Интернет-математика: конкурсы по машинному обучению

Павел Браславский, Андрей Гулин,
Павел Карпович, Денис Расковалов

Москва
2010

Аннотация

В докладе будет рассказано о двух конкурсах по машинному обучению: ИМАТ-2009 и ИМАТ-2010. В рамках первого конкурса(ИМАТ-2009) перед участниками была поставлена задача построения функции ранжирования поисковой системы на основе оценок релевантности для пар "запрос-документ". В 2010 году целью конкурса будет прогнозирование загруженности дорог мегаполиса M. в часы пик. Мы расскажем о самих задачах, способах оценки решений участников и опыте проведения ИМАТ-2009.

Содержание

- ИМАТ-2009
  - Задача конкурса - ранжирование поисковой системы
  - Оценка участников
  - Победитель
- ИМАТ-2010
  - Задача конкурса - предсказание загруженности дорог мегаполиса M
  - Оценка участников

ИМАТ-2009 - Задача и данные

Задача: построить функцию ранжирования, которая упорядочивает документы по степени их соответствия поисковому запросу.

Данные:

- Факторы ранжирования для пар "запрос, документ". Для каждой пары приведено по 245 числовых признаков.
- Оценки релевантности для каждой пары - числа из диапазона $[0, 4]$(0 - нерелевантно).
- Данные не содержат оригинальных запросов и ссылок на оригинальные документы.

ИМАТ-2009 - Задача и данные

Задача: построить функцию ранжирования, которая упорядочивает документы по степени их соответствия поисковому запросу.

Данные:

- Факторы ранжирования для пар "запрос, документ". Для каждой пары приведено по 245 числовых признаков.
- Оценки релевантности для каждой пары - числа из диапазона $[0, 4]$(0 - нерелевантно).
- Данные не содержат оригинальных запросов и ссылок на оригинальные документы.

ИМАТ-2009 - Задача и данные

Задача: построить функцию ранжирования, которая упорядочивает документы по степени их соответствия поисковому запросу.

Данные:

- Факторы ранжирования для пар "запрос, документ". Для каждой пары приведено по 245 числовых признаков.
- Оценки релевантности для каждой пары - числа из диапазона $[0, 4]$(0 - нерелевантно).
- Данные не содержат оригинальных запросов и ссылок на оригинальные документы.

ИМАТ-2009 - Задача и данные

Задача: построить функцию ранжирования, которая упорядочивает документы по степени их соответствия поисковому запросу.

Данные:

- Факторы ранжирования для пар "запрос, документ". Для каждой пары приведено по 245 числовых признаков.
- Оценки релевантности для каждой пары - числа из диапазона $[0, 4]$(0 - нерелевантно).
- Данные не содержат оригинальных запросов и ссылок на оригинальные документы.

ИМАТ-2009 - Задача и данные

Данные разделены на два файла: обучающее и тестовое множества.

- В обучающем множестве представлены факторы и оценки релевантности для 97 290 пар "запрос, документ"(9 124 запроса)
- Тестовое множество: 115 643 пар "запрос, документ". Оценки для тестового множества не предоставляются. Тест разделен на два подмножества:
  - публичная оценка - 21 103 пары,
  - финальная оценка - 94 540 пар.

Загружать можно любое число решений. Результат на "публичном тесте" виден сразу. Оценка на финальном множестве открывается один раз в конце проведения конкурса и по ней выявляется победитель.

ИМАТ-2009 - Задача и данные

Данные разделены на два файла: обучающее и тестовое множества.

- В обучающем множестве представлены факторы и оценки релевантности для 97 290 пар "запрос, документ"(9 124 запроса)
- Тестовое множество: 115 643 пар "запрос, документ". Оценки для тестового множества не предоставляются. Тест разделен на два подмножества:
  - публичная оценка - 21 103 пары,
  - финальная оценка - 94 540 пар.

Загружать можно любое число решений. Результат на "публичном тесте" виден сразу. Оценка на финальном множестве открывается один раз в конце проведения конкурса и по ней выявляется победитель.

ИМАТ-2009 - Задача и данные

Конкурс проводился с 10 марта по 15 мая 2009 года.

После завершения конкурса данные и интерфейс для загрузки решений остаются открытыми. Финальная оценка обновляется один раз в неделю по последнему загруженному решению.

