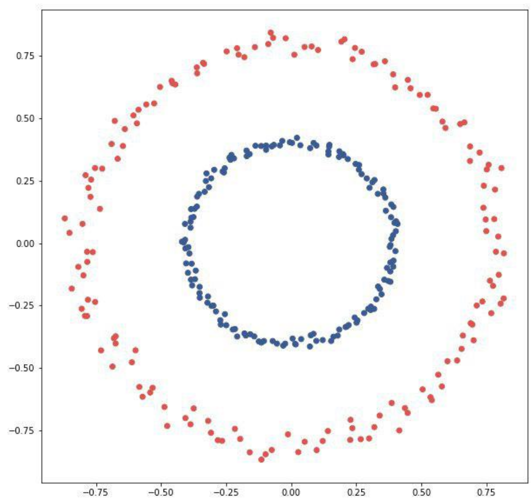
Школа анализа данных
Машинное обучение, часть 2
Домашнее задание №1

Кузина Е.М.

Решите предложенные задачи. Решения необходимо оформить в виде PDF документа. Каждая задача должна быть подробно обоснована, задачи без обоснования не засчитываются. Решения пишутся в свободной форме, однако так, чтобы проверяющие смогли разобраться. Если проверяющие не смогут разобраться в решении какой-нибудь задачи, то она автоматически не засчитывается. Дедлайн очников 15 октября 2018 09:00MSK, дедлайн заочников и филиалов +2 суток.

Задача 1 (0.5 балла) Нейронные сети.
Дана выборка из двух концентрических окружностей:



Допустим, что для классификации нужно обучить нейронную сеть — причем доступны только следующие слои: линейный $L(n, m)$ ($Wx + b,\ x \in \mathbb{R}^n,\ b \in \mathbb{R}^m$) и активация A (сигмоида или tanh), которые разрешено последовательно ставить друг после друга.

Вопрос: какие из приведенных ниже архитектур будут способны разделить выборку со 100% accuracy? Почему?

1. $L(2,2) \to A \to L(2,1)$
2. $L(2,2) \to A \to L(2,2) \to A \to L(2,1)$
3. $L(2,3) \to L(3,1)$
4. $L(2,3) \to A \to L(3,1)$
5. $L(2,3) \to L(3,3) \to L(3,1)$

Решение: Понятно, что архитектуры под номерами 3 и 5 не подходят для данной задачи, так как они строят линейную разделяющую поверхность. Архитектуры под пунктами 1 и 2 тоже не будут распознавать окружности со 100% accuracy, потому что для получения такой оценки нужно "выгибать"двухмерное пространство (Рис. 2) в пространство большей размерности, а в этих пунктах преобразуется двухмерное пространство в двухмерное (Рис.1). Поэтому ответом будет архитектура под номером 4.

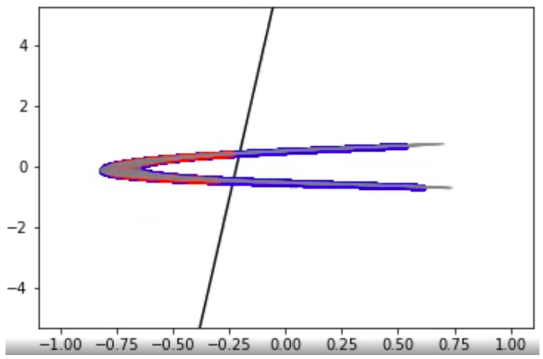


Рис. 1

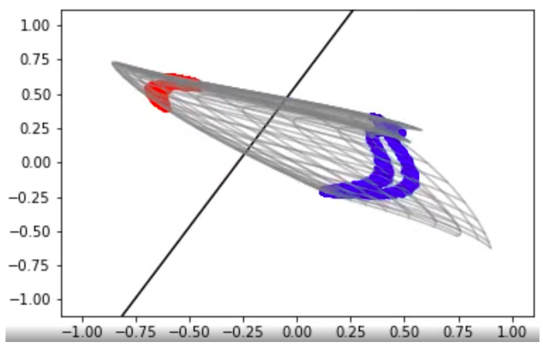


Рис. 2

Задача 2 (1.5 балла) Нейронные сети, back-prop.

