Применение методов обучения с подкреплением к задаче алгоритмической торговли

Сокурский Юрий
Научный руководитель: Ветров Д.П.

МГУ им. М.В. Ломоносова, факультет ВМК

24 Апреля 2015

Алгоритмическая торговля, постановка задачи

Основные понятия:

- заявка
- книга заявок
- сделка

Покупка	Цена	Продажа
1	1 609,11	
1	1 609,31	
10	1 609,33	
22	1 609,50	
80	1 609,51	
27	1 610,00	
77	1 610,01	
55	1 610,51	
280	1 610,52	
	1 610,99	15
	1 611,00	532
	1 611,01	20
	1 611,05	42
	1 611,50	25
	1 611,53	70
	1 611,90	1 300
	1 612,00	1 235
	1 612,78	1
	1 612,90	600

Мотивация

Почему методы обучения с подкреплением подходят для задачи алгоритмической торговли?

- вознаграждение приходит с задержкой
- online обучение

Обучение с подкреплением, постановка задачи

S - множество состояний среды

A - множество действий агента

Игра агента со средой:

- Инициализация стратегии $\pi_1(a|s)$, и состояния среды s_1
- для всех t = 1,...,T,..
- агент выбирает действие согласно стратегии $a_t \sim \pi_t(a|s_t)$
- среда генерирует вознаграждение $r_{t+1} \sim p(r|a_t, s_t)$ и новое состояние $s_{t+1} \sim p(s|a_t, s_t)$
- агент корректирует стратегию $\pi_{t+1}(a|s)$

Цель: максимизировать вознаграждение: $R_t = \sum_{\tau=t}^{+\infty} \gamma^{\tau-t} r_\tau$

$\gamma \in (0, 1)$ - коэффициент дисконтирования, дальновидность агента

Алгоритмическая торговля, в контексте обучения с подкреплением

Соответствие задаче обучения с подкреплением.

- Среда - рынок
- Агент - торгующий алгоритм
- Действие агента - выставление заявки на покупку или продажу
- Состояние среды - переменная, описывающая рынок

Предположение: действия агента не влияют на среду:
$s_{t+1} \sim p(s|s_t)$

Алгоритм Q-Learning.

Функция ценности действия a в состоянии s при стратегии π:

$Q^{\pi}(s,a) = \mathbb{E}_{\pi}(R_t|s_t = s, a_t = a) =$

$\mathbb{E}(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}) = \mathbb{E}(r_{t+1} + \gamma Q^{\pi}(s_{t+1}, a_{t+1}))$

Игра агента со средой:

- Инициализация стратегии $\pi_1(a|s)$, и состояния среды s_1
- для всех t = 1,...,T,..
- агент выбирает действие $a_t \sim \pi_t(a|s_t)$
- среда генерирует вознаграждение $r_{t+1} \sim p(r|a_t, s_t)$ и новое состояние $s_{t+1} \sim p(s|a_t, s_t)$
- $Q(s_t, a_t) :=$ $Q(s_t, a_t) + \alpha_t(r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t))$

Функция потерь, при аппроксимация функции ценности:

$L_i(\theta_i) = \mathbb{E}_{s,a,r}[\mathbb{E}_{s'}(r + \gamma \max_{a'} Q(s', a'|\theta_{i-1}) - Q(s, a|\theta_i))^2] \rightarrow \min$

