Design and analysis of large scale experiments

- "Analytics" or "data science" are often called upon to evaluate the potential impact of proposed management actions before they are carried out or to find "optimal" such actions.
- This can be framed as estimation of causal effect of interventions.
- Techniques for estimation of causal effects without experimentation are quite fragile and not well suited to routine industrial applicaton.

Mathematical model

A useful mathematical formalism is where repeatedly:

1. Context vector x is revealed to us.

2. Policy π chooses action $a \in A$, conditioned on x: $a \sim \pi(a \mid x)$.

3. Reward r_a is revealed and added to our cumulative reward.

The goal is to chose a policy π to maximize the cumulative (equivalently, average) reward.

The main distinction from the general model-free control setting is the absense of hidden state, past actions do not affect future contexts or rewards.

The distinction from classical analysis of experiments is the dependence of the choice of action on side information and explicit notion of reward.

"Offline" version

In the offline (as opposed to "online" or sequential) scenario, we are given a log $(x_i, a_i, r_i, p_i)_{i=1}^T$ of observed actions, contexts and rewards collected under some known randomised policy.

We wish to do two things:

- Estimate the average reward under a given arbitrary policy by reusing logged data. We also wish to calculate the variability in this estimate.

- Choose the best policy to use going foward by optimizing over average reward. Estimate its statistial properties.

This problem has been studied largely independently in a few different settings, e.g. dynamic treatment regime optimisation, web advertising (contextual bandits), "uplift modelling", revenue optimisation and reinforcement learning.

Policy estimation

Suppose we want to use logged data $(x_i, a_i, r_i, p_i)_{i=1}^T$ to estimate the average reward of a new policy π'. These presents a challenge in that a_i does not neccessarily equal $\pi'(x_i)$.

A naive approach (direct method) is to use the log to construct a reward estimator $\hat{r}(a, x)$ (e.g. a simple regression model) and then to evaluate average reward of π' as follows:

$$V_{\mathsf{DM}}(\pi') = \frac{1}{T} \sum_{i=1}^{T} \sum_{a \in A} \pi'(a \,|\, x_i) \hat{r}(a, x_i).$$

This estimate, however, can suffer from (undetectable) bias.

Policy estimation — continued

An unbiased estimate can, however, be obtained using inverse propensity scoring:

$$V_{\mathsf{IPS}}(\pi') = \frac{1}{T}\sum_{i=1}^{T} r_i \frac{\pi'(a_i \,|\, x_i)}{p_i}.$$

This also suggests that it is sufficient to run a (uniformly) randomised policy in the real world to credibly evaluate the effectivenes of *any* rule governing the choices of interventions.

Doubly robust estimator

Both the direct method and the inverse propensity scoring method can be combined to form the "doubly robust" estimate[1]:

$$V_{\mathsf{DR}}^{\pi} = \frac{1}{T}\sum_{i=1}^{T}\left[\sum_{a'\in A}\pi'(a'\,|\,x_i)\hat{r}(a',x_i) + \frac{\pi(a_i\,|\,x_i)}{p_i}\big(r_i - \hat{r}(x_i,a_i)\big)\right].$$

This estimator can be shown to be unbiased if either of V_{IPS} or V_{DM} is unbiased and generally has lower variance than V_{IPS}.

[1] http://arxiv.org/abs/1103.4601

Policy optimisation

Mirroring the earlier development for policy evaluation, we can estimate $\pi(x)$ is by writing it as:

$$\pi(x) = \underset{a \in A}{\operatorname{argmax}}\ \hat{r}(a, x),$$

where $\hat{r}(a, x)$ is a suitable reward estimator. This approach can suffer from dramatic undetectable bias, especially when combined with the "direct method" of evaluation.

Another approach is to directly estimate the policy $\pi(x)$ by reformulating the task as a multiclass classification problem with weighted $0/1$ loss.

Logged data needs to be in the form $\left\{\left(x_i, a_i, \frac{r_i}{p(a_i|x)}\right)\right\}_{i=1}^{T}$, where a_i are the labels and the ratios $\frac{r_a}{p(a|x)}$ are instance weights. This allows us to leverage

reasonably well understood methods for controlling model complexity to avoid overfitting.

Finally, we can use a hybrid method. We replace every instance (x_i, a_i, r_i, p_i) collected under the exploration policy with $|A|$ instances of the form:

$$\left(x_i, a', \frac{(r_i - \hat{r}(a_i, x_i))\mathbb{I}(a' = a_i)}{p(a_i|x_i)} + \hat{r}(a', x_i) \right), \quad a' \in A$$

and again use a weight sensitive multiclass classifier with a as the label.

Conclusions

- A powerful framework for analysing certain types of business problems in insurance with the ability to directly estimate financial effects and perhaps should become a part of standard actuarial methodology.

- Certain extensions are of particular relevance, e.g. the ability to deal with delayed feedback and dependencies/risk accumulation.