Теоретическое задание 2

Кириленко Елена

29 октября 2018 г.

1 *Задача 1*

1)

$$E_{x,y}(y-\mu(x))^2 = E_{x,y}(y-E(y|x)+E(y|x)-\mu(x))^2 =$$
$$= E_{x,y}(y-E(y|x))^2 + E_{x,y}(E(y|x)-\mu(x))^2 + 2E_{x,y}\left((y-E(y|x))(E(y|x)-\mu(x))\right)$$

Рассмотрим последнее слагаемое, обозначим его через A. Заметим, что мы можем переписать его в следующем виде :

$$A = 2\ E_x\Big(\ E_y\ [\ (y-E(y|x))\ (E(y|x)-\mu(x))\ |\ x\ ]\ \Big)$$

Теперь заметим, что $(E(y|x)-\mu(x))$ не зависит от y так как $E(y|x)$ не зависит от y и $\mu(x)$ тоже. Тогда можем вынести этот множитель за знак мат. ожидания по y:

$$A = 2\ E_x\Big(\ (E(y|x)-\mu(x))\ *\ E_y\,[\ (y-E(y|x))\ \ |\ x\ ]\ \Big)$$

Рассмотрим последний множитель под знаком мат. ожидания по x.

$$E_y\,[\ (y-E(y|x))\ \ |\ x\ ] = E_y(y|x) - E_y(E(y|x)\ |\ x) = E_y(y|x) - E_y(y|x) = 0$$

Таким образом, получили, что

$$A = 2\ E_x\,(\ (E(y|x)-\mu(x))\ *\ 0\ ) = 0$$

И в итоге :

$$E_{x,y}(y-\mu(x))^2 = E_{x,y}(y-E(y|x))^2 + E_{x,y}(E(y|x)-\mu(x))^2$$

2) Подставим полученное в $L(\mu)$:

$$L(\mu) = E_{X^l,y^l}\Big(E_{x,y}(y-E[y|x])^2\Big) + E_{X^l,y^l}\Big(E_{x,y}(E[y|x]-\mu(x))^2\Big)$$

3) Заметим, что y и $E[y|x]$ не зависят от X^l, y^l, а поэтому $E_{x,y}(y-E[y|x])^2$ не зависит от X^l, y^l и мат. ожидание по X^l, y^l можно убрать :

$$L(\mu) = E_{x,y}(y-E[y|x])^2\ + E_{X^l,y^l}\Big(E_{x,y}(E[y|x]-\mu(x))^2\Big)$$

Видим, что первое слагаемое полученного выражения это шумовая компонента.

4) Рассмотрим второе слагаемое предыдущего разложения.

$$E_{X^l,y^l}\Big(E_{x,y}(E[y|x]-\mu(x))^2\Big) = E_{x,y}\Big(E_{X^l,y^l}(E[y|x]-\mu(x))^2\Big) =$$

$$= E_{x,y}\Big(\ E_{X^l,y^l}\big(\ E[y|x] - E_{X_l,y_l}(\mu(x)) + E_{X_l,y_l}(\mu(x)) - \mu(x)\ \big)^2\ \Big) =$$

$$= E_{x,y}\Big(\ E_{X^l,y^l}\big[E_{x_l,y_l}(\mu(x)) - \mu(x)\big]^2\ \Big) + E_{x,y}\Big(\ E_{X^l,y^l}\big[E(y|x) - E_{X_l,y_l}(\mu(x))\big]^2\ \Big) +$$

$$+2\ E_{x,y}\Big(\ E_{X^l,y^l}\big[(E_{X_l,y_l}(\mu(x)) - \mu(x)) * (E(y|x) - E_{X_l,y_l}(\mu(x)))\big]\ \Big)$$

5) Рассмотрим последнее слагаемое предыдущего выражения. Для этого рассмотрим выражение под мат.ожиданием:

$$E_{X^l,y^l}\Big[(E_{X_l,y_l}(\mu(x)) - \mu(x)) * (E(y|x) - E_{X_l,y_l}(\mu(x)))\Big] =$$