Проблемы и решения

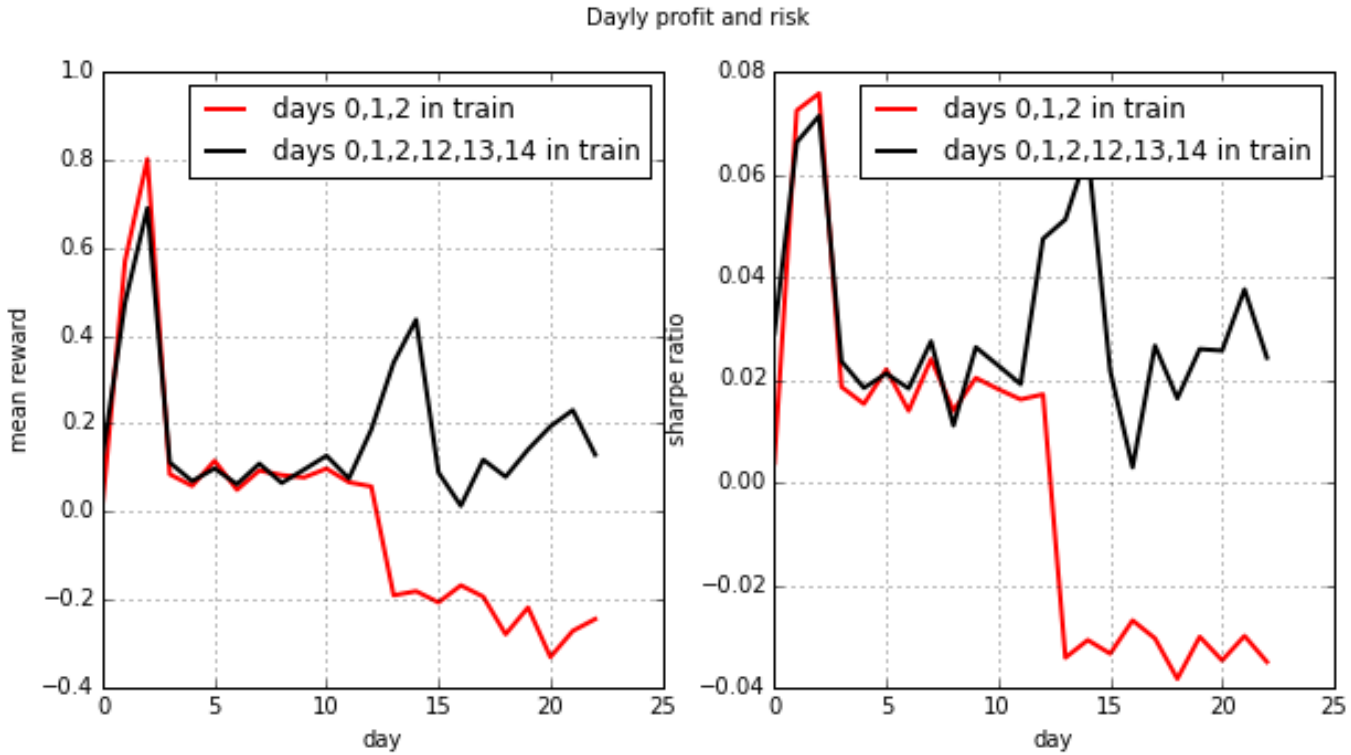
1. Проблема: генерация случайных реализаций троек (s,a,r) ресурсоемка
   Решение: отсутствие влияния на рынок ($s_{t+1} \sim p(s|s_t)$) и возможность тестироваться на исторических данных позволяет перебирать на каждом шагу все действия
2. Проблема: переобучение на стратегиях с длинными сделками
   Решение: разделение стратегии на две части: открывающую и закрывающую сделку; ограничение на промежуток между открытием и закрытием сделки (t_{max})

Проблемы и решения

1. Проблема: не устойчивая сходимость функции потерь у стратегии, закрывающей сделку
   Решение: введена функция ценности Q_t для закрытия сделки в момент времени $t \in [1, t_{max}]$
2. Проблема: переобучение на глобальные тренды
   Решение: добавление перевчрнутой выборки

Эксперимент 1

Подбор объема обучающей выборки



На защиту выносится

- Предложен и реализован метод обучения с подкреплением для задачи алгоритмической торговли.
- Экспериментальное исследование алгоритма и сравнение с альтернативными методами на реальных данных.

TO DO

- сравнение с базовой стратегией
- дообучение стратегии (окно по дням)
- обучение на разных торговых инструментах

Алгоритм Q-Learning.

Функция ценности действия a в состоянии s при стратегии π:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}(R_t|s_t = s, a_t = a) = \mathbb{E}(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s, a_t = a)$$

Построение стратегии:

- $\epsilon - greedy$: $A_t(s) = argmax_{a \in A} Q_t(s,a)$.
  $\pi_{t+1}(a|s) = \frac{1-\epsilon}{|A_t(s)|}[a \in A_t] + \frac{\epsilon}{A/A_t}$
- SoftMax: $\pi_{t+1}(a|s) = \frac{exp(Q_t(s,a)/\tau)}{\sum_{a'} exp(Q_t(s,a')/\tau)}$