ИМАТ-2009 - Задача и данные

Конкурс проводился с 10 марта по 15 мая 2009 года.

После завершения конкурса данные и интерфейс для загрузки решений остаются открытыми. Финальная оценка обновляется один раз в неделю по последнему загруженному решению.

ИМАТ-2009 - Оценка

DCG - discounted cumulative gain

$$DCG(ranking\ for\ query\ q) = \sum_{j=1}^{N_q} \frac{rel_j}{log_2 j + 1}$$

N_q - число документов для запроса, rel_j - релевантность документа на позиции j.

ИМАТ-2009 - Результаты

Рейтинг

Таблица объединяет финальный рейтинг конкурса (на 15.05.2009) и новые результаты. Подробнее о задаче и методике оценки конкурсе и раздел Задачи и данные.

Команда	Время последней загрузки	Количество попыток	Последний результат (публичная оценка)	Финальный результат
Всем чмоке в этом чате :)	14.02.2010 (19:03 GMT+03)	4	4.283924	*4.133886*
Joker	05.09.2009 (05:07 GMT+03)	2	4.283317	*4.151528*
Сам	21.02.2010 (20:57 GMT+03)	187	4.282790	*4.147628*
Simple	21.02.2010 (21:18 GMT+03)	531	4.282597	*4.135317*
-F	02.12.2009 (16:44 GMT+03)	2	4.281325	*4.145202*
alexeigor	07.05.2009 (17:02 GMT+03)	118	4.280676	*4.141230*
depechemode	29.10.2009 (01:19 GMT+03)	30	4.278378	*4.142855*
Победа	17.03.2009 (16:25 GMT+03)	3	4.276001	*4.139854*
ACGT	15.05.2009 (14:03 GMT+03)	21	4.274666	*4.128807*
stohastic	25.10.2009 (23:37 GMT+03)	819	4.274414	*4.129173*
RelevanceDoesMatter	22.02.2010 (17:54 GMT+03)	378	4.272542	—
WoodWeb	22.04.2009 (23:09 GMT+03)	12	4.267894	*4.127512*
Nordic	15.05.2009 (23:37 GMT+03)	4	4.266904	*3.857102*
stohastic	15.05.2009 (23:43 GMT+03)	176	4.266712	*4.118830*
Test	15.05.2009 (23:45 GMT+03)	58	4.264024	*3.859052*
ZENIT	15.05.2009 (23:20 GMT+03)	206	4.259964	*4.117877*
MysteriousGuest	18.01.2010 (16:00 GMT+03)	8	4.259078	*4.134784*
Euclid	08.05.2009 (21:46 GMT+03)	40	4.257802	*4.122558*

ИМАТ-2009 - Победитель

ИМАТ-2009

Подобные коркурсы:

- LETOR
  http://research.microsoft.com/enus/um/beijing/projects/letor/
- Yahooo challenge 2010
  http://learningtorankchallenge.yahoo.com/

ИМАТ-2010 - Задача и данные

Задача: предсказание загруженности дорог мегаполиса M.

Данные:

- Преобразованный граф дорог мегаполиса M.
- Архив с данными о загруженности за 30 дней: время с 16.00 до 22.00. В архиве содержатся данные о скорости на части дорог.
- Данные о движении в 31-й день: время с 16.00 до 18.00.

Задание состоит в предсказании загруженности дорог с 18.00 до 22.00 в 31-й день.

ИМАТ-2010 - Задача и данные

Задача: предсказание загруженности дорог мегаполиса M.

Данные:

- Преобразованный граф дорог мегаполиса M.
- Архив с данными о загруженности за 30 дней: время с 16.00 до 22.00. В архиве содержатся данные о скорости на части дорог.
- Данные о движении в 31-й день: время с 16.00 до 18.00.

Задание состоит в предсказании загруженности дорог с 18.00 до 22.00 в 31-й день.

ИМАТ-2010 - Задача и данные

Задача: предсказание загруженности дорог мегаполиса M.

Данные:

- Преобразованный граф дорог мегаполиса M.
- Архив с данными о загруженности за 30 дней: время с 16.00 до 22.00. В архиве содержатся данные о скорости на части дорог.
- Данные о движении в 31-й день: время с 16.00 до 18.00.

Задание состоит в предсказании загруженности дорог с 18.00 до 22.00 в 31-й день.

