Школа анализа данных
Машинное обучение, часть 1
Домашнее задание №3

Федор Ерин

Задача 1 (1.5 балла). Метрики качества.

Костя участвует в конкурсе по анализу данных, в котором нужно решить задачу бинарной классификации, в которой для оценивания качества используется функционал ошибки Mean Absolute Error (MAE):

$$Q(\vec{y}, \tilde{\vec{y}}) = \frac{1}{l}\sum_{i=1}^{l}|y_i - \tilde{y}_i|, \quad \vec{y} \in \{0,1\}^l, \quad \tilde{\vec{y}} \in [0,1]^l,$$

где l — количество обучающих объектов, $\vec{y}$ — вектор истинных классов объектов, $\tilde{\vec{y}}$ — вектор предсказанных «степеней принадлежности» классу 1, качество которого и оценивается. Костя заметил, что качество предсказания на скрытой выборке, которая доступна только организаторам конкурса, часто улучшается, если сдвинуть прогноз $\tilde{y}_i$ в один из концов отрезка $[0,1]$ для каждого объекта i. Объясните, почему Костины действия привели к улучшению качества предсказания. При каких обстоятельствах Костя мог провернуть такой трюк? Всегда ли такое возможно?

Подсказка: Сделайте предположение, что объект с номером i принадлежит классу 1 с истинной вероятностью p_i.

* Оцениваться будут рассуждения, каким условиям должны удовлетворять предсказания, чтобы Костина идея сработала, а также подтверждающие и/или опровергающие примеры.

- ***Модель прогнозирует "степень принадлежности" к классу 1, но это может быть не настоящей вероятностью, особенно в задаче с сильным дизбалансом классов. В таком случае используется калибровка, которая на отложенной выборке (которую не видела модель) ищет преобразование значений выхода модели с некоторым распределением вероятностей, например, схожим с нормальным или равномерным, в то, которое истинно. В случае редкого класса 1, это будет распределение сильно смещенное к нулю. Если положить, что истинная вероятность принадлежности некоторого объекта с номером i равна p_i, а исходная некалиброванная модель выдала $\tilde{p}_i$ (причем если значения близки к 1, то $\tilde{p}_i < p_i$, или наоборот), то сместив прогноз в ближайший край отрезка $[0,1]$ можно улучшить метрику MAE.***

Более того, в случае задачи классификации оптимально с т.з. MAE будет округлять прогноз до ближайшего целого 0/1, так как мы можем предсказывать достаточно точно выдавать для истинного 1 высокие вероятности, а для 0 низкие, но MAE будет накапливать отклонения. В данной задаче такой эффект улучшения метрики может достигаться при некалиброванной неуверенной модели и дисбалансе классов, для уверенной модели такой эффект будет минимален. Такое возможно не всегда, например, когда баланс идеален и тогда такие смещения будут делать модель излишне уверенной (что будет видно на калибровочной кривой) и тогда распределение вероятностей в прогнозах модели исказит реальную картину в скрытой выборке.

Задача 2 (1.5 балла). Метрические методы, kNN, проклятие размерности..

Рассмотрим N точек, распределенных равномерно по объему D-мерного единичного шара с центром в нуле. Предположим, что мы хотим применить метод ближайшего соседа для точки начала координат. Зададимся вопросом, на каком расстоянии будет расположен ближайший объект. Для ответа на этот вопрос выведите выражение для **медианы** расстояния от начала координат до ближайшего объекта. Чтобы проинтерпретировать полученный результат, подставьте в формулу конкретные значения: $N = 500$ и $D = 10$. Покажите, как будет меняться значение медианы при дальнейшем увеличении размерности пространства при фиксированном количестве точек и постройте график этой зависимости. Поясните, в чем состоит проклятие размерности и почему полученная для медианы формула наглядно его демонстрирует. Для размерности D посчитайте, сколько точек $N = f(D)$ необходимо взять, чтобы побороть проклятье размерности.

- ***TODO***

Задача 3 (1.5 балла). Решающие деревья, индекс Джини.

