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

Кошман Дмитрий

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

В случае с малым количеством объектов больше подходит LOO-CV, поскольку в этом случае увеличение размера обучающей выборки на 1 объект сильно влияет на качество модели. То есть при увеличении тестовой выборки на 1 качество модели падает сильнее, чем возрастает качество оценки этой модели. - уменьшается разброс между качеством на тесте и трейне, но и качество в среднем падает. Еще следует отметить, что поскольку объектов мало, то и обучить модель столько раз, сколько объектов в выборке можно за адекватное количество времени.

В случае с большим количеством объектов обучать модель столько раз, сколько объектов в выборке займет слишком много времени, поэтому LOO-CV не подходит. Если она очень большая, то даже KFold-CV будет работать слишком долго и лучше просто пользоваться HoldOut. Если нет, KFold-CV скорее всего успеет сойтись на 4/5 объектов и среднее качество не будет сильно отличаться от качества при обучении на всех объектах, но при этом будет хорошая оценка на тестовом фолде, ведь в нем тоже много объектов.

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

1. (0.5 балла) Пусть классы имеют метки $\{-1, 1\}$. Если выборка линейно разделима, это означает, что существует такое w', что $\forall i : y_i \langle w', x_i \rangle < 0$. Предположим, что существует w, максимизирующее правдоподобие вероятностной модели логистической регрессии, то есть для w достигается максимум

$$L(w, X, y) = \sum_{i=1}^{N} \log \left(1 + e^{-y_i \langle w, x_i \rangle}\right)$$

   Но тогда если взять $w + w'$, правдоподобие увеличится - противоречие.

2. (0.3 балла) Предложите, как можно модифицировать модель, чтобы оптимум достигался.
   Если в вероятностной модели предположить не только существование истинной зависимости между признаками и вероятностью положительного класса, но и априорное распределение на параметрах модели, объясняющееся неточностью измерений, представлений или наличием шума, то модель модифицируется таким образом:

$$p(X, Y, w; \sigma) = p(X, Y \mid w)\, p(w; \sigma)$$

   Принцип максимума совместного правдоподобия данных и модели:

$$L_\sigma(w, X, Y) = \ln p(X, Y, w; \sigma) = \sum_{i=1}^{n} \ln p(x_i, y_i \mid w) + \ln p(w; \sigma) \to \max_w$$

Если предположить, что w имеет нормальное распределение $w \sim N\left(0, \sigma^2\right)$, где σ - гиперпараметр, то получаем:

$$\ln p(w;\sigma) = \ln\left(\frac{1}{(2\pi\sigma)^{n/2}}\exp\left(-\frac{\|w\|^2}{2\sigma}\right)\right) = -\frac{1}{2\sigma}\|w\|^2 + \mathrm{const}(w)$$

Тогда L_σ - непрерывная по w функция, стремящаяся к минус бесконечности при $w \to \infty$, значит она достигает максимума.

3. (0.7 балла) В случае L2- регуляризации логистической регрессии решение всегда единственно. Посмотрим на матрицу вторых производных:

$$\nabla L = -\frac{1}{\sigma}E - X^T D X$$

где D - матрица с положительными числами на диагонали. Тогда:

$$u^T \nabla L u = -\|u\|^2 - \|D^{1/2}Xu\|^2 < 0$$

значит максимум единственный.

Задача 3 (0,5 балла). L^2**-регуляризация.**
Модифицируем данные следующим образом:

$Y' = X^T Y$
$X' = X^T X + \lambda I$

Тогда решение задачи наименьших квадратов для этих данных такого:
$w = (X'^T X')^{-1} X'^T Y'$
Заметим, что X' симметрична и положительна определена. Значит,

$w = X'^{-1} Y' = (X^T X + \lambda I)^{-1} X^T Y$

Что есть решение для L^2-регуляризованной линейной регрессии.

Задача 4 (1.5 балла). L^1**-регуляризация.**

1. (0.5 балла)

   Поскольку L^1-регуляризованная линейная регрессия представляет собой сумму квадрата L^2 нормы линейного по w выражения и L^1 нормы w, домноженной на положительную константу, то в целом выражение является выпуклой по w функцией как сумма выпуклых функций. Значит, если минимум выпуклой функции достигается в двух точках, то между этими точками значение функции также равно минимуму, и точек минимума континуум.

