Projeto final – Algoritmos e Incerteza
Smart "Predict, then Optimize"

Tomás Gutierrez

PUC - Rio

26 de novembro de 2018

Sumário

Referência

- Artigo: Smart "Predict,then Optimize"
- Autores:
  - Adam N. Elmachtoub (IEOR – Columbia University, NY)
  - Paul Grigas (IEOR – University of California, CA)

Introdução

- O que é feito em geral? Um passo de cada vez.
- Técnicas de ML, estatística etc para realizar previsões.
- Modelos de otimização para tomada de decisão (quando possível ...).
- Exemplo de roteamento de veículos:
  - Modelo de ML é calibrado para predizer tempo de rota em cada aresta da rede.
  - Otimização para encontrar rotas próximas do "ótimo".
- O que não é feito em geral?

Introdução

- A estrutura do problema de otimização não é levada em consideração.
- Nova metodologia fundamentalmente mantêm o paradigma sequencial (prever depois otimizar), porém não separa as etapas por completo.
- Agora, qualidade medida pela consequência da decisão tomada: *SPO loss function*.

O framework SPO

- Problema de otimização de interesse com função objetivo linear.
- Vetor de custos desconhecido mas estimado a partir dos dados, por um modelo de ML (modelo de previsão).
- Modelo de previsão selecionado de forma a minimizar alguma *loss function*, que quantifica o erro de previsões incorretas.
- Dados históricos para previsão.

Formalmente

- $$\begin{aligned} P(c) : z^*(c) := \text{minimizar}_w \quad & c^T w \\ \text{sujeito a} \quad & w \in S \end{aligned}$$

- $w \in \mathbb{R}^d$: variáveis de decisão,
- $c \in \mathbb{R}^d$: vetor custo,
- $S \subseteq \mathbb{R}^d$: conjunto não vazio compacto e convexo; região viável
- $P(c)$: problema pode ser descrito pelo vetor custo.
- Oráculo retorna $w^*(c)$: solução para cada input.

Temos ainda

- *training data* : $(x_1, c_1), (x_2, c_2), \ldots, (x_n, c_n)$, $x_i \in \mathcal{X}$ representando informações auxiliares associadas com c_i.
- Uma classe $\mathcal{H}$ de modelos de previsão de c, $f : \mathcal{X} \to \mathbb{R}^d$, para a qual temos uma previsão $\hat{c} := f(x)$ associada ao input x.
- Função de perda $l(.,.) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ que quantifica o erro resultante de prever $\hat{c}$ quando o correto é c.

Logo,

Devemos buscar resolver o seguinte problema de otimização:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), c_i)$$

Dado um modelo f^* escolhido, nossa regra de decisão é dada, portanto, por $w^*(f^*(x))$.

O que existe

- Em geral, $l(.,.)$ é completamente independente do problema de otimização sequencial à previsão.
- A estrutura de $P(c)$ não é levada em consideração.
- Por exemplo: $l(\hat{c}, c) = \frac{1}{2}||\hat{c} - c||_2^2$. Ainda, se $\mathcal{H}$ é um conjunto de funções lineares, nos reduzimos ao problema clássico de MQO.

Notação

- p: dimensão do vetor de *features*
- d: dimensão do vetor de decisão e do vetor de custos
- n: tamanho do *training sample*
- $W^*(c) := \arg\min_{w \in S} c^T w$: conjunto de soluções ótimas
- $w^*(.) : \mathbb{R}^d \to S$ um oráculo para solucionar $P(c)$.
  Observemos que $w^*(c)$ é um elemento arbitrário de $W^*(c)$.
- Seja ainda $\xi_S(.) : \mathbb{R}^d \to \mathbb{R}$ a função suporte de S, definido por $\xi_S(c) := \max_{w \in S}\{c^T w\}$.
- Notemos que $\xi_S(c) = -z^*(-c) = c^T w^*(-c)$, convexa.