Пусть имеется построенное решающее дерево для задачи многоклассовой классификации. Рассмотрим лист дерева с номером m и объекты R_m, попавшие в него. Обозначим за p_{mk} долю объектов k-го класса в листе m. *Индексом Джини* этого листа называется величина

$$\sum_{k=1}^{K} p_{mk}(1 - p_{mk}),$$

где K — общее количество классов. Индекс Джини обычно служит мерой того, насколько хорошо в данном листе выделен какой-то один класс

1. Поставим в соответствие листу m классификатор $a(x)$, который предсказывает класс случайно, причем класс k выбирается с вероятностью p_{mk}. Покажите, что матожидание частоты ошибок этого алгоритма на объектах из R_m равно индексу Джини.

- ***Распишем матожидание частоты ошибок такого алгоритма:***

$$\mathbb{E}\left(\frac{\sum_{x_i \in R_m}[y_i \neq a(x_i)]}{n}\right) = \frac{1}{n}\sum_{x_i \in R_m}\mathbb{E}[y_i \neq a(x_i)] =$$

$$= \frac{1}{n}\sum_{x_i \in R_m}(1 - p_{my_i}) = \sum_{k=1}^{K}\frac{\sum_{x_i \in R_m}[y_i = k]}{n}(1 - p_{mk}) =$$

$$= \sum_{k=1}^{K} p_{mk}(1 - p_{mk}).$$

2. *Дисперсией класса* k назовем дисперсию выборки $\{[y_i = k] : x_i \in R_m\}$, где y_i — класс объекта x_i, $[f]$ — индикатор истинности выражения f, равный 1 если f верно, и нулю в противном случае, а R_m — множество объектов в листе. Покажите, что сумма дисперсий всех классов в заданном листе равна его индексу Джини.

   - ***Распишем сумму дисперсий всех классов в заданном листе:***

$$\mathbb{D} = \sum_{k=1}^{K}\frac{\sum_{i=0}^{n}([y_i = k] - p_{mk})^2}{n} =$$

$$= \sum_{k=1}^{K}\frac{(1 - p_{mk})^2 \cdot p_{mk} \cdot n + p_{mk}^2 \cdot (1 - p_{mk}) \cdot n}{n} =$$

$$= \sum_{k=1}^{K}(1 - p_{mk}) \cdot p_{mk} \cdot (1 - p_{mk} + p_{mk}) = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}).$$

Задача 4 (1.5 балла). LDA

Для задачи бинарной классификации с классами 0 и 1 рассмотрим две модели: 1) Логистическую регрессию (без регуляризации); 2) LDA Докажите, что в модели LDA

$$\log\left[\frac{P(y=0|x)}{P(y=1|x)}\right] = a_0 + a^T x,$$

где a_0 и a — это число и вектор, выражающиеся через матожидание и матрицу ковариации распределений $p(x|y)$. Собственно, как и в логистической регрессии. Так правда ли, что LDA для двух классов и логистическая регрессия — это разные модели? Объясните, почему. Сравните эти две модели.

- ***TODO***

Задача 5 (1.5 балла). Градиентный бустинг

1. Какой функции потерь будет соответствовать градиентный бустинг, который на каждой итерации настраивается на разность между вектором истинных меток и текущим вектором предсказанных меток?

- ***Градиентный бустинг при использовании функции потерь*** $MSE(y, \tilde{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2$ ***настраивается на градиент функции потерь по ответу модели, чтобы сделать шаг в сторону этого вектора со знаком минус (антиградиент). Производная MSE с точностью до константы равна разности истинных меток и прогнозных, поэтому в данном случае речь идет о функции потерь MSE.***

2. Градиентный бустинг обучается на пяти объектах с функцией потерь для одного объекта

$$\mathcal{L}(y, y) = (\tilde{y} - y)^4$$

На некоторой итерации полученная композиция дает ответ $(5, 10, 6, 3, 0)$. На какой вектор ответов будет настраиваться следующий базовый алгоритм, если истинный вектор ответов равен $(6, 8, 6, 4, 1)$?

