Московский государственный университет имени М. В. Ломоносова

Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Методы повышения эффективности обучения, основанные на ансамблях промежуточных решений»

Выполнил:

студент 3 курса 317 группы

Королев Николай Сергеевич

Научный руководитель:

д.ф-м.н., в. науч. сотр. ВЦ РАН

Сенько Олег Валентинович

Москва, 2019

Содержание

Аннотация

Данный документ является образцом оформления дипломной работы для студентов кафедры Математических методов прогнозирования ВМК МГУ. Приведённые ниже рекомендации взяты из статьи «Написание отчётов и статей (рекомендации)» на вики-ресурсе `www.MachineLearning.ru`. Студенты, готовящие дипломную работу к защите, могут найти много полезной информации также в статьях «Научно-исследовательская работа (рекомендации)», «Подготовка презентаций (рекомендации)», «Защита выпускной квалификационной работы (рекомендации)» на том же ресурсе.

Аннотация обычно содержит краткое описание постановки задачи и полученных результатов, одним абзацем на 10–15 строк. Цель аннотации — обозначить в общих чертах, о чём работа, чтобы человек, совершенно не знакомый с данной работой, понял, интересна ли ему эта тема, и стоит ли читать дальше. Аннотация собирается в последнюю очередь путем легкой модификации наиболее важных и удачных фраз из введения и заключения.

1 Введение

Во введении рассказывается, где возникает данная задача, и почему её решение так важно. Вводится на неформальном уровне минимум терминов, необходимый для понимания постановки задачи. Приводится краткий анализ источников информации (литературный обзор): как эту задачу решали до сих пор, в чем недостаток этих решений, и что нового предлагает автор. Формулируются цели исследования. В конце введения даётся краткое содержание работы по разделам; при этом отмечается, какие подходы, методы, алгоритмы предлагаются автором впервые. При упоминании ключевых разделов кратко формулируются основные результаты и наиболее важные выводы.

Цель введения: дать достаточно полное представление о выполненном исследовании и полученных результатах, понятное широкому кругу специалистов. Большинство читателей прочтут именно введение и, быть может, заключение. Во введении автор решает сложную оптимизационную проблему: как сообщить только самое важное, потратив минимум времени читателя, да так, чтобы максимум читателей поняли, о чём вообще идёт речь.

Введение лучше писать напоследок, так как в ходе работы обычно происходит переосмысление постановки задачи. Если же введение писать, когда работа еще не готова, задача усложняется вдвойне. В конце обычно приходит понимание, что всё получилось совсем не так, как планировалось в начале, и исходный вариант введения всё равно придётся переписывать. Кстати, к таким «потерям» надо относиться спокойно — в хорошей работе почти каждый абзац многократно переделывается до неузнаваемости.

Введение имеет много общего с текстом доклада на защите, поэтому имеет смысл готовить их одновременно.

1.1 Постановка задачи

$x_1, x_2, \ldots, x_N$ – точки в некотором векторном пространстве; $y_1, y_2, \ldots, y_N$ – значения, соответствующие этим точкам. При этом $y_i = y(x_i) = f(x_i) + \varepsilon$, где ε – случайная величина с нулевым математическим ожиданием и дисперсией σ^2. Пусть также задана функция $\hat{f}_1(x)$, которая приближает функцию $f(x)$. Стоит задача нахождения $m-1$ функции $\hat{f}_2(x), \hat{f}_3(x), \ldots, \hat{f}_m(x)$, а также функции-ансамбля $\hat{f}(x) = a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_m(x))$, приближающей функцию $f(x)$ лучше чем любая из функций $\hat{f}_i(x),\ i = 1, 2, \ldots, m$.

