Школа анализа данных
Машинное обучение, часть 1
Теоретическое домашнее задание №2

Федор Ерин

Задача 1 (0.5 балла) Кроссвалидация, LOO, k-fold.

Объясните, стоит ли использовать оценку leave-one-out-CV или k-fold-CV с небольшим k в случае, когда:

- обучающая выборка содержит очень малое количество объектов;

  Здесь нужно использовать leave-one-out-CV, потому что в этом случае модель вычислительно легко обучить много раз, так как данных мало, а так же мы будем каждый раз оставлять в трейне максимум данных, из которых выявим закономерности, а проверим качество на небольшом тесте из 1 объекта. Тем самым мы не оставим в тесте слишком много данных, которые следовало бы показать модели.

- обучающая выборка содержит очень большое количество объектов.

  Здесь нужно использовать k-fold-CV с небольшим k, потому что чем больше k, тем более вычислительно трудоемко провести такую кросс-валидацию, нужно много раз обучить заново модель, а данных у нас много. Поэтому следует ограничиться небольшим кол-вом сплитов, и каждый раз в трейне будет достаточно данных, чтобы выявить закономерности.

Задача 2 (1.5 балла). Логистическая регрессия, решение оптимизационной задачи.

1. (0.5 балла) Докажите, что в случае линейно разделимой выборки не существует вектора параметров (весов), который бы максимизировал правдоподобие вероятностной модели логистической регрессии в задаче двухклассовой классификации.

   В данном случае модель будет переобучаться, все больше приближая выход сигмоиды к значениям 0 и 1, а вектор весов w при этом будет устремляться в бесконечность. Алгоритм будет бесконечно шагать по пространству признаков, бесконечно увеличивая правдоподобие.

2. (0.3 балла) Предложите, как можно модифицировать модель, чтобы оптимум достигался.

   Можно использовать регуляризацию и ограничить рост нормы весов w, тогда начиная с какого-то момента увеличивать веса не будет целесообразно с т.з. прироста правдопободия и алгоритм сойдется.

3. (0.7 балла) Что можно сказать о единственности решения L2-регуляризованной задачи? Почему?

   При L2-регуляризации решение задачи единственно, потому что минимизация функционала ошибки с L2 соответствует поиску среднего арифметического, а это значение однозначно и единственно. Если же рассмотреть с другой

стороны, с L2-регуляризацией мы имеем сумму двух квадратичных фукнций, каждая из которых выпукла и имеет один глобальный минимум, а значит их сумма тоже.

Если добавить для сравнения ситуация с L1, то там мы ищем медиану, которая неоднозначна, если число элементов четно, а значит решение будет не единственно, ведь любая точка между двумя средними элементами будет давать одинаковое значение функционала ошибки.

Задача 3 (0,5 балла). L^2-регуляризация.
Докажите, что L^2-регуляризованную линейную регрессию можно переписать в виде задачи наименьших квадратов для модифицированных данных.

Требуется совершить переход от

$$L(f_w, X, y) = \|Xw - y\|^2 + \lambda \|w\|^2$$

к

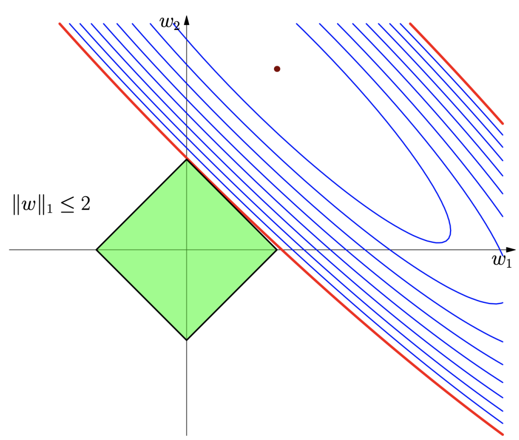
$$L'(f_w, X', y') = \|X'w - y'\|^2,$$

где X', y' - модифицированные данные. Добавим (np.vstack) к матрице X квадратную матрицу размера (n, n), где n - кол-во признаков, состоящую из нулей и элементов $\sqrt{\lambda}$ на главной диагонали. Также к вектору y допишем n нулей. В этом случае, при перемножении Xw мы будем получать исходную часть суммы функционала ошибки без $L2$ регуляризации и плюс $\lambda \|w\|^2$. Таким преобразованием мы перешли к задаче МНК.