- $\mathcal{L}'_0(y, y) = 4(\tilde{y_0} - y_0)^3$***, где*** $\tilde{y_0} = (5, 10, 6, 3, 0), y_0 = (6, 8, 6, 4, 1)$ ***- бустинг будет настраивать следующий алгоритм на данный вектор со знаком минус, вычислим его:***

$$-\mathcal{L}'_0 = -4((5, 10, 6, 3, 0) - (6, 8, 6, 4, 1))^3 = -4((-1, 8, 0, -3, 1))^3 =$$

$$= (4, -32, 0, 4, 4)$$

3. Рассмотрим задачу бинарной классификации, $Y = \{0, 1\}$. Будем считать, что все алгоритмы из базового семейства A возвращают ответы из отрезка $[0, 1]$, которые можно интерпретировать как вероятности принадлежности объекта к классу 1. В качестве функции потерь возьмем отрицательный логарифм правдоподобия (negative log-likelihood):

$$\mathcal{L}(y, z) = -(y \log z + (1 - y)\log(1 - z)),$$

где y — правильный ответ, а z — ответ алгоритма. На какой таргет и с какой функцией потерь настраиваются базовые алгоритмы градиентного бустинга.

- ***Таргетом обучения базовых алгоритмов градиентного бустинга является антиградиент функции потерь:***

  $$-\mathcal{L}'(y, z) = \frac{y}{z} - \frac{1 - y}{1 - z}$$

  Функцией потерь базовых алгоритмов может быть MSE или косинусное расстояние, оптимизируем эту функцию независимо от функционала исходной задачи, так как вся информация о функции потерь $\mathcal{L}$ ***содержится в векторе сдвигов*** $s = (s_1, ..., s_n)$***.***

Задача 6 (1 балл) ROC AUC.

Определим понятие доли дефектных пар ответов классификатора. Пусть дан классификатор $a(x)$, который возвращает оценки принадлежности объектов классам: чем больше ответ классификатора, тем более он уверен в том, что данный объект относится к классу «+1». Отсортируем все объекты по неубыванию ответа классификатора a: $x_{(1)}, \dots, x_{(\ell)}$. Обозначим истинные ответы на этих объектах через $y_{(1)}, \dots, y_{(\ell)}$. Тогда доля дефектных пар записывается как

$$\mathrm{DP}(a, X^\ell) = \frac{2}{\ell(\ell-1)} \sum_{i<j}^{\ell} \left[y_{(i)} > y_{(j)}\right].$$

Как данный функционал связан с AUC (площадью под ROC-кривой)? Ответ должен быть дан в виде формулы, связывающей DP и AUC.

- ***Обозначим ℓ^+и ℓ^- кол-во объектов положительного и отрицательного классов соответственно ($\ell^+ + \ell^- = \ell$). При построении ROC-кривой по мере невозрастания оценок классификатора (от большей к меньшей) мы делаем шаг вверх на $1/\ell^+$, если у объекта истинная метка «1», иначе - на $1/\ell^-$ вправо, в этом случае AUC увеличивается на величину:***

$$\sum_{j=i+1}^{\ell} \frac{\left[y_{(j)} = 1\right]}{\ell^+\ell^-}.$$

Тогда распишем AUC:

$$\mathrm{AUC} =$$

$$= \sum_{j=i+1}^{\ell} \frac{\left[y_{(j)} = 1\right]}{\ell^+\ell^-} \sum_{i=1}^{\ell} \left[y_{(i)} = 0\right] =$$

$$= \frac{1}{\ell^+\ell^-} \sum_{i=1}^{\ell} \sum_{j=i+1}^{\ell} \left[y_{(i)} < y_{(j)}\right] =$$

$$= \frac{1}{\ell^+\ell^-} \sum_{i<j}^{\ell} \left(1 - \left[y_{(i)} \geqslant y_{(j)}\right]\right) =$$

