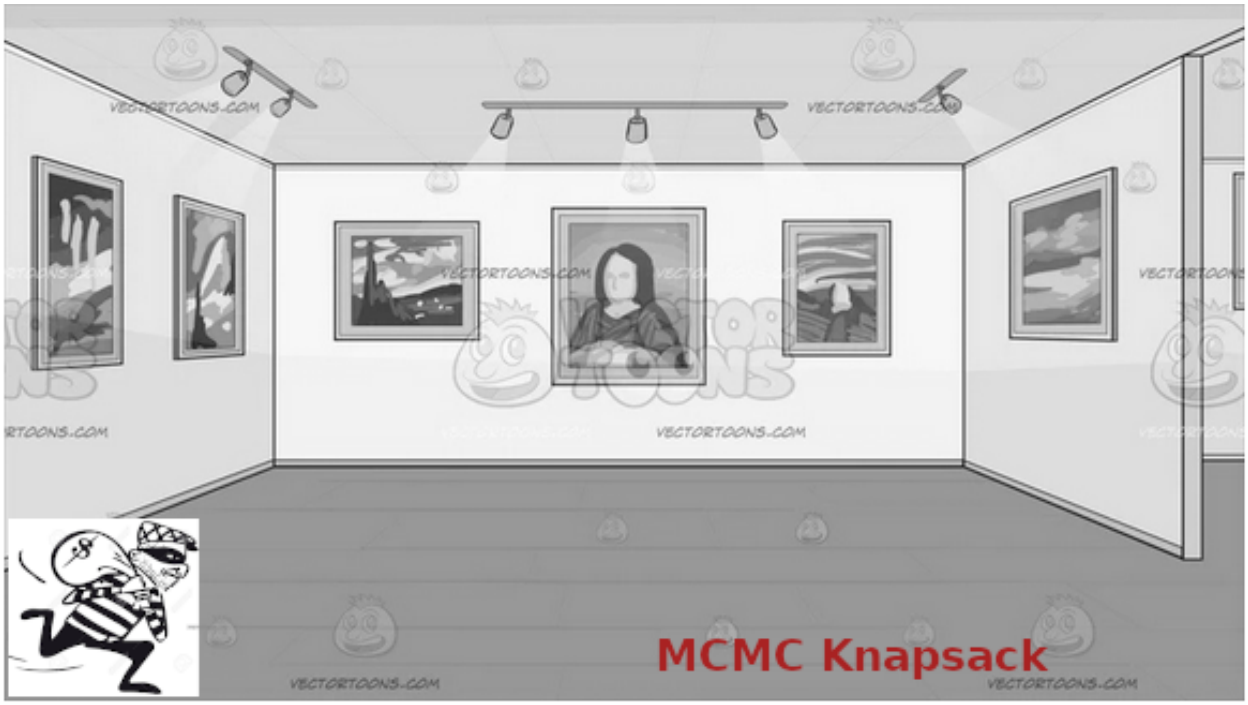


MCMC Knapsack

Krishnan Raman, Dept. of Statistics, Purdue

Introduction

Perhaps the most intensively studied problem in the discipline of Combinatorial Optimization, Knapsack Problems are a family of a NP-hard problems first proposed in the 1950s by the mathematician and father ***Tobias Dantzig,*** first solved by his son ***George Dantzig*** in 1957***!*** Most real-world selection processes such as selecting stocks[finance], food[nutrition], raw materials[manufacturing], keys[cryptography], end up as some variant of Knapsack. ***Richard Karp*** in 1972 listed Knapsack among the 21 most important NP-complete problems in his widely cited work ***Reducibility Among Combinatorial Problems***. ***David Pisinger's*** monumental 1995 Computer Science Ph.D. Thesis ***Algorithms for Knapsack Problems*** is considered to be a Knapsack Bible. In a famous comparative study of CS algorithms, Professor ***Steven Skiena*** of Stony Brook found Knapsack to be the third most useful algorithm in all of Computer Science, and the 19th most popular!

In this work, we propose an *entirely novel statistical variant* of Knapsack that is amenable to MCMC! We investigate the efficacy of our results via a visual interactive online solution.

Formal Problem Statement

A burglar breaks into an art gallery with a knapsack (a bag). The capacity of his knapsack is fixed. The burglar can steal any number of paintings he chooses, so long as the sum of the weights of the paintings do not exceed his knapsack carrying capacity. Every painting has a price. The burglar might naturally choose to steal the most expensive paintings, subject to

the weight constraint. However, it is entirely possible to pick numerous inexpensive paintings that are not heavy which add up to a huge dollar amount while satisfying the weight constraint! In other words, picking the most expensive paintings may not be the optimal strategy, as perhaps these may be heavy and he might end up with fewer dollars overall. Can we devise an optimal algorithm if *the price & weight distributions were known*?