Задача 4 (1.5 балла). L^1-регуляризация.
Рассмотрим задачу L^1-регуляризованной линейной регрессии, в которой ранг матрицы X меньше числа признаков D.

1. (0.5 балла) Докажите, что если у задачи более одного решения, то решений бесконечно много.

   Если ранг X меньше числа признаков, значит среди признаков если линейно зависимые или сильно скоррелированные. В этом случае эллипсоиды линий уровня будут растягиваться, постепенно превращаясь в параллельные линии и в какой-то момент практически совпадут с ребром ромбом $\|w\| = t$, где в силу ограниченной точности вычислений могут получиться равные значения в нескольких точках, и тогда не только 2 общие точки могут быть, но и бесконечно много. Визуально это может выглядеть так (лекции New York University, 2017 г.):



2. (0.5 балла) Докажите, что в этом случае для всех решений $\widehat{w}$ значение $X\widehat{w}$ одно и то же.

   Так как решение множество, значит при всех них достигается один и тот же минимум функционала ошибки на train'е, а следовательно Xw будет давать одинаковые прогнозы для train'а.

3. (0.5 балла) Докажите, что L^1-нормы всех решений $\widehat{w}$ одинаковы.

   Поскольку значения Xw одинаковые и $loss$ одинаковый для всех решений w, то есть равны результаты сумму двух слагаемых и первые из слагаемых, следовательно, равны и вторые слагаемые (L^1 нормы): $\lambda\|w\|$.

Задача 5 (1 балл). Обобщённая линейная модель.

Напомним, что гамма-распределение задаётся функцией плотности:

$$p(y \mid a, b) = \frac{1}{\Gamma(a)b^a} y^{a-1} e^{-\frac{y}{b}}$$

где $\Gamma(a)$ — гамма-функция

1. (0.3 балла) Докажите, что семейство гамма-распределений относится к экспоненциальному классу.

   Экспоненциальное семейство распределение задается в следующем виде:

$$f(y) = \frac{1}{h(\theta)} g(y) e^{\theta^T u(y)},$$

$$h(\theta) \in \mathrm{R}^1, y \in \mathrm{R}^m, \theta \in \mathrm{R}^\alpha, \alpha \ll m, u(y) = (u_1(y), ..., u_\alpha(y))^T$$

   Принимая $\theta = (-\frac{1}{b}, a-1), u(y) = (y, \ln y)^T, g(y) = 1, h(\theta) = \Gamma(a)b^a$, получаем в точности функцию плотности гамма-распределения.

2. (0.7 балла) Как будет выглядеть соответствующая гамма-распределению обобщённая линейная модель? Найдите каноническую функцию связи и функционал, который надо оптимизировать, чтобы найти веса обобщённой линейной модели.

   Ищем такую функцию g, для которой $g(\mathbf{E}(y|x)) = \langle x, w \rangle$, это функция связи. Перепишем функцию гамма-распределения в виде:

$$p(y \mid a, b) = \frac{1}{\Gamma(a)b^a} y^{a-1} e^{-\frac{y}{b}} = exp(\frac{-y - \ln\Gamma(a)b^{a+1}}{b} + (a-1)\ln y))$$

   $u_1(y) = -\frac{y}{b}$, откуда $\mu = -b\mathrm{E}u_1(y) = -b(\ln\Gamma(a)b^a)'_a = f'(a)$

   Положим $a = \langle x, w \rangle$, а поскольку $\mathbf{E}(y|x) = g^{-1}(\langle x, w \rangle)$ и $\mathbf{E} = f'(a)$, то получаем каноническую функцию связи: $g = (f')^{-1}$.