Заметим, что $E(y|x) - E_{X_l,y_l}(\mu(x))$ не зависит от X^l, y^l, поэтому можем вынести его за знак мат. ожидания по X^l, y^l. Получим:

$$= (E(y|x) - E_{X_l,y_l}(\mu(x))) * E_{X^l,y^l}\Big[(E_{X_l,y_l}(\mu(x)) - \mu(x))\Big]$$

Последний множитель здесь равен 0 :

$$E_{X^l,y^l}\Big[(E_{X_l,y_l}(\mu(x)) - \mu(x))\Big] = E_{X_l,y_l}(\mu(x)) - E_{X_l,y_l}(\mu(x)) = 0$$

В итоге получаем, что

$$2\ E_{x,y}\Big(\ E_{X^l,y^l}\big[(E_{x_l,y_l}(\mu(x)) - \mu(x)) * (E(y|x) - E_{X_l,y_l}(\mu(x)))\big]\ \Big) = 2\ E_{x,y}\Big((E(y|x) - E_{X_l,y_l}(\mu(x))) * 0\Big) = 0$$

И в конце концов получим:

$$L(\mu) = E_{x,y}(y - E[y|x])^2 + E_{x,y}\Big(\ E_{X^l,y^l}\big[E_{X_l,y_l}(\mu(x)) - \mu(x)\big]^2\ \Big) + E_{x,y}\Big(\ E_{X^l,y^l}\big[E(y|x) - E_{X_l,y_l}(\mu(x))\big]^2\ \Big)$$

Также видим, что в последнем слагаемом можно убрать мат. ожидание по X^l, y^l так как выражение под этим мат. ожиданием не зависит от X^l, y^l.
Тогда получим следующее:

$$L(\mu) = E_{x,y}(y - E[y|x])^2 + E_{x,y}\Big(\ E_{X^l,y^l}\big[E_{X_l,y_l}(\mu(x)) - \mu(x)\big]^2\ \Big) + E_{x,y}\Big(\ \big[E(y|x) - E_{X_l,y_l}(\mu(x))\big]^2\ \Big)$$

В итоге получили bias-variance разложение, в котором первое слагаемое - шум, второе - разброс и третье - смещение.

2 *Задача 2*

См. в прикрепленном питон-буке.

3 *Задача 3*

$$h(x) = \alpha * h_1(x) + (1-\alpha)h_2(x) = h_2(x) + \alpha(h_1(x) - h_2(x))$$

Пусть у нас есть набор документов $d_1, ..., d_N$.

Чтобы понять, при каком α достигается наилучшее NDCG нам нужно как-то рассмотреть различные перестановки документов, которые мы можем получить при разных α. Напомню решающее правило : $h(d_1) > h(d_2) \Rightarrow d_1$ стоит выше d_2.

Заметим, что $\forall d : h_1(d) = const$ и $h_2(d) = const. \Rightarrow h_1(d) - h_2(d) = const$. Кроме того $h_1(d) - h_2(d)$ может быть либо больше 0, либо меньше, либо равно. Тепер вспомним, что $h(x) = h_2(x) + \alpha(h_1(x) - h_2(x))$ и получим, что:
1) в случае, если $h_1(x) - h_2(x) > 0$, то при увеличении α $h(d)$ увеличивается.
2) если $h_1(x) - h_2(x) < 0$, то при увеличении α $h(d)$ уменьшается.
3) если $h_1(x) - h_2(x) = 0$, то $h(d)$ не меняется

Также заметим, что при если при каком-то α_1 было верно, что $h(d_1) > h(d_2)$, а при каком-то другом α_2 (без ограничения общности, пусть $\alpha_1 < \alpha_2$) верно $h(d_1) < h(d_2)$, то существует такой $\alpha \in (\alpha_1, \alpha_2) : h(d_1) = h(d_2)$ (это следует из линейности h). То есть мы получили, что если порядок ранжирования документов d_1 и d_2 меняется, то это происходит при каком-то α.