Рассмотрим двуслойную полносвязную нейронную сеть, применяемую для задачи классификации. На вход нейронной сети подается вектор признаков x размерности n, полносвязный слой с матрицей весов W размерности $n \times d$ преобразует вектор x в скрытое представление h некоторой размерности d:

$$h = xW$$

Функции активации нет, еще один полносвязный слой с матрицей весов W' размерности $d \times m$ преобразует скрытое представление в вектор оценок a принадлежности к каждому классу. Чтобы получить из этих оценок вероятности, используется softmax. Например, вероятность того, что объект, описываемый вектором признаков x, относится к классу j согласно нейронной сети выглядит так:

$$p_j = \frac{\exp(a_j)}{\sum_{k=1}^{m} \exp(a_k)}$$

В качестве функции потерь используется cross-entropy loss:

$$\mathcal{L} = -\sum_{j=1}^{m} y_j \log p_j,$$

где y – one-hot encoding истинной метки объекта.

Итак, мы полностью описали проход по нейронной сети вперед: как по входному вектору x найти вероятности классов p_j и вычислить значение функции потерь, зная ответ y на рассматриваемом объекте. Опишите обратный проход по нейронной сети: выпишите формулы изменения матриц весов W и W' в стохастическом градиентном спуске для метода обратного распространения ошибки (backpropagation).

Решение:

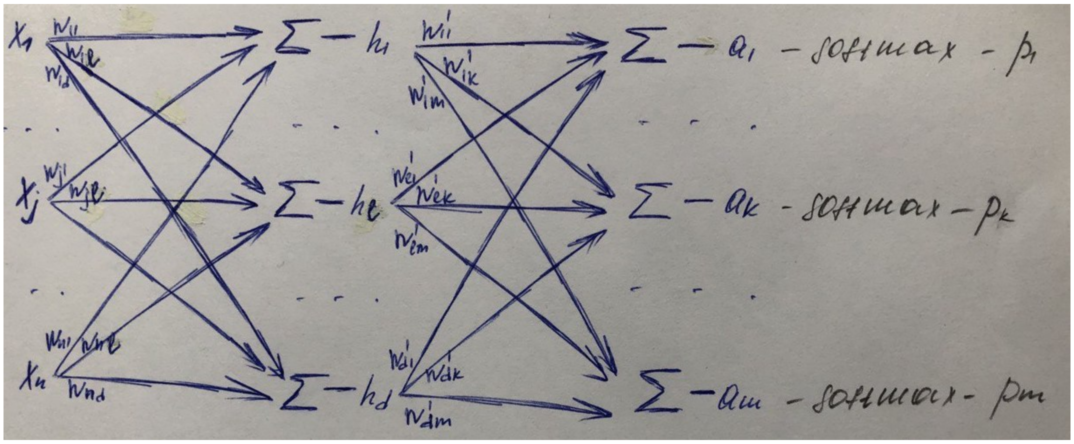


Рис. 3

$$h_l = \sum_{i=1}^{n} x_i w_{il} \qquad a_k = \sum_{i=1}^{d} h_i w'_{ik} \qquad p_k = \frac{exp(a_k)}{\sum\limits_{i=1}^{m} exp(a_i)} \qquad \mathcal{L} = -\sum_{i=1}^{m} y_i \log(p_i)$$

$$\frac{\partial \mathcal{L}}{\partial w'_{lk}} = \frac{\partial \mathcal{L}}{\partial p_k}\frac{\partial p_k}{\partial a_k}\frac{\partial a_k}{\partial w'_{lk}} = -\frac{y_k}{p_k}\frac{exp(a_k)\sum\limits_{i=1}^{m} exp(a_i) - exp(2a_k)}{(\sum\limits_{i=1}^{m} exp(a_i))^2}h_l = -\frac{y_k(\sum\limits_{i=1}^{m} exp(a_i) - exp(a_k))}{\sum\limits_{i=1}^{m} exp(a_i)}h_l$$

$$\frac{\partial \mathcal{L}}{\partial w_{jl}} = \frac{\partial \mathcal{L}}{\partial h_l}\frac{\partial h_l}{\partial w_{jl}} = -\sum_{i=1}^{m}\frac{y_i(\sum\limits_{j=1}^{m} exp(a_j) - exp(a_i))}{\sum\limits_{j=1}^{m} exp(a_j)}w'_{li}x_j$$

Обратный проход:
$w'_{lk} = w'_{lk} - \mu \frac{\partial \mathcal{L}}{\partial w'_{lk}}$
$w_{jl} = w_{jl} - \mu \frac{\partial \mathcal{L}}{\partial w_{jl}}$
μ - размер шага, x_j - j-ый признак x

Задача 3. Нейронные сети, инициализация весов.