ИМАТ-2010 - Задача и данные

Задача: предсказание загруженности дорог мегаполиса M.

Данные:

- Преобразованный граф дорог мегаполиса M.
- Архив с данными о загруженности за 30 дней: время с 16.00 до 22.00. В архиве содержатся данные о скорости на части дорог.
- Данные о движении в 31-й день: время с 16.00 до 18.00.

Задание состоит в предсказании загруженности дорог с 18.00 до 22.00 в 31-й день.

ИМАТ-2010 - Задача и данные

Задача: предсказание загруженности дорог мегаполиса M.

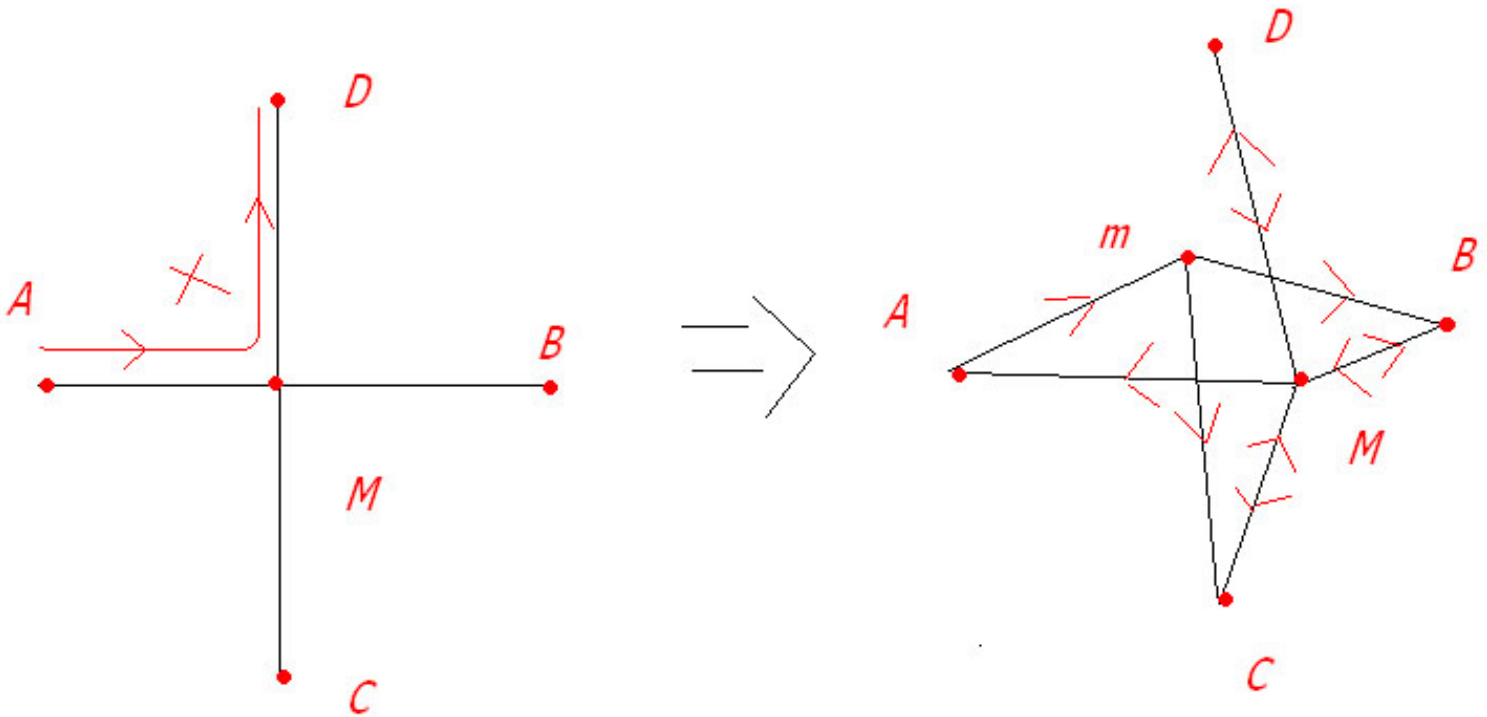
Данные:

- Преобразованный граф дорог мегаполиса M.
- Архив с данными о загруженности за 30 дней: время с 16.00 до 22.00. В архиве содержатся данные о скорости на части дорог.
- Данные о движении в 31-й день: время с 16.00 до 18.00.