SPO Loss functions

De acordo com o racional desenvolvido até aqui, a *true SPO loss function*, dado um oráculo w^* e uma previsão $\hat{c}$ é dada por:

$$l_{SPO}^{w^*}(\hat{c}, c) := c^T w^*(c) - z^*(c)$$

obs: A dependência em relação ao oráculo em geral não é problema, visto que esperamos que $W^*(c)$ seja um *singleton*, i.e., solução única. Porém, podemos reescrever:

$$l_{SPO}(\hat{c}, c) := \max_{w \in W^*(\hat{c})} \{c^T w\} - z^*(c)$$

Na prática

Objetivamente, queremos o modelo de previsão que minimize o seguinte problema de otimização (SPO loss function empírica):

$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l_{SPO}(f(x_i), c)$

- Difícil solução, teórica e prática
- Necessária aproximação viável

Aproximando a SPO loss function

- Dada a intratabilidade da função original, desenvolveremos agora uma aproximação, a qual chamamos SPO_+ *loss function*: l_{SPO_+}
- Idealmente, ao minimizarmos a l_{SPO_+} empírica, desejamos estar minimizando a l_{SPO} empírica.

- $l_{SPO+}(\hat{c}, c) = \xi_S(c - 2\hat{c}) + 2\hat{c}^T w^*(c) - z^*(c)$

Aproximando a SPO loss function

- Proposição 1: A função de perda 0-1 associada com a classificação binária é um caso especial da *SPO loss function*.
- Proposição 2 : Para um vetor de custos fixo c, $l_{SPO_+}(\hat{c}, c)$ é uma função convexa em $\hat{c}$. Além disso, l_{SPO_+} é um upper bound para l_{SPO}:

  $l_{SPO} \leq l_{SPO_+} \forall \hat{c} \in \mathbb{R}^d$
- Proposição 3: A *hinge loss* é equivalente a SPO_+ *loss*, nas mesmas condições em que a *0-1 loss* é equivalente a *SPO loss*. Prova.

Consistência de Fisher

Sob algumas condições, minimizar a SPO_+ empírica é a mesma coisa que minimizar a SPO empírica.

- Distribuição de c é contínua.
- Distribuição de c é simétrica em torno da média $\bar{c}$.
- A média $\bar{c}$ tem solução única; $W^*(\bar{c})$ é um singleton.
- A região viável S não é um singleton.

Por exemplo, a distribuição Normal.

Network flow

- Minimum cost network flow problem: decisão é quanto fluxo enviar por cada aresta da rede.
- Por hipótese, temos o grafo. S representa a conservação de fluxo, capacidade e restrições necessárias no grafo.
- c não é conhecido com exatidão, mas pode ser estimado de informações como tempo, dia, comprimento, custos observados mais recentemente etc.

Shortest Path

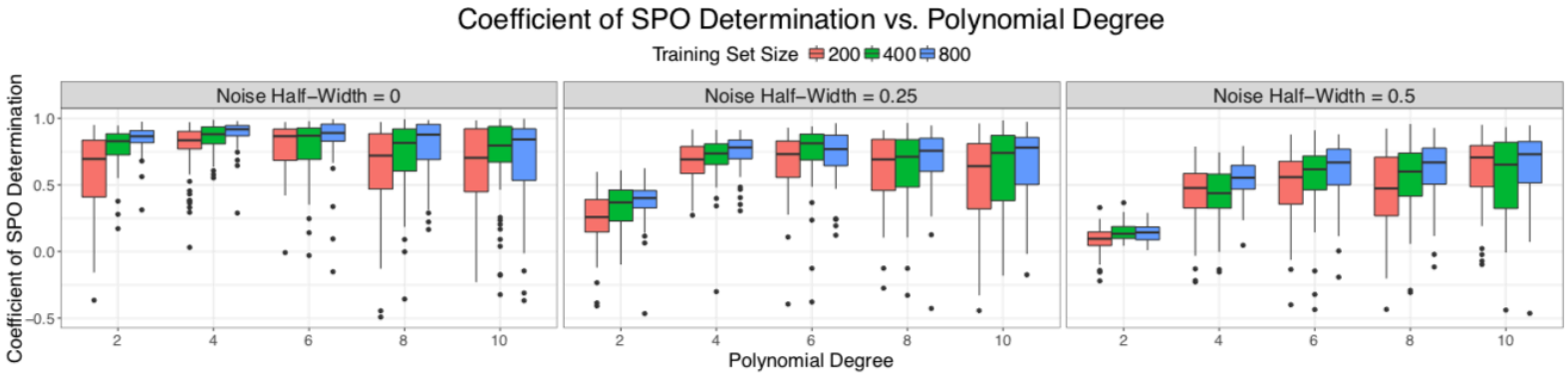
- Tradicionalmente, $\mathcal{H} = f : f(x) = Bx, B \in \mathbb{R}^{d \times p}$ e $l(\hat{c}, c) = \frac{1}{2}||\hat{c} - c||_2^2$
- Determina-se B^*, portanto, a partir de:
- $\min_B \frac{1}{n} \sum_{i=1}^{n} ||Bx_i - c_i||_2^2$
- Observemos que regularizar é restringir a classe de hipótese ainda mais.
- A regra de decisão para encontrar o fluxo ideal dado x é, portanto, $w^*(B^*x)$.

