Машинное обучение: задание 3

Составитель: Виктор Кантор

12 марта 2017 г.

Организационные вопросы

Дедлайн

На выполнение задания дается две недели, последний срок сдачи — 27 марта 23:59.

Сдача задания

1. Пришлите на почту ml.course.mipt@gmail.com письмо с решениями заданий в ipynb//py//pdf файлах (где что более уместно). Тему письма укажите в формате «ML2017_fall <номер_группы> <фамилия> <имя>, Задание 3 (Trees ensembles)», например: «ML2017_fall 492 Страуструп Бьярн, Задание 3 (Trees ensembles)»
2. Если есть принципиальное желание оформлять теоретические задачи в ipynb в Markdown-ячейках, это не запрещается. Также не запрещается решать теоретические задачи на бумаге, оформлять их аккуратно и разборчиво, и присылать pdf со сканами, вместо того, чтобы набирать в LaTeX.

Контрольные вопросы

Ниже приводится список вопросов, с ответами на которые может быть полезно разобраться для понимания темы.

Деревья

1. Как выглядит решающее дерево? Как применяется уже построенное для задачи классификации дерево? А для задачи регрессии?
2. Как строятся решающие деревья? (рекомендуется обратиться к материалам лекций или документации sklearn)
3. Как выглядят энтропийный критерий, критерий Джини и среднеквадратичное отклонение, используемое как критерий в задаче регрессии?
4. Что такое node impurity и goodness of split? Как они связаны?
5. Какие преимущества и недостатки есть у деревьев? (полезно как подумать самостоятельно, так и обратиться к документации sklearn)
6. Есть ли разница (с точки зрения вида получаемого в итоге дерева): строить каждое разбиение в дереве, максимизируя информативность, или строить каждое разбиение, минимизируя «ошибку», как было предложено на первой лекции про деревья?

Общие идеи построения композиций

1. Что такое bagging, blending, stacking, boosting?
2. Нужно ли как-то делить выборку, чтобы избежать переобучения, при реализации стэкинга?
3. В чем преимущества и недостатки бустинга и бэггинга?

Градиентный бустинг

1. В чем основная идея градиентного бустинга?
2. Как выглядит алгоритм градиентного бустинга в самом общем виде — с произвольной функцией потерь в функционале ошибки и произвольным функционалом, оценивающим качество приближения антиградиента?
3. Как выглядит алгоритм градиентного бустинга с квадратичными функциями потерь? На что настраиваются базовые алгоритмы?
4. Как выглядит алгоритм градиентного бустинга в случае задачи бинарной классификации?
5. Какие параметры есть у классификаторов и регрессоров на основе градиентного бустинга над деревьями в sklearn и XGBoost? Какие параметры стоит настраивать в первую очередь?
6. Какая высота деревьев оправдана в градиентном бустинге над деревьями? Почему?
7. Есть ли у градиентного бустинга склонность к сильному переобучению при увеличении количества деревьев? С какими еще параметрами алгоритма эффект переобучения может быть связан?
8. Какие есть методы борьбы с переобучением, применяемые в градиентном бустинге?

Случайный лес

1. Как работает Random Forest?
2. Зачем в Random Forest делается рандомизация с выбором подмножества признаков в каждом сплите?
3. Какой высоты деревья стоит строить в Random Forest?
4. Какие параметры есть у Random Forest в sklearn? Какие параметры стоит настраивать в первую очередь?
5. Есть ли у Random Forest склонность к сильному переобучению при увеличении количества деревьев? С какими еще параметрами алгоритма эффект переобучения может быть связан?
6. Как работает ExtraTreesClassifier из sklearn?

1 Контест на прогнозирование спроса

60% баллов за задание

В этом контесте вам нужно решить задачу прогнозирования количества проданных товаров, при этом количество получаемых баллов будет определять качество на private leaderboard (для этого потребуется не только построить хорошую модель, но и не переобучиться на public leaderboard).

Подробное описание - на странице соревнования, присоединиться можно по ссылке:
https://kaggle.com/join/demand_prediction_mipt

В названии команды обязательно указывайте номер своей группы, имя и фамилию.