A Trivial Example

Suppose the burglar's knapsack can only carry 5 pounds.
Paintings in the gallery: {A[$4, 5 lb], B[$1, 2 lb], C[$2, 1 lb], D[$2,2 lb]}
Most expensive painting = A, which weighs 5 pounds. If the burglar stole A, he would satisfy the knapsack weight constraint right away, but only end up with $4.
A better strategy is to steal {B,C,D} as the total weight is still {2+1+2 = 5lbs} but the dollar amount {1 + 2 + 2 = $5 > $4} is bigger than stealing the most expensive painting!

Concrete Problem Statement

An art gallery holds 100 paintings. Some are cheap, others are expensive. Some are heavy, others are light in weight. A burglar with a fixed capacity knapsack breaks into the gallery.

a. The paintings are arranged in a price pattern: cheaper paintings are displayed first, then the expensive ones. Change point (between the two sets of prices) is unknown.
b. The paintings are also arranged in a weight pattern: lighter paintings are displayed before the heavier ones. Change point (between the two sets of weights) is unknown.
c. The cheap paintings come from a probability distribution with unknown price parameters. The expensive ones come from a (different) distribution as well.
d. The heavy paintings have their own weight distribution. So do the lighter ones.
e. The price & weight of every painting is known ahead of time.

The burglar must grab as many paintings as he wants so as not to exceed his knapsack carrying capacity, while maximizing the total dollar value of the stolen paintings. How so?

Model parametrization (Unknown model parameters: α, β, σ, θ, a, b)

1. There are n (= 100) paintings, belonging to 2 price categories.
   1a. $X[j] \sim Poi(\alpha)$, j = 1.. a, are cheap paintings
   1b. $X[j] \sim Poi(\beta)$, j = a + 1 .. n, are expensive paintings

2. The n paintings also belong to 2 weight categories.
   2a. $X[j] \sim Poi(\sigma)$, j = 1 .. b, are light weight.
   2b. $X[j] \sim Poi(\theta)$, j = b+1 .. n, are quite heavy.

3. $\alpha \mid rate\_price \sim Gamma(5, rate\_price)$, $\beta \mid rate\_price \sim Gamma(5, rate\_price)$
4. $\sigma \sim Gamma(5, rate\_weight)$, $\theta \sim Gamma(5, rate\_weight)$
5. $a \sim Uniform[1.. n]$, $b \sim Uniform[1.. n]$
6. $rate\_price \sim Gamma(10,10)$, $rate\_weight \sim Gamma(10,10)$

Novelty

Traditionally, Knapsack does not involve any element of statistics. It is an exercise in pure combinatorial optimization. However, in this novel formulation, we have introduced several statistical components that might help us obtain an optimal solution. These are:

a. There is a price distribution. Cheap paintings are sampled from a Poisson model. Expensive paintings have their own Poisson distribution.
b. There is a weight distribution. Heavy paintings are sampled from a Poisson model. Lighter paintings have their own Poisson distribution.
c. Paintings are displayed in a pattern: first the cheap ones, then the expensive ones. The cutoff point is unknown, but comes from a discrete Uniform distribution.
d. Paintings are displayed in a pattern: first the lighter ones, then the heavier ones. The cutoff point is unknown, but comes from a discrete Uniform distribution.
e. The model parameters (mean of the 4 Poisson distributions) are unknown. We resort to a standard Bayesian analysis & impose a Gamma prior on each parameter.

Analysis

We have chosen conjugate Gamma priors. Via a simple application of Bayes, it can be shown that the posterior distribution of the model parameters are Gamma as well. In particular,

1. Posterior(α | rate_price) ~ Gamma(shape=5+sum(j=1..a),rate=rate_price+a)
2. Posterior(β | rate_price) ~ Gamma(shape=5+ sum(j=a+1..n), rate=rate_price+n-a)

Similar results hold for α, σ, rate_price, rate_weight.

(We have skipped the details since our focus here is on the application of MCMC/Gibbs to Knapsack, not on inference of model parameters. However, the conditional distributions necessary to carry out Gibbs Sampling for the Change Point models have been derived elsewhere: See Problem 6.4 & Problem 7.6 ***Computational Statistics, Givens & Hoeting***)

MCMC Knapsack Strategy

Suppose we implement a Gibbs Sampler to estimate the posterior distribution of the model parameters. After convergence over a suitably large number of iterations, we obtain:

a. E(α) = the sample mean of α, which is the unknown model parameter of the Poisson distribution pertaining to the price of the cheap paintings.
b. E(β) = the sample mean of β, which is the unknown model parameter of the Poisson distribution pertaining to the price of the expensive paintings.
c. E(a) = the sample mean of a, which is the cut-off point, i.e. paintings from 1 thru a are the cheap paintings, paintings from a+1 thru n=100 are the valuable paintings.