Shortest Path

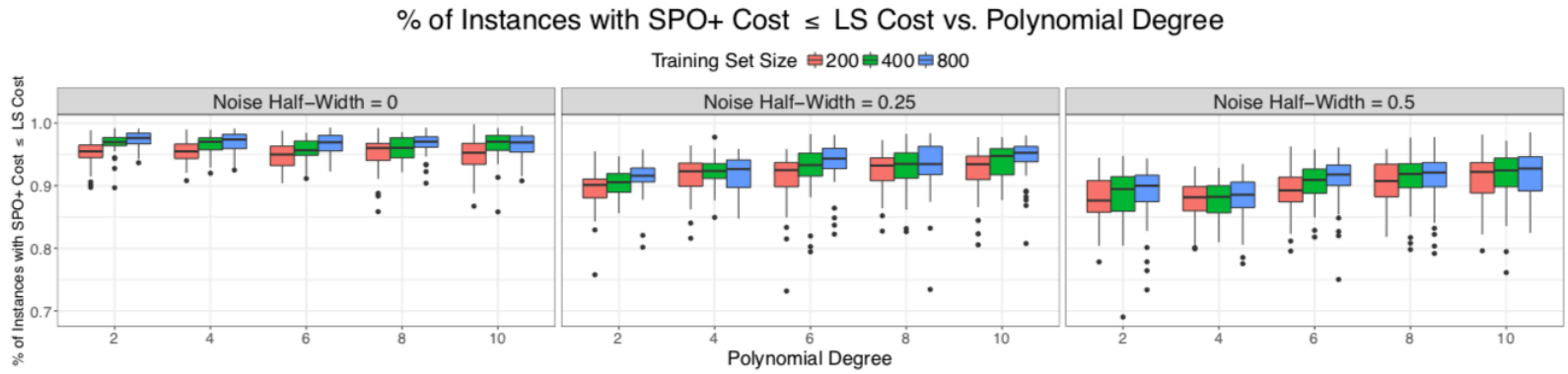
- Experimento dos autores com um grid 5×5 como topologia da rede.
- Objetivo é ir do sudoeste para o nordeste.
- 40 arestas, vão apenas para norte ou leste.
- Custos das arestas gerados sinteticamente através de um modelo sofisticado matematicamente.
- $p = 3$ features, o que leva a $pd = 120$ parâmetros a serem estimados.
- Variam o tamanho do training size $n \in \{200, 400, 800\}$.
- Variam a complexidade do modelo.
- Rodam diversas simulações para avaliar performance.

Resultados numéricos

Figure 1 Shortest path problem.



Resultados numéricos



Note. Figure showing (i) the "coefficient of SPO determination" $1 - \frac{\text{SPO+}_{\text{test}}}{\text{LS}_{\text{test}}}$ and (ii) the percentage of test set instances where the SPO+ model yields a solution with true cost less than or equal to the cost of the solution produced by the least squares model, versus the polynomial degree parameter deg, for different values of the training set size n and different values of the noise half-width parameter $\bar{\varepsilon}$.