Попробуем найти такие α и соответствующие изменения в ранжировании при прохождении через этот α (то есть пару документов, ранжирование которой поменяется при прохождении через α).

Для этого заметим, что график $h(d) = h_2(d) + \alpha(h_1(d) - h_2(d)$ от α при фиксированном d - наклонная прямая (следует из того, что $h'_\alpha(d) = h_1(d) - h_2(d) = const h_2(d) = const$). Поэтому понимаем, что графики для двух разных документов d_1, d_2 либо не пересекаются, либо пересекаются в одной точке. В случае пересечения в точке α, мы как раз находим тот α: при котором происходит изменение в порядке ранжирования документов d_1 и d_2.

Для того, чтобы рассмотреть все такие α достаточно рассмотреть все пары документов d_1 и d_2 и найти такое $alpha$, что $h(d_1) = h(d_2)$. Решение такого уравнения потребует $O(1)$ времени (просто линейное уравнение на α). Всего таких пар документов $\frac{N(N-1)}{2}$, что $O(N^2)$.
В итоге после этого имеем массив из таких "переломных"α. Кроме того, для каждой такой α запомним номера двух документов, в которых происходит перестановка при переходе через α.

Отсуртируем полученный массив из α, что займет $O(N^2 * logN^2) = O(N^2 logN)$ так как данный массив имеет длину порядка N^2.

Теперь рассмотрим $\alpha = 0$. Посчитаем NDCG для данного α. Для этого нам по-

требуется отсортировать документы по значениям $h(d)$, что $O(NlogN)$. После считаем значение NDCG за линию.
Далее рассмотрим следующий по возрастанию "переломный"α (берем из отсортированного массива α) и считаем NDCG для него (точнее не для него, а для промежутка, который лежит после этого α и до следующего в массиве переломного $alpha$). Для того, чтобы посчитать NDCG при данном α нам не нужно сортировать. Для того, чтобы получить отсортированный массив документов достаточно лишь свопнуть 2 документа, которые соответствуют данному α (находили ранее) и посчитать значение NDCG за линию.
И так далее для всех "переломных"α.
По пути выбираем $alpha$ с самым оптимальным NDCG.

В итоге при таком прохождении рассмотрим $O(N^2)$ разных "переломных"α, произведем сортировку при $\alpha = 0$ и для остальных α найдем NDCG за $O(N)$.

Итоговая сложность $O(N^2) + O(N^2 logN) + O(NlogN) + O(N) * O(N^2) = O(N^3)$.

4 *Задача 4*

$$Q = argmin_{Q \in \mathbb{R}^{r*m}} \sum_{(u,i) \in known} (r_{ui} - p_u^T q_i)^2$$

Заметим, что мы можем оптимизировать независимо по разным i. То есть можем оптимизировать отдельно по столбцам матрицы Q. Получим :

$$Q_i = argmin_{q_i \in \mathbb{R}^r} \sum_{u:(u,i) \in known} (r_{ui} - p_u^T q_i)^2$$

Для минимизации данного выражения возьмем производную по q_i и приравняем к 0.

$$\sum_{u:(u,i) \in known} (p_u^T q_i)^2 = -2 \sum_{u:(u,i) \in known} p_u (r_{ui} - p_u^T q_i)) = 0$$

Отсюда получим выражение для оптимального Q_i :

$$Q_i = \left(\sum_{u:(u,i) \in known} p_u p_u^T \right)^{-1} \left(\sum_{u:(u,i) \in known} p_u r_{ui} \right)$$

То есть получили выражение и для всей матрицы Q.

