Sampling

Браженко Дмитрий

December 5, 2017

План

Планирование эксперимента

- $X = \{x_i\}_{i=1}^{N}$ – план эксперментa
- $D = (X, Y = f(X)) = ((x_1, f(x_1)), (x_2, f(x_2)), ..., (x_N, f(x_N)))$

Что такое семплирование?

Идея

Семплирование – метод выбора подмножества наблюдаемых величин из множества, с целью выделения неких свойст исходного множества.

Пример применения

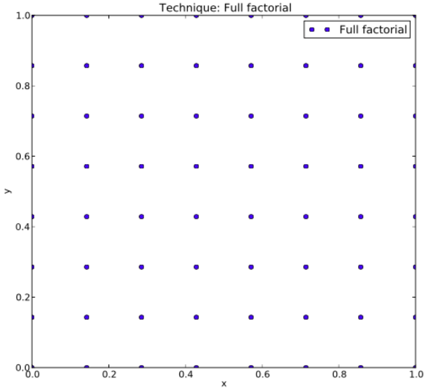
- Оценивание математического ожидания сложных вероятностных распределений:

$$E[f] = \int f(z)p(z)dz$$

$$\hat{f} = \frac{1}{L}\sum_{i=1}^{L} f(z^{(i)})$$

- **Как выбрать $\{z^{(l)}\}$?**

Полный факторный план эксперимента



Полный факторный план эксперимента

Преимущества

- Хорошо заполняет пространство
- Просто генерировать

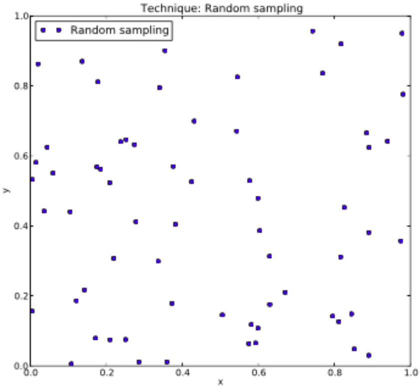
Недостатки

- Требуется очень много точек

Простая случайная выборка

Суть метода

Равномерная генерация точек в гиперкубе



Особенности простой случайной выборки

Преимущества

- Универсальность и гибкость.
- Возможность расширения с помощью добавления точек.

Недостатки

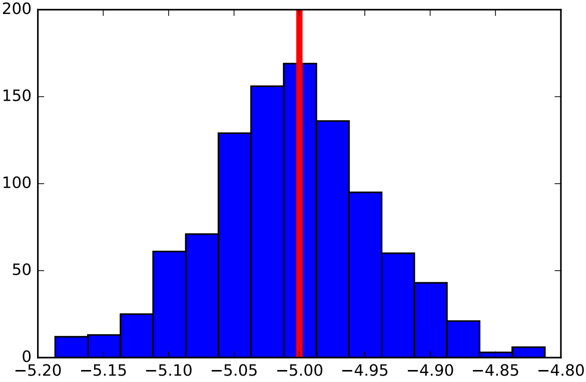
- Неравномерное заполнение

Численный пример

- Имеется 10^7 точек из распределения $N(-5, 20)$.
- Мы не можем узнать значения всех точек. Мы можем узнать значения значения 10^5 точек.
- Выбираем случайные точки из большой выборки.
- Среднее значение по подвыборке в эксперименте получилось: -4.975.

Проверка

- Эксперимент был повторен 1000 раз.



Районированная выборка

- Выборка может быть разнородной, поэтому простая случайная выборка может дать смещенные результаты.
- Выборка может состоять из нескольких кластеров разного размера. Внутри кластеров может наблюдаться однородность.
- Лучше рассмотреть пример

Численный пример

Описание

Имеется город, в котором живет 600 тысяч людей. Есть цель – узнать сколько процентов людей довольны работой мэра города. Все население делится на 3 группы:

- 100 тысяч человек – студенты (50% довольны)
- 300 тысяч человек – офисные рабочие (40% довольны)
- 200 тысяч человек – пенсионеры (90% довольны)
- *Спойлер для проверки: реально дольны работой мэрии* $58,33\%$