Рассмотрим полносвязный слой нейронной сети с матрицей весов W и свободным членом b, получающий на вход вектор x размерности n и вычисляющиюющий скрытое представление размерности m

$$h = Wx + b.$$

Предложите, из какого невырожденного вероятностного распределения надо выбирать веса W и b, чтобы активации h имели нормальное распределение $N(0, \sigma^2)$, если

(a) **(1 балл)** Все признаки независимы и распределены по стандартному нормальному закону.

(b) **(2 балла)** Все признаки независимы и распределены равномерно от 0 до a.

Распределения W и b не обязаны совпадать, они могут быть из разных семейств.

Задача 4 (1.5 балла) Композиции алгоритмов, бустинг, AdaBoost.
Обозначим через $\tilde{w}^{(N)}$ нормированный вектор весов на N-й итерации алгоритма AdaBoost. Покажите, что взвешенная ошибка базового классификатора b_N относительно весов со следующего шага $\tilde{w}_i^{(N+1)}$ равна 1/2:

$$\sum_{i=1}^{\ell} \tilde{w}_i^{(N+1)}[b_N(x_i) \neq y_i] = \frac{1}{2}.$$

Решение:
Теория к решению и обозначения взяты здесь

$$\sum_{i=1}^{\ell} \tilde{w}_i^{(N+1)}[b_N(x_i) \neq y_i] = \sum_{i=1}^{\ell} \frac{w_i^{(N+1)}[b_N(x_i) \neq y_i]}{\sum_{j=1}^{\ell} w_j^{(N+1)}} = \sum_{i=1}^{\ell} \frac{w_i^{(N)} e^{-\alpha_t y_i b_N(x_i)}[b_N(x_i) \neq y_i]}{\sum_{j=1}^{\ell} w_j^{(N)} e^{-\alpha_t y_j b_N(x_j)}} =$$

$$= \frac{e^{\alpha_t} \sum_{i=1}^{\ell} w_i^{(N)}[b_N(x_i) \neq y_i]}{e^{\alpha_t} \sum_{j=1}^{\ell} w_j^{(N)}[b_N(x_j) \neq y_j] + e^{-\alpha_t} \sum_{j=1}^{\ell} w_j^{(N)}[b_N(x_j) = y_j]} = \frac{\sqrt{\frac{1-Q}{Q}}Q}{\sqrt{\frac{1-Q}{Q}}Q + (1-Q)\sqrt{\frac{Q}{1-Q}}} = \frac{1}{2}$$

Последнее равенство возможно, потому что $\sum_{j=1}^{\ell} w_j^{(N)}[b_N(x_j) = y_j] + \sum_{j=1}^{\ell} w_j^{(N)}[b_N(x_j) \neq y_j] = = \sum_{j=1}^{\ell} w_j^{(N)} = 1$, т.к. w - нормированный вектор.

Задача 5 (2 балла) Градиентный бустинг.
Примечание: в этой и 6 задаче используется теория из лекций Соколова ссылка

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

   Решение:

   В градиентном бустинге остаток $s_i^{(N)} = -\frac{\partial \mathcal{L}}{\partial z}\big|_{z=a_{N-1}(x_i)}$. Нужно найти такую функцию потерь $\mathcal{L}$, чтобы $s_i^{(N)} = y_i - a_{N-1}(x_i)$. Очевидно, что тогда $\mathcal{L}(\tilde{y}_i, y_i) = \frac{1}{2}(\tilde{y}_i - y_i)^2$, где y_i - истинный ответ на i-ом объекте, $\tilde{y}_i$ - его предсказание.

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

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

   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 $(6, 8, 6, 4, 1)$?

   Решение:

   $s_i^{(N)} = -\frac{\partial \mathcal{L}}{\partial z}\big|_{z=a_{N-1}(x_i)} = -4(\tilde{y}_i - y_i)^3 => s^N = (4, -32, 0, 4, 4)$

   Тогда вектор ответов, на который будет настраиваться следующий базовый алгоритм, равен сумме ответа на данной итерации и остатка: $\tilde{y} + s = (5, 10, 6, 3, 0) + (4, -32, 0, 4, 4) = (9, -22, 6, 7, 4)$.