Теперь поймем асимптотику.
Рассмотрим сначала первую скобку.
Для подсчета $p_u p_u^T$ требуется $O(r^2)$ операций так как $p_u p_u^T$ - матрица $r*r$. Такие $p_u p_u^T$ нам нужно просумировать по $u : (u, i) \in known$, то есть в силу равномерного распределения известных оценок по матрице получаем $O(\alpha n)$ слагаемых. Для того, чтобы обратить

матрицу $r * r$ потребуется $O(r^3)$. Таким образом для конкретного i первая скобка считается за $O(r^2) * O(\alpha n) + O(r^3) = O(\alpha n r^2 + r^3)$

Теперь рассмотрим вторую скобку.
Для подсчета $p_u r_{ui}$ требуется $O(r)$ времени. Таких слагаемых у нас опять же $O(\alpha n)$, как и в первой скобке. В итоге получаем, что на вычисление второй скобки нужно $O(r) * O(\alpha n) = O(\alpha n r)$.

Теперь нужно перемножить 2 скобки - $O(r^2)$, так как первая скобка - матрица размера $r * r$, а вторая скобка - вектор размера r.
В итоге на вычисление Q_i потребуется $O(\alpha n r + \alpha n r^2 + r^3 + r^2) = O(\alpha n r^2 + r^3)$

Таким Q_i у нас m штук, поэтому итоговая асимптотика : $O(m\alpha n r^2 + m r^3) = O(r^2 m(\alpha n + r))$.

5 *Задача 5*

$$a(x) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=i+1}^{d} x_i x_j < v_i, v_j >$$

Заметим, что первая сумма считается за $O(rd)$ и поэтому мы можем ее не трогать. Прибавление w_0 - $O(1)$.
Будем оптимизировать последнюю сумму. Заметим, что

$$\sum_{i=1}^{d} \sum_{j=i+1}^{d} x_i x_j < v_i, v_j >= \sum_{i=1}^{d} \sum_{j>i} x_i x_j < v_i, v_j >= \sum_{i=1}^{d} \sum_{j<i} x_i x_j < v_i, v_j >$$

Отсюда получим, что

$$\sum_{i=1}^{d} \sum_{j=i+1}^{d} x_i x_j < v_i, v_j >= \frac{\sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j < v_i, v_j > - \sum_{i=1}^{d} x_i x_i < v_i, v_i >}{2}$$

Заметим, что вторая сумма считается за $O(d)$ и поэтому можем ее не оптимизировать. Рассмотрим первую сумму.

$$\sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j < v_i, v_j >= \sum_{i=1}^{d} \sum_{j=1}^{d} < x_i v_i, x_j v_j >$$

Давайте для начала заранее посчитаем $u_i = x_i v_i \forall i \in [0, 1, .., d]$, что займет $O(dr)$, так как подсчет одного u_i требует $O(r)$. Тогда для того, чтобы посчитать сумму нужно посчитать $\sum_{i=1}^{d} \sum_{j=1}^{d} < u_i, u_j >$.

Разложим эту сумму :

$$\sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{r} u_{ik} * u_{jk} = \sum_{k=1}^{r} \sum_{i=1}^{d} \sum_{j=1}^{d} u_{ik} * u_{jk} = \sum_{k=1}^{r} \left(\sum_{i=1}^{d} u_{ik} \right)^2$$

Подсчет данной суммы займет как раз $O(rd)$. В итоге получим, что пожем все посчитать за $O(rd)$.

6 *Задача 6*

У Дениса странная задача : он хочет предсказывать покупки пользователей. Это неплохая задача, но это совсем не рекомендательная система : алгоритм просто понимает, что человек скорей всего и так купит и нам нет смысла ему это рекомендовать эти товары. Хочется порекомендовать то, что будет полезно, но то, что он бы сам не купил. С другой стороны, данный алгоритм может выучить такие зависимости как : с этим товаром часто покупают другой товар; люди, интересующиеся тем-то часто покупают что-то. Поэтому в случае, если человек не купит то, что было предложено алгоритмом, то порекомендовав ему то, что было предложено алгоритмом, он скорей всего это в итоге купит.

У Андрея идея кажется более логичной так как наша метрика - удовлетворенность пользователями рекомендациями, что в принципе должно быть видно при A-B тестировании. Однако не совсем корректо выбирать правильное решение только лишь по критерию "в какой группе больше". Для этого нужно провести какой-либо статистический тест.

В остальном идея Андрея кажется хорошей.