Мы можем опросить 1200 человек, мы можем выбирать их из различных групп.
Как провести эксперимент? *(Учитывая, что вероятность попадания пенсионера в выборку –* 0.4*, а студента или рабочего – по* 0.3*.)*

Оценивание

Простая случайная выборка

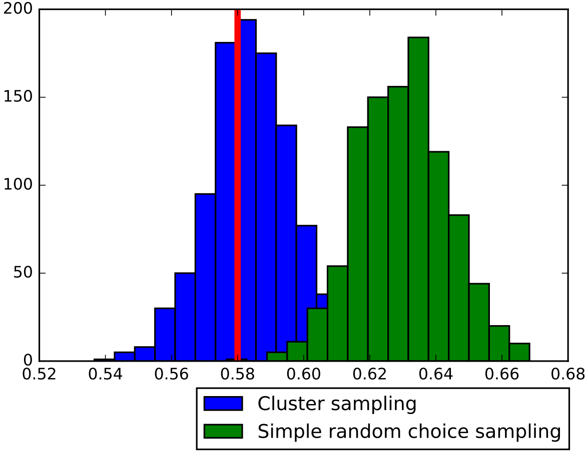
- Берем 1200 произвольных людей города и выясняем их мнение.
- В проведенном эксперименте получилось, что доля одобряющих – 63.33%.

Районированная выборка

- Берем 200 произвольных студентов, 600 офисных рабочих и 400 пенсионеров и выясняем их мнение.
- В проведенном эксперименте получилось, что доля одобряющих – в каждой группе соответсвенно: 0.515, 0.392, 0.895.
- Получается, что оценка доли людей, которые одобряют работу мэрии составляет 58.01%

Проверка наших результатов

Для проверки каждый из экспериментов был повторен 1000 раз.



Исследуется взаимосвязь между несколькими независимыми переменными и зависимой переменной. Строится поверхность следующего вида:

$$\hat{f}(x) = \alpha_0 + \sum_{i=1}^{d} \alpha_i x^i + \sum_{i,j=1,i\leq j}^{d} \beta_{ij} x^i x^j$$

$$\boldsymbol{x} = (x^1, x^2, \ldots, x^d)$$

Параметры α_i, β_{ij} настраиваются по обучающей выборке по заранее заданному набору $\boldsymbol{X} = \{x_i\}$

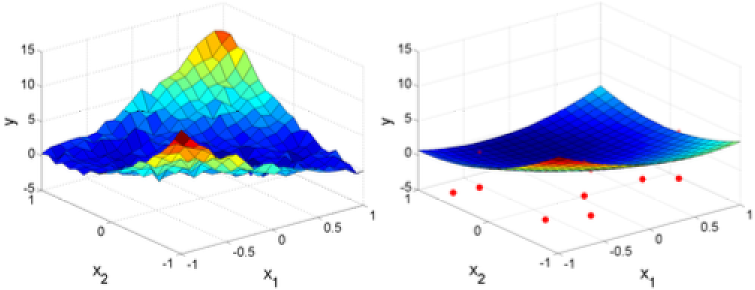
Оптимизация модели

Оптимальный дизайн RSM это оптимальный $X = \{x_i\}_{i=1}^{N}$, минимизирующий **одну** из двух величин:

- Дисперсию оценки параметров модели **(D-optimality)**
- Дисперсию прогноза модели **(IV-optimality)**