Задача 6 (4 балла) Обратное распространение ошибки.

В этой задаче вам нужно будет сделать простую вещь: написать формулы обратного распространения ошибки для нескольких слоёв.

В контесте сданы все пункты:

Время посылки	ID	Задача	Компилятор	Вердикт	Тип посылки	Время	Память	Тест	Баллы	
11 ноя 2021, 00:44:11	56865099	A	(make) yandexdataschool	OK	-	452ms	23.11Mb	-	4	отчёт

1. **(0.5 балла)** LeakyReLU;

   (a) $output = \begin{cases} x & \text{если } x > 0, \\ slope \cdot x & \text{иначе.} \end{cases}$ x – вход слоя (input), $slope$ - коэф-т угла наклона

   (b) $grad\_input = output\_grad \odot \begin{cases} 1 & \text{если } x > 0, \\ slope & \text{иначе.} \end{cases}$

2. **(0.5 балла)** SoftPlus;

   (a) $output = \ln(1 + e^x)$

   (b) $grad\_input = output\_grad \odot \sigma(x)$, σ - логистическая функция (сигмоида)

3. **(1 балл)** LogSoftMax;

   (a) $output = \ln(softmax(x - max(x, axis = 1)))$, где $\text{softmax}(x)_i = \frac{\exp x_i}{\sum_j \exp x_j}$

   (b) $grad\_input = output\_grad - softmax(x) \odot sum(output\_grad, axis = 1)$

4. **(1 балл)** нестабильную версию Negative log likelihood;

   (a) $output = -sum(target \odot \ln x)/N$, где N - кол-во объектов в выборке

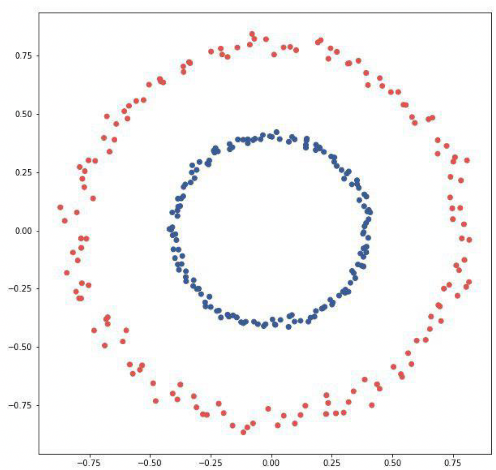
   (b) $grad\_input = -(target/x)/N$

5. **(1 балл)** более стабильную версию Negative log likelihood.

   (a) $output = -sum(target \odot \ln(e^x))/N$

   (b) $grad\_input = -((target/e^x) \odot e^x)/N$

Задача 7 (0.5 балла) Нейронные сети.
Дана выборка из двух концентрических окружностей:



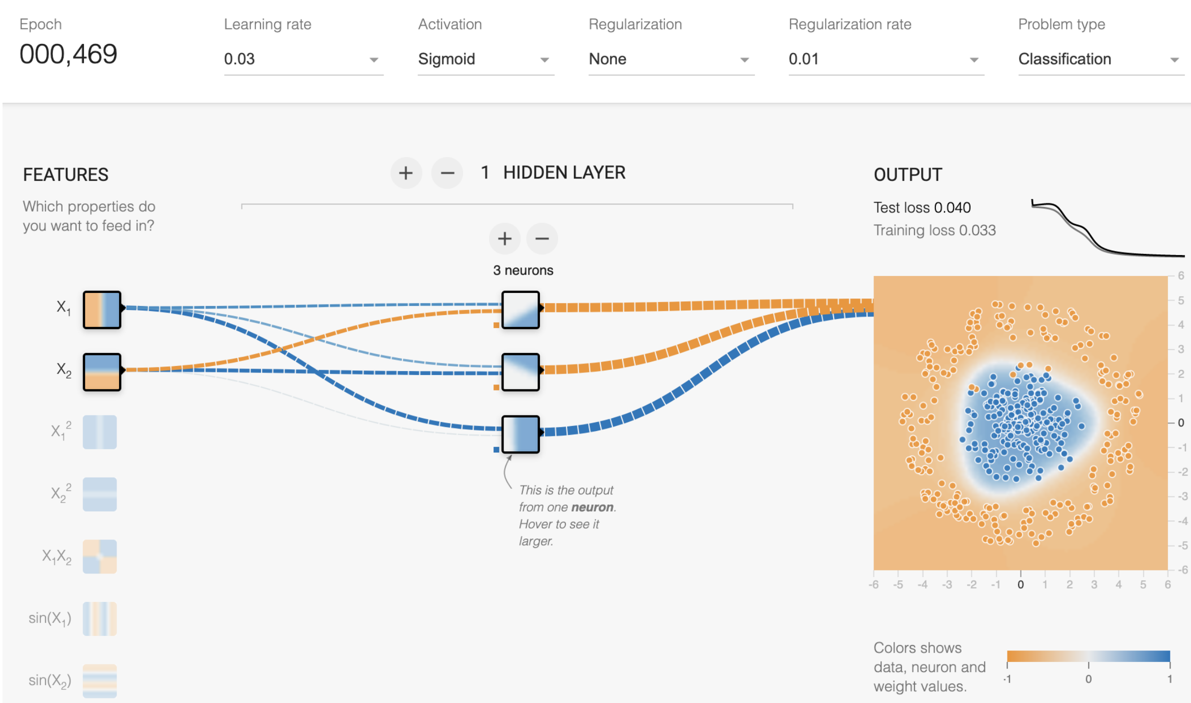
Допустим, что для классификации нужно обучить нейронную сеть — причем доступны только следующие слои: линейный $L(n, m)$ ($Wx + b,\ x \in \mathbb{R}^n,\ b \in \mathbb{R}^m$) и активация A (сигмоида или tanh), которые разрешено последовательно ставить друг после друга.

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

1. $L(2,2) \to A \to L(2,1)$ - ***нет***
2. $L(2,2) \to A \to L(2,2) \to A \to L(2,1)$ - ***нет***
3. $L(2,3) \to L(3,1)$ - ***нет***
4. $L(2,3) \to A \to L(3,1)$ - ***да***
5. $L(2,3) \to L(3,3) \to L(3,1)$ - ***нет***

Выпуклую область можно получить, проведя, как минимум, 3 линии, за которые отвечают 3 нейрона, и выполнив логическое "И" выходным нейроном. Таким образом мы задаем условие, при котором точки (объекты) должны лежать по определенную сторону от каждой из прямой. В этом случае достаточно одного скрытого слояиз 3 нейронов. При этом важно наличие нелинейной функции активации, так как выборка линейно неразделима, линейная модель здесь не подойдет. Под данные критерии подходит только архитектура номер 4.

Если объекты обучающей выборки располагаются в виде круга, то очертив 3 линиями треугольник, нейроны будут не очень уверены в себе (длины векторов будут небольшими), из-за этого границы будут размываться, а углы треугольника будут сглаживаться, и он превратится в область, схожую с окружностью. Проверить это можно, например, на сайте $playground.tensorflow.org$***:***



Задача 8 (0.5 балла) Нейронные сети, калибровка.

Глубокие нейронные сети часто являются плохо скалиброванными моделями. Что с ними не так? Почему?

Глубокие нейронные сети - очень сильные классификаторы, при обучении на метки классов (а не на вероятности) они выучиваются давать ответы, наиболее близкие к 0 и 1. При обучении нейросеть смотрит только за точностью (например, accuracy), но не за уверенностью. Данную проблему можно

было бы увидеть на диаграмме калибровки, а превратить выход модели в вероятности можно с помощью соответствующих методов (гистограммный, изотонная регрессия, калибровка Платта и др.)