3. Рассмотрим задачу бинарной классификации, $Y = \{0, 1\}$. Будем считать, что все алгоритмы из базового семейства $\mathcal{A}$ 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):

$$L(y, z) = -\big(y \log z + (1 - y) \log(1 - z)\big),$$

где y — правильный ответ, а z — ответ алгоритма.

Выпишите формулы для поиска базовых алгоритмов b_n и коэффициентов γ_n в градиентном бустинге.

Решение:

Преобразуем метки класссов в -1 и 1: $y_{new} = 2y - 1$, тогда функция потерь L(y, z) преобразуется в $\tilde{L}(y_{new}, a(x_i)) = \log(1 + \exp(-2y_{new}a(x_i)))$, где $a(x_i) = \frac{1}{2}\log\frac{z(x_i)}{1-z(x_i)}$ ссылка.

Теперь будем строить градиентный бустинг, где $a_N(z) = \sum\limits_{n=0}^{N} \gamma_n b_n(x)$.

Пусть b_0 возвращает самый популярный класс и $\gamma_0 = 1$.

$b_n = \underset{b \in A}{argmin} \sum\limits_{i=1}^{l} (b(x_i) - s_i)^2 = \underset{b \in A}{argmin} \sum\limits_{i=1}^{l} (b(x_i) + \frac{\partial \tilde{L}(y_{new\,i}, a_{n-1})}{\partial a_{n-1}})^2 = \underset{b \in A}{argmin} \sum\limits_{i=1}^{l} (b(x_i) - \frac{2y_{new\,i}}{1+exp(2y_{new\,i}a_{n-1}(x_i))})^2$

$\gamma_n = \underset{\gamma \in R}{argmin} \sum\limits_{i=1}^{l} \tilde{L}(y_{new\,i}, a_{n-1}(x_i) + \gamma b_n(x_i)) = \underset{\gamma \in R}{argmin} \sum\limits_{i=1}^{l} \log\Bigg(1 + \exp\big(-2(2y_i - 1)(a_{n-1}(x_i) +$
$+ \gamma b_n(x_i))\big)\Bigg)$

Задача 6 (1.5 балла) Композиции, устойчивость к шуму.

1. Рассмотрим алгоритм AdaBoost — бустинг с экспоненциальной функцией потерь

$$\mathcal{L}(M) = \exp(-M),$$

где M — отступ объекта. Покажите, что алгоритм неустойчив к шуму, т.е. возможен неограниченный рост отношения весов шумовых объектов по отношению к весам пороговых объектов.

Решение: В задачах классификации $s_i = y_i w_i = -\frac{\partial \mathcal{L}(y_i a_{N-1}(x_i))}{\partial a_{N-1}(x_i)}$, где $y_i a_{N-1}(x_i) = M$ - отступ объекта.

$$s_i = y_i exp(-y_i a_{N-1}(x_i)) => w_i = exp(-y_i a_{N-1}(x_i))$$

$\dfrac{w_{\text{шум. объект}}}{w_{\text{порог. объект}}} = \dfrac{exp(-M_{\text{шум}})}{exp(-M_{\text{порог.}})}$ неограниченно возрастает, так как числитель не ограничен при

больших положительных значениях (минус большой отрицательный отступ = большому положительному числу), а знаменатель приблизительно равен 1, так как отступ на нем положительный и порядка нуля.

2. Покажите, что бустинг с логистической функцией потерь

$$\mathcal{L}(M) = \log(1 + \exp(-M))$$

устойчив к шуму в описанном выше смысле смысле.

Решение:

$$s_i = y_i \frac{exp(-M)}{1 + exp(-M)} = y_i \frac{1}{1 + exp(M)} => w_i = \frac{1}{1 + exp(M)}$$

$$\frac{w_{\text{шум. объект}}}{w_{\text{порог. объект}}} = \frac{1 + exp(M_{\text{порог.}})}{1 + exp(M_{\text{шум}})} \approx 2 \text{ (аналогичные рассуждения с пунктом 1)}$$

Примечание. Пороговые объекты — это те, для которых значение отступа положительно и порядка нуля, то есть они лежат близко к границе между классами и в своем классе. Шумовые объекты лежат глубоко в чужом классе, на них отступ принимает большие отрицательные значения.