1.2 Теоретическая часть

Из [3] известно разложение математического ожидания квадратичной ошибки на смещение и дисперсию:

$$\mathbb{E}\left[\left(\hat{f}(x) - y\right)^2\right] = \left(\mathbb{E}\,\hat{f}(x) - f(x)\right)^2 + \left(\mathbb{E}\,\hat{f}^2(x) - \left(\mathbb{E}\,\hat{f}(x)\right)^2\right) + \sigma^2$$

Для уменьшения как смещения, так и дисперсии $\hat{f}(x)$ используются ансамбли:

$$\hat{f}(x) = a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_m(x))$$

В [1] было доказано, что в случае, когда функция $a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_m(x))$ является выпуклой комбинацией своих аргументов:

$$a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_m(x)) = \sum_{i=1}^{m} c_i \hat{f}_i(x),$$

$$\sum_{i=1}^{m} c_i = 1; \quad c_i \geq 0, \quad i = 1, 2, \ldots, m;$$

выполнено

$$\mathbb{E}\left[\left(\hat{f}(x) - y\right)^2\right] = \sum_{i=1}^{m} c_i\,\mathbb{E}\left[\left(\hat{f}_i(x) - y\right)^2\right] - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j\,\mathbb{E}\left[\left(\hat{f}_i(x) - \hat{f}_j(x)\right)^2\right] \quad (1)$$

Соответственно, для уменьшения среднеквадратичной ошибки необходимо уменьшать среднеквадратичную ошибку каждого предиктора $\hat{f}_i(x)$, а также увеличивать расхождение между прогнозами различных предикторов.

1.3 Существующие методы

В большинстве случаев функция-ансамбль $\hat{f}(x) = a(\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_m(x))$ является линейной комбинацией функций предикторов $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_m(x)$. Коэффициенты ищутся при помощи методов регрессионного анализа:

- гребневая регрессия [4]
- метод Лассо [5]
- эластичная сеть [6]
- регрессионная модель, отбирающая признаки, наиболее коррелирующие с откликом [2]

Функции $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_m(x)$ в таких случаях обычно независимы друг от друга и получены методом оптимизации некоторого функционала.

2 Ансамбль промежуточных решений

Пусть:
X - матрица из N строк, i-ая строка равна x_i;
Y - вектор из N элементов, i-ый элемент равен y_i.

Пусть $\hat{f}_1(x)$ представима в виде функции с параметрами $\tilde{f}(x, \theta)$ и была получена методом оптимизации функционала среднеквадратичной ошибки $MSE(\hat{f}_1(x), X, Y) = \dfrac{1}{N}\sum\limits_{i=1}^{N}(\hat{f}_1(x_i) - y_i)^2$:

$$\hat{f}_1(x) = \tilde{f}(x, \theta_1)$$

2.1 Вычисление промежуточных решений

Будем искать функции $\hat{f}_2(x), \hat{f}_3(x), \ldots, \hat{f}_m(x)$ в виде $\tilde{f}(x, \theta)$, минимизируя следующий функционал:

$$\mathcal{L}(\hat{f}_k(x), X, Y) = MSE(\tilde{f}(x, \theta_k), X, Y) - \frac{\alpha}{k-1}\sum_{i=1}^{k-1}\|\theta_k - \theta_i\|^2, \quad k = 2, 3, \ldots, m,$$

где $\alpha \geq 0$ является гиперпараметром.

Данный функционал поощряет функции $\hat{f}_i(x)$ иметь различные параметры θ_i. Таким образом мы добиваемся максимального расхождения значений функций. Впоследствии это уменьшит среднеквадратичную ошибку функции-ансамбля $\hat{f}(x)$, что следует из уравнения (1).