Задание состоит в предсказании загруженности дорог с 18.00 до 22.00 в 31-й день.

ИМАТ-2010 - Данные (граф)

"Естественный" граф дорог мегаполиса преобразован в ориентированный граф G, в котором продублированы некоторые вершины и добавлены новые дуги для учета правил дорожного движения. Преобразованный граф G обладает тем свойством, что в нем нет запрещенных маршрутов - можно ездить везде.



ИМАТ-2010 – Данные (граф)

Описание графа содержится в трех файлах:

- Файл вершин(vertices.txt): *<id_vertex> <id_group>*. Вершины из одной группы обозначают один и тот же перекресток. (139 241 виртуальных - 33 029 реальных)
- Файл дуг(edges.txt):

  <id_edge> <id_edge_group> <start_vert> <end_vert>

  Дуги из одной группы соответствуют одной и той же улице. (206 260 виртуальных - 86 249 реальных)
- Файл данных про дуги(edges_data.txt):

  <id_edge_group> <length> <avg_speed>

  Длина предоставляется в метрах.

ИМАТ-2010 – Данные (граф)

Описание графа содержится в трех файлах:

- Файл вершин(vertices.txt): *<id_vertex> <id_group>*. Вершины из одной группы обозначают один и тот же перекресток. (139 241 виртуальных - 33 029 реальных)
- Файл дуг(edges.txt):

  <id_edge> <id_edge_group> <start_vert> <end_vert>

  Дуги из одной группы соответствуют одной и той же улице. (206 260 виртуальных - 86 249 реальных)
- Файл данных про дуги(edges_data.txt):

  <id_edge_group> <length> <avg_speed>

  Длина предоставляется в метрах.

ИМАТ-2010 – Данные (граф)

Описание графа содержится в трех файлах:

- Файл вершин(vertices.txt): *<id_vertex> <id_group>*. Вершины из одной группы обозначают один и тот же перекресток. (139 241 виртуальных - 33 029 реальных)
- Файл дуг(edges.txt):

  <id_edge> <id_edge_group> <start_vert> <end_vert>

  Дуги из одной группы соответствуют одной и той же улице. (206 260 виртуальных - 86 249 реальных)
- Файл данных про дуги(edges_data.txt):

  <id_edge_group> <length> <avg_speed>

  Длина предоставляется в метрах.

ИМАТ-2010 – Данные (граф)

Описание графа содержится в трех файлах:

- Файл вершин(vertices.txt): *<id_vertex> <id_group>*. Вершины из одной группы обозначают один и тот же перекресток. (139 241 виртуальных - 33 029 реальных)
- Файл дуг(edges.txt):

  <id_edge> <id_edge_group> <start_vert> <end_vert>

  Дуги из одной группы соответствуют одной и той же улице. (206 260 виртуальных - 86 249 реальных)
- Файл данных про дуги(edges_data.txt):

  <id_edge_group> <length> <avg_speed>

  Длина предоставляется в метрах.

ИМАТ-2010 – Данные ("пробки")

Пробки лежат в файле "jams.txt"(29 226 208 строк):

<id_edge_group> <day> <time> <speed>

Скорость 0 - улица стоит.

Файл с заданием:

<id_edge_group> <day> <time> ??

Предсказываем скорость.

По аналогии с ИМАТ-2009 задание делится на публичный тест и финальный.

ИМАТ-2010 – Данные ("пробки")

Пробки лежат в файле "jams.txt"(29 226 208 строк):

<id_edge_group> <day> <time> <speed>

Скорость 0 - улица стоит.

Файл с заданием:

<id_edge_group> <day> <time> ??

Предсказываем скорость.

По аналогии с ИМАТ-2009 задание делится на публичный тест и финальный.

ИМАТ-2010 - Оценка

Метрика качества - взвешенное отклонение от реальной скорости на участке дороги.

$$Err(a) = \frac{1}{n}\sum_{i=1}^{n} l_i t_i |a_i - v_i|$$

n - число предсказываемых случаев,

a_i - предсказание для скорости,

v_i - реальная скорость,

l_i - коэффициент длины (отношение длины улицы к средней длине улиц в графе),

t_i - коэффициент времени

$$1 + 0.1 < number\ of\ minutes\ from\ 18:00 > /4$$

Тише едешь - дальше будешь ?