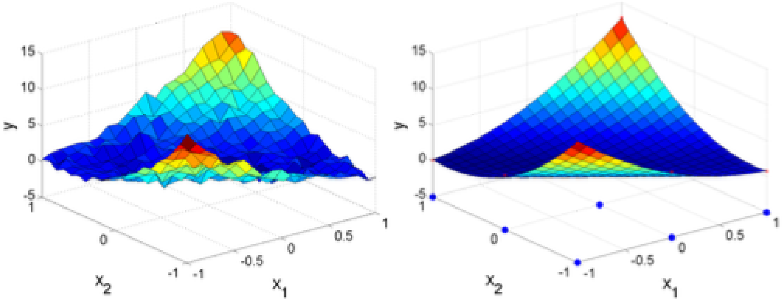
Пример расчета

Настоящая функция и RSM, обученный на случайной выборке



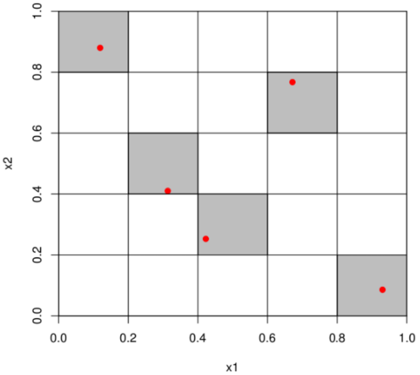
Пример расчета

Настоящая функция и RSM, обученная на **(D-)**оптимальном дизайне



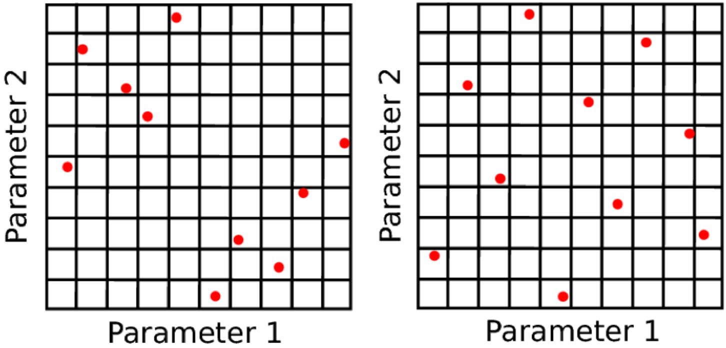
Семплирование латинским гиперкубом

Семплирование латинским гиперкубом выполняется с помощью разделения значений каждой компоненты дизайна на N равных интервалов, в каждый из которых попадает по одной точке.

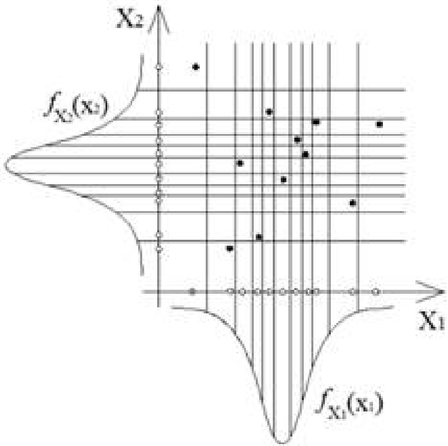


Возможные проблемы

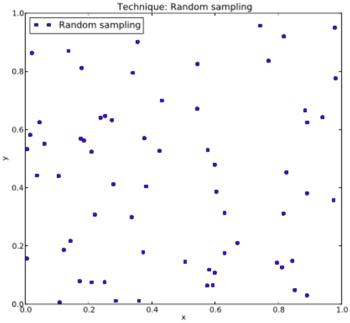
"Дыры в пространстве"...

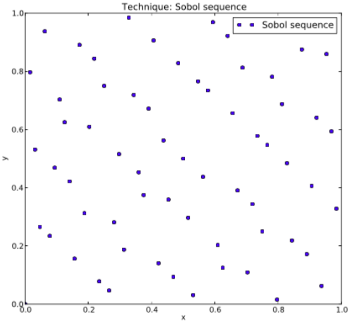


Оценивание двумерного распределения



О равномернсти





Критерии равномернсти

- Максиминное расстояние между точками

$$\rho(X) = \max_i \left(\min_{j, j \neq i} ||x_j - x_i|| \right)$$

- ϕ-метрика

$$\phi_p(X) = \left(\sum_{i<j} ||x_i - x_j||^{-p} \right)^{1/p}$$

Оптимизированный латинский гиперкуб

- Латинский гиперкуб может давать нежелательный результат
- Оптимзированный латинский гиперкуб (OLHS) генерирует много случайных гиперкубов, после чего выбирает наилучший среди них.

Преимущества

- Простота
- Равномерность на оси

Недостатки

- Медленный поиск
- Невозможно дополнить дизайн эксперимента, не нарушая исходное правило.

Выводы

- Для построения хорошей модели необязательно брать всю возможную выборку
- Важна равномерность точек
- Для дальнейшего анализа часто используется метод bootstrap. (Об этом в следующий раз)