Стоит отметить, что при больших k вычисление данного функционала может быть вычислительно затратным, так как $\sum\limits_{i=1}^{k-1} \|\theta_k - \theta_i\|^2$ в общем случае требует $O(k)$ времени. Если же все θ_i лежат в некотором евклидовом, данное выражение переписывается в виде:

$$\sum_{i=1}^{k-1} \|\theta_k - \theta_i\|^2 = (k-1) \|\theta_k\|^2 - 2\langle\theta_k, \sum_{i=1}^{k-1} \theta_i\rangle + \sum_{i=1}^{k-1} \|\theta_i\|^2 .$$

Значения выражений $\sum\limits_{i=1}^{k-1} \theta_i$ и $\sum\limits_{i=1}^{k-1} \|\theta_i\|^2$ можно поддерживать в течении всего процесса поиска функции $\hat{f}_k(x)$ и пересчитывать за $O(1)$ времени при переходе к поиску $\hat{f}_{k+1}(x)$.

3 Вычислительные эксперименты

Цель данного раздела: продемонстрировать, что предложенная теория работает на практике; показать границы её применимости; рассказать о новых экспериментальных фактах.

Чисто теоретические работы могут вообще не содержать раздела экспериментов (не работает, ну и не надо — зато теория красивая). Кстати, теоретики имеют право не догадываться, где, кому и когда их теории пригодятся.

3.1 Исходные данные и условия эксперимента

Описывается прикладная задача, параметры анализируемых данных (например, сколько объектов, сколько признаков, каких они типов), параметры эксперимента (например, как производился скользящий контроль).

3.2 Результаты эксперимента

Результаты экспериментов представляются в виде таблиц и графиков. Объясняется точный смысл всех обозначений на графиках, строк и столбцов в таблицах.

3.3 Обсуждение и выводы

Приводятся выводы: в какой степени результаты экспериментов согласуются с теорией? Достигнут ли желаемый результат? Обнаружены ли какие-либо факты, не нашедшие объяснения, и которые нельзя списать на «грязный» эксперимент?

Обсуждаются основные отличия предложенных методов от известных ранее. В чем их преимущества? Каковы границы их применимости? Какие проблемы удалось решить, а какие остались открытыми? Какие возникли новые постановки задач?

4 Заключение

В квалификационных работах последний раздел нужен для того, чтобы конспективно перечислить основные результаты, полученные лично автором.

Результатами, в частности, являются:

- Предложен новый подход к…
- Разработан новый метод…, позволяющий…
- Доказан ряд теорем, подтверждающих (опровергающих), что…
- Проведены вычислительные эксперименты…, которые подтвердили / опровергли / привели к новым постановкам задач.

Цель данного раздела: доказать квалификацию автора. Даже беглого взгляда на заключение должно быть достаточно, чтобы стало ясно: автору удалось решить актуальную, трудную, ранее не решённую задачу, предложенные автором решения обоснованы и проверены.

Иногда в Заключении приводится список направлений дальнейших исследований.

Список литературы

[1] *Докукин А. А., Сенько О. В.* Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности // *Ж. вычисл. матем. и матем. физ.* — 2011. — Т. 51. — С. 1751–1760.

[2] *Докукин А. А., Сенько О. В.* Регрессионная модель, основанная на выпуклых комбинациях, максимально коррелирующих с откликом // *Ж. вычисл. матем. и матем. физ.* — 2015. — Т. 55. — С. 530–544.

[3] *Domingos Pedro.* A unified bias-variance decomposition for zero-one and squared loss // Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. — AAAI Press, 2000. — Pp. 564–569.

[4] *Ng Andrew Y.* Feature selection, l1 vs. l2 regularization, and rotational invariance // Proceedings of the Twenty-first International Conference on Machine Learning. — ICML '04. — New York, NY, USA: ACM, 2004. — Pp. 78–.

[5] *Tibshirani Robert.* Regression shrinkage and selection via the lasso // *Journal of the Royal Statistical Society. Series B (Methodological).* — 1996. — Vol. 58, no. 1. — Pp. 267–288.

[6] *Zou Hui, Hastie Trevor.* Regularization and variable selection via the elastic net // *Journal of the Royal Statistical Society, Series B.* — 2005. — Vol. 67. — Pp. 301–320.