За соревнование вы получаете 60% от всех баллов за это задание, они распределяются следующим образом:

1. от 5% до 45% баллов за преодоление порогов по SMAPE на private leaderboard:
   (a) 5% - при $40\% \leq \text{SMAPE} \leq 45\%$
   (b) 10% - при $30\% \leq \text{SMAPE} < 40\%$
   (c) 15% - при $24\% \leq \text{SMAPE} < 30\%$

(d) 45% - при SMAPE $< 24\%$

2. до 15% баллов за отчет в свободной форме о решении контеста, включающий:
   (a) Краткую формулировку задачи
   (b) Описание итогового решения: как готовились данные, что использовалось в качестве таргета, какой алгоритм обучался и все комментарии, которые могут быть полезны для воспроизведения вашего решения
   (c) Рассказ о подходах, которые вы пробовали, и том, как реализация тех или иных идей сказывалась на качестве вашего решения
   (d) Код вашего итогового решения
   (e) Описание того, как вы оценивали качество при решении контеста: как делали кросс-валидацию, насколько она коррелировала с результатом по leaderboard

Также за соревнование можно получить дополнительные баллы (до 50% от основного балла за всё третье задание вместе с теоразадчами). Количество дополнительных баллов рассчитывается на основе места в private leaderboard по следующей формуле:

$$40\% \cdot 0.5^{k-1} + 10\% \cdot (N - k)$$

N - количество участников конкурса, получивших SMAPE на private leaderboard меньше 24%, k - место в private leaderboard. В случае, если несколько участников делят места с m-того по $(m+n)$-ное, для рассчетов используется $k = m + n$, так что участники не заинтересованы в идеально совпадающих решениях друг у друга.

2 Теоретические задачи

2.1 Bias-variance-noise decomposition

15% баллов

Покажите справедливость bias-variance-noise decomposition:

$$\mathbf{E}_{x,y}\mathbf{E}_{X^\ell}\left(y - a_{X^\ell}(x)\right)^2 = \mathbf{E}_{x,y}\left(y - \mathbf{E}(\mathbf{y}|\mathbf{x})\right)^2 +$$

$$+\mathbf{E}_{x,y}\left(\mathbf{E}(\mathbf{y}|\mathbf{x}) - \mathbf{E}_{X^\ell}a_{X^\ell}(x)\right)^2 + \mathbf{E}_{x,y}\mathbf{E}_{X^\ell}\left(a_{X^\ell}(x) - \mathbf{E}_{X^\ell}a_{X^\ell}(x)\right)^2$$

Слагаемые в этом разложении называются соответственно шумом (noise), смещением (bias) и разбросом (variance).

2.2 Смещение и разброс в бэггинге

15% баллов

Выясните как соотносятся смещение и разброс для композиции с теми же параметрами для базовых алгоритмов, если композиция строится с помощью бэггинга:

$$a(x) = \frac{1}{M}\sum_{m=1}^{M} a_m(x)$$

Можете считать, что ответы всех базовых алгоритмов распределены одинаково.

2.3 Корреляция ответов базовых алгоритмов

10% баллов

Покажите, что если есть M одинаково распределенных случайных величин с дисперсией σ^2, любые две из которых имеют положительную корреляцию ρ, то дисперсия их среднего будет равна:

$$\rho\sigma^2 + (1-\rho)\frac{\sigma^2}{M}$$

3 Дополнительные вопросы

до 15% дополнительных баллов

1. Как вы думаете, как был сформирован sample_submission.tsv для контеста?
2. Выясните и опишите, как строятся решающие деревья в sklearn: как выбираются разбиения в вершинах, как выполняется перебор разбиений, делается ли прунинг, на какой из известных алгоритмов это похоже и в чем отличия?
3. Попробуйте с помощью BaggingClassifier сделать аналог Random Forest, но с выбором случайного подмножества признаков не в каждом сплите, а перед построением всего дерева. Сравните результаты с обычным Random Forest.