$$= \frac{1}{\ell^+\ell^-} \sum_{i<j}^{\ell} \left(1 - \left[y_{(i)} = y_{(j)}\right]\right) - \frac{1}{\ell^+\ell^-} \sum_{i<j}^{\ell} \left[y_{(i)} > y_{(j)}\right] =$$

$$= \frac{1}{\ell^+\ell^-} \frac{\ell(\ell-1)}{2} - \frac{\ell^+(\ell^+-1)}{2\ell^+\ell^-} - \frac{\ell^-(\ell^--1)}{2\ell^+\ell^-} - \frac{1}{\ell^+\ell^-} \sum_{i<j}^{\ell} \left[y_{(i)} > y_{(j)}\right] =$$

$$= 1 - \frac{1}{\ell^+\ell^-} \sum_{i<j}^{\ell} \left[y_{(i)} > y_{(j)}\right] =$$

$$= 1 - \frac{\ell(\ell-1)}{2\ell^+\ell^-} \mathrm{DP}$$

Следовательно, DP связан с AUC следующим образом:

$$\mathrm{DP} = \frac{2\ell^+\ell^-}{\ell(\ell-1)} \left(1 - AUC\right).$$

Задача 7 (1.5 балл) SVD. Для двух заданных матриц A и B одного размера найдите ортогональную матрицу Q, для которой норма Фробениуса разности $||QA - B||_F$ минимальна.

Напомним, что норма Фробениуса определяется, как

$$X_F = \sqrt{\sum_{i,j} x_{ij}^2}$$

Эту задачу можно решать по-разному, но наиболее эффективное решение использует сингулярное разложение (а какой именно матрицы — вам предстоит выяснить самим))

- ***Для удобства рассмотрим квардрат нормы Фробениуса и распишем ее через след:***

$$||QA-B||_F^2 = \mathrm{tr}\left[(QA-B)(QA-B)^T\right] = \mathrm{tr}\left[(QA-B)(A^TQ^T-B^T)\right] =$$

$$= \mathrm{tr}\left[QAA^TQ^T\right] - \mathrm{tr}\left[QAB^T\right] - \mathrm{tr}\left[BA^TQ^T\right] + \mathrm{tr}\left[BB^T\right] =$$

$$= ||A||_F^2 - 2\mathrm{tr}\left[QAB^T\right] + ||B||_F^2.$$

Так как мы минимизируем $||QA-B||_F^2$, нужно максимизировать tr $\left[QAB^T\right]$, а нормы матриц A, B - константы. Применим SVD разложение для AB^T и распишем след:

$$\mathrm{tr}\left[QAB^T\right] = \mathrm{tr}\left[Q(U\Sigma V^T)\right] = \mathrm{tr}\left[(QU\sqrt{\Sigma})(\sqrt{\Sigma}V)^T\right] =$$

$$= \left\langle QU\sqrt{\Sigma}, \sqrt{\Sigma}V\right\rangle \leqslant ||QU\sqrt{\Sigma}||_F^2 \cdot ||\sqrt{\Sigma}V||_F^2 =$$

$$= ||\sqrt{\Sigma}||_F^2 \cdot ||\sqrt{\Sigma}V||_F^2 = \mathrm{tr}\left[\sqrt{\Sigma}\sqrt{\Sigma}\right] = \mathrm{tr}\Sigma.$$

Здесь использовали следующие факты: $\sqrt{\Sigma}$ диагональна, норма Фробениуса произведения диагональной и ортогональной матрицы равна норме диагональной, QU и V ортогональны, а также применили неравенство Коши-Буняковского. В полученном выражении получается равенство, если положить $Q = VU^T$:

$$\mathrm{tr}\left[QU\Sigma V^T\right] = \mathrm{tr}\left[(VU^T)U\Sigma V^T\right] = \mathrm{tr}\left[V^TV\Sigma\right] = \mathrm{tr}\left[\Sigma\right].$$

Следовательно, норма Фробениуса разности $||QA-B||_F$ минимальна при $Q = VU^T$, где $U\Sigma V^T$ - SVD разложение матрицы AB^T.