2. (0.5 балла)

   Рассмотрим два решения $\widehat{w}$ и w. По рассуждениям из предыдущего пункта в точках вида $w + \alpha(\widehat{w} - w)$ функция потерь также равна минимуму для $\alpha \in [0, 1]$. Тогда скалярная по α функция

$$|X(w + \alpha(\widehat{w} - w)) - y|_2^2 + \lambda|w + \alpha(\widehat{w} - w)|_1$$

   Представляет собой сумму полинома второй степени и непрерывной кусочно-линейной функции, которая может постоянна на отрезке $[0, 1]$ только если коэффициент перед квадратом равен нулю, то есть если

$$|X(\widehat{w} - w)|_2^2 = 0 \Rightarrow X(\widehat{w} - w) = 0 \Rightarrow X\widehat{w} = Xw$$

3. (0.5 балла)

   Поскольку для решений $\widehat{w}$ и w значения равны $X\widehat{w}$ и Xw, то

$$min = |Xw - y|_2^2 + \lambda|w|_1 = |X\widehat{w} - y|_2^2 + \lambda|w|_1 = |X\widehat{w} - y|_2^2 + \lambda|\widehat{w}|_1$$

$$|w|_1 = |\widehat{w}|_1$$

1. (0.3 балла)

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

   Если положить $u_1(x) = \ln x, u_2(x) = -x, \theta_1 = a - 1, \theta_2 = \frac{1}{b}$, то можно переписать функцию плотности:

$$p(y \mid \theta) = \frac{exp\{\theta^T u(y)\}}{\Gamma(\theta_1 + 1)\theta_2^{-\theta_1 - 1}} = \frac{exp\{\theta^T u(y)\}}{h(\theta)}$$

   Откуда видно, что гамма-распределение относится к экспоненциальному классу.

2. (0.7 балла)

   Положим $\phi = -\frac{1}{a}, \theta = \frac{1}{ab}$. Тогда $a = -\frac{1}{\phi}, b = -\frac{\phi}{\theta}$ и плотность вероятности представляется в виде

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

$$= exp\{(-\frac{1}{\phi} - 1)\ln y + \frac{\theta y}{\phi} - \ln\Gamma(-\frac{1}{\phi}) + \frac{1}{\phi}\ln -\frac{\phi}{\theta}\} =$$

$$= exp\{\frac{\theta y - \ln\theta}{\phi} + (-\frac{1}{\phi} - 1)\ln y - \ln\Gamma(-\frac{1}{\phi}) + \frac{1}{\phi}\ln -\phi\} =$$

$$= exp\{\frac{\theta y - h(\theta)}{\phi} + g(y, \phi)\}$$

   где $h(x) = \ln x, h'(x) = \frac{1}{x}$. Значит, мы рассматриваем параметризованное семейство $\Gamma(\frac{1}{\sigma}, \frac{\sigma}{<x,w>})$ с фиксированной σ, каноничекая функция связи равна $g(x) = (h'(x))^{-1} = \frac{1}{x}$, и нужно оптимизировать функционал

$$\sum_i |g^{-1}(< x_i, w >) - y_i|^2 = \sum_i \left|\frac{1}{< x_i, w >} - y_i\right|^2$$

Задача 7 (0.5 балла) Нейронные сети.

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. $L(2,3) \to L(3,3) \to L(3,1)$

   Простым линейным преобразованием нельзя сделать так, чтобы две вложенные окружности оказались на разных полуплоскостях, поскольку линейное преобразование сохраняет порядок на прямой. Значит, варианты 3 и 5 не подходят.

   В первом варианте похожая проблема. После любого первого преобразования $Wx + b$ всегда найдутся такие точки x, y из внешней окружности и z из внутренней, что $w_1x+b_1 < w_1z+b_1 < w_1y+b_1$. Поскольку сигмоида монотонна, то этот порядок на первых координатах сохранится и последний линейный слой не сможет разделить эти три точки. Во втором варианте то же самое, поскольку второй линейный слой будет работать с искаженными, но все еще вложенными окружностями, для которых приведенное рассуждение остается верным.

   Остается только четвертый вариант. Он сможет разделить эти две окружности, в первом слое проведя три прямые, продолжающие стороны вписанного равностороннего треугольника во внешнюю окружность, а в последнем слое взять сумму координат. Изменяя масштаб коэффициентов в первом слое и границу деления при формировании ответов, можно добиться полного разделения двух классов. Если же окружности лежали плотнее, то для их разделения понадобилось бы больше нейронов в слое активации.

Задача 8 (0.5 балла) Нейронные сети, калибровка.

Ответы нейронной сети на задачу классификации часто не пропорциональны истинным вероятностям классов, поскольку сети обучаются, основываясь на точность предсказаний, а не калибровку. Также глубина сети напрямую связана со сложностью модели, а сложные модели легко переобучаются. Вообще нелинейные преобразования плохо интерпретируются, и без явно поставленной задачи хорошей калибровки нет причин ожидать калиброванную модель на выходе.