prediction quality

- If predictor works better out-of-sample, won't it work better in production?
- Better how? More revenue? Longer usage time? More clicks? More likes? More comments? Better comments?
- Usually many metrics, why should loss-lowering improve all of them?
- Often, even the predictions don't work in production
  - Missing *counterfactuals* in fit set
  - I.e., fit on data from old controller, run in new controller
  - FB ML Field Guide, Ep. 6: "online-offline gap" [2]

Experimentation alternatives

Simulation

Experiment	Simulation
Measures Business Metrics	Measures Business Metrics
Slow	Fast
Business risks (ex., damage, losing money)	No business risks
Accurate	Biased

There's the catch

Experimentation alternatives
Simulation

- Controllers optimized on simulation fail in the real world due to simulator bias
- Bias due to
  - approximation / modeling
  - missing counterfactual data
- Evolutionary robotics: “reality gap”
- Reinforcement learning: “out of task”
- Tesla Autonomy Day: “unknown unknowns”

Experimentation alternatives

Complementary

- Domain knowledge: use to generate hypothesis, watch out for big risks
- Prediction quality: useful sub-goal, weed out bad ideas quickly; can mitigate missing counterfactual problem (but still need to translate into business metrics)
- Simulation: useful sub-goal, weed out bad ideas quickly

Experimental methods

Most of this course

- Reduce experimentation cost, increase quality of comparisons
- Methods:
    - A/B testing
    - Multi-armed bandits
    - Response surface modeling
    - Contextual bandits
    - Bayesian optimization

Summary

- Business metric improvement is your goal.
- Measure improvements in business metrics with experiments.
- Experimental methods minimize experimentation costs.
- Experiments are the most accurate and reliable way to decide whether to modify a system.

References

[1] https://ai.stanford.edu/~ronnyk/ExPThinkWeek2009Public.pdf

[2] https://research.facebook.com/blog/2018/05/the-facebook-field-guide-to-machine-learning-video-series/