Knowing these parameters, our burglar can estimate the following quantities:

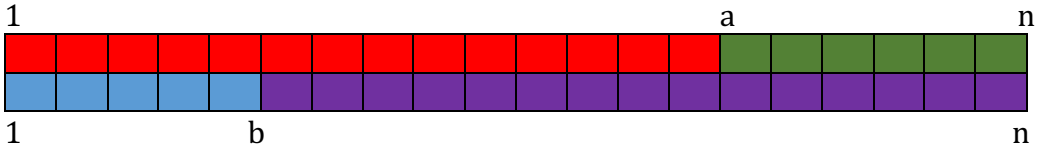
$a\alpha$ = estimated earnings from stealing all the cheap paintings.

$(n-a)*\beta$ = estimated earnings from stealing all the expensive paintings.

d. $E(\sigma)$ = the sample mean of σ, which is the unknown model parameter of the Poisson distribution pertaining to the weight of the lighter paintings.
e. $E(\theta)$ = the sample mean of θ, which is the unknown model parameter of the Poisson distribution pertaining to the weight of the heavier paintings.
f. $E(b)$ = the sample mean of b, which is the cut-off point, i.e. paintings from 1 thru b are lightweight, paintings from b+1 thru n=100 are heavy.

Knowing these parameters, our burglar can estimate the following quantities:
$b\sigma$ = estimated weight of all the cheap paintings.
$(n-b)^* \theta$ = estimated weight of all the expensive paintings.

Our burglar can formulate his strategy thusly:
Assume the weight of his knapsack = W.
The burglar tries to fill the knapsack with mostly expensive paintings.



We display the same n paintings above, in 2 different ways:

By Price:
Red = Cheap Paintings.
Green = Expensive Paintings.
a is the cutoff point between Cheap & Expensive.

By Weight:
Blue = Light Paintings.
Purple = Heavy Paintings.
b is the cutoff point between Light & Heavy.

MCMC Knapsack Algorithm

Step 1. If $W > (n-a)^* \theta$, the burglar can steal all the expensive paintings!
The remaining space i.e. $W - (n-a)^* \theta = W'$,
can be allotted to as many cheap paintings as he can steal.
Approximate number of cheap paintings he can steal= $W'/\sigma = c$

Step 2. Skip max(0, (a-c)) paintings. Steal the rest of the paintings in the gallery.

Step 3. On the other hand, if $W < (n-a)^* \theta$, the burglar cannot steal all the expensive paintings. Now the burglar must be really careful. He should skip the first ***a*** paintings entirely, because they are cheap, so not worth stealing. He can then steal as many of the ***(n-a)*** expensive paintings as he can, until he hits the weight constraint W.

A Remarkable Coincidence

It is quite remarkable that the cutoff point ***b*** does not have a major role in the algorithm, but the weight parameters θ & σ, which are separated by ***b***, are vital to the decision-making process! Equally remarkable is the fact that α and β have no role to play, but the cutoff point ***a*** that separates α and β is vital to the decision-making process!

These coincidences can be explained by the symmetry of the problem space. The n paintings can be partitioned by price, at cutoff point ***a***. Or they can be partitioned by weight, at cutoff point ***b***. Cutoff point ***a*** is vital since we hope to maximize our earnings, so it is important to know how many expensive paintings there are, and where they begin. On the other hand, the price of the paintings themselves don't matter, because we only care about the total sum. So, α and β have no role to play. But the amount of weight we can carry is wholly dependent on weight parameters θ & σ, which are separated by ***b.*** So, while ***b*** doesn't matter**,** θ & σ do!

Implementation

To demonstrate the efficacy of the MCMC Knapsack algorithm, we have 2 burglars who break into the gallery. The first burglar uses a Gibbs Sampler to estimate the model parameters as described, and follows the steps in the MCMC Knapsack algorithm. The second burglar is a Statistical skeptic, and does not use any algorithms. He simply grabs paintings at random until his knapsack is full. We compare the earnings of the knapsack burglar to the skeptic.

The Gibbs Sampler and the MCMC Knapsack algorithm have been implemented from scratch in JavaScript and can be run at: ***https://krishnanraman.github.io//mcmcknapsack.html***
Total lines of code: ~400 lines

References

1. *George B Dantzig, Discrete-Variable Extremum Problems, Operations Research, 1957*
2. *Richard Karp, Reducibility Among Combinatorial Problems, 1972*
3. *David Pisinger, Algorithms for Knapsack Problems, Ph.D. Thesis, 1995*
4. *Steven Skiena, Who is Interested in Algorithms and Why? Lessons from the Stony Brook Algorithm Repository, 1999*