Машинное обучение, ФКН ВШЭ
Теоретическое домашнее задание №1
Линейные модели

Задача 1. Скоро первая самостоятельная работа. Чтобы подготовиться к ней, ФКН ест конфеты и решает задачи. Число решённых задач y зависит от числа съеденных конфет x. Если студент не съел ни одной конфеты, то он не хочет решать задачи. Поэтому для описания зависимости числа решённых задач от числа съеденных конфет используется линейная модель с одним признаком без константы $y_i = w \cdot x_i$. В аналитическом виде найдите оценки параметра w, минимизируя следующие функции потерь:

1. Линейная регрессия без штрафа: $Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - wx_i)^2$;

   Ответ: Чтобы минимизировать Q(w), найдём его производную и приравняем к нулю:

$$Q'(w) = \frac{-2}{\ell}\sum_{i=1}^{\ell} x_i(y_i - wx_i) = 0$$

$$\sum_{i=1}^{\ell} x_i y_i = w\sum_{i=1}^{\ell} x_i^2$$

$$\hat{w} = \frac{\sum_{i=1}^{\ell} x_i y_i}{\sum_{i=1}^{\ell} x_i^2}$$

2. Ridge-регрессия: $Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - wx_i)^2 + \lambda w^2$;

$$Q'(w) = 2\lambda w + \frac{-1}{\ell}\sum_{i=1}^{\ell} x_i(y_i - wx_i) = 0$$

$$2\lambda w + \frac{2}{\ell}\sum_{i=1}^{\ell} wx_i^2 = \frac{2}{\ell}\sum_{i=1}^{\ell} x_i y_i$$

$$w(2\lambda + \frac{2}{\ell}\sum_{i=1}^{\ell} x_i^2) = \frac{2}{\ell}\sum_{i=1}^{\ell} x_i y_i$$

$$\hat{w}_R = \frac{\frac{1}{\ell}\sum_{i=1}^{\ell} x_i y_i}{\lambda + \frac{1}{\ell}\sum_{i=1}^{\ell} x_i^2}$$

3. LASSO-регрессия: $Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - wx_i)^2 + \lambda|w|$;

   Если λ не слишком большая, то решение будет такое же как в обычной линейной регрессии, но если λ будет огромным, то $\lambda|w|$ будет доминировать над суммой квадратов, из-за чего выгоднее будет занулить $|w|$, то есть сделать $w = 0$

4. Пусть решения этих задач равны $\hat{w}, \hat{w}_R$ и $\hat{w}_L$ соответственно. Найдите пределы

$$\lim_{\lambda\to 0}\hat{w}_R, \quad \lim_{\lambda\to\infty}\hat{w}_R, \quad \lim_{\lambda\to 0}\hat{w}_L, \quad \lim_{\lambda\to\infty}\hat{w}_L.$$

$$\lim_{\lambda\to 0}\hat{w}_R = \frac{\sum_{i=1}^{\ell} x_i y_i}{\sum_{i=1}^{\ell} x_i^2}$$

$$\lim_{\lambda\to\infty}\hat{w}_R = 0$$

$$\lim_{\lambda\to 0}\hat{w}_L = \frac{\sum_{i=1}^{\ell} x_i y_i}{\sum_{i=1}^{\ell} x_i^2}$$

$$\lim_{\lambda\to\infty}\hat{w}_L = 0$$

5. Как можно проинтерпретировать гиперпараметр λ?

Hint: в случае Lasso-регрессии придётся повозиться с модулем. Обратите внимание на то, что $Q(w)$ парабола, это поможет корректно найти аналитическое решение. Подумайте, с чем возникнут проблемы, если у нас будет не один параметр, а сотня.

Задача 2. Вася измерил вес трёх покемонов, $y_1 = 6$, $y_2 = 6$, $y_3 = 10$. Вася хочет спрогнозировать вес следующего покемона с помощью константной модели $y_i = w$. Для оценки параметра w Вася использует целевую функцию

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - w)^2 + \lambda w^2$$

1. Найдите оптимальное w при произвольном λ.

$$Q'(w) = \frac{-2}{\ell}\sum_{i=1}^{\ell}(y_i - w) + 2\lambda w = 0$$
$$\frac{-1}{\ell}\sum_{i=1}^{\ell}(y_i - w) + \lambda w = 0$$
$$w - \frac{1}{\ell}\sum_{i=1}^{\ell} y_i + \lambda w = 0$$
$$w(1+\lambda) = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i$$
$$w = \frac{\frac{1}{\ell}\sum_{i=1}^{\ell} y_i}{1+\lambda}$$
$$w = \frac{\frac{1}{3}(6+6+10)}{1+\lambda} = \frac{22}{3(1+\lambda)}$$

2. Подберите оптимальное λ с помощью кросс-валидации leave one out («выкинь одного»). На первом шаге мы оцениваем модель на всей выборке без первого наблюдения, а на первом тестируем её. На втором шаге мы оцениваем модель на всей выборке без второго наблюдения, а на втором тестируем её. И так далее ℓ раз. Чтобы найти λ_{CV} мы минимизируем среднюю ошибку, допущенную на тестовых выборках.

$$w_1 = w_2 = \frac{\frac{1}{2}(6+10)}{1+\lambda} = \frac{8}{1+\lambda}$$
$$w_3 = \frac{\frac{1}{2}(6+6)}{1+\lambda} = \frac{6}{1+\lambda}$$
$$Q_1 = Q_2 = \left(6 - \frac{8}{1+\lambda}\right)^2 + \left(\frac{8}{1+\lambda}\right)^2\lambda = 36 - \frac{96}{1+\lambda} + (1+\lambda)\left(\frac{8}{1+\lambda}\right)^2 =$$
$$= 36 + \frac{64-96}{1+\lambda} = 36 - \frac{32}{1+\lambda}$$
$$Q_3 = \left(10 - \frac{6}{1+\lambda}\right)^2 + \left(\frac{6}{1+\lambda}\right)^2\lambda = 100 - \frac{120}{1+\lambda} + (1+\lambda)\left(\frac{6}{1+\lambda}\right)^2 =$$
$$= 100 + \frac{36-120}{1+\lambda} = 100 - \frac{84}{1+\lambda}$$
$$Q_{CV} = \frac{Q_1+Q_2+Q_3}{3} = \frac{2\left(36 - \frac{32}{1+\lambda}\right) + 100 - \frac{84}{1+\lambda}}{3} = \frac{172 - \frac{148}{1+\lambda}}{3}$$
$$Q'_{CV} = \frac{148}{3(1+\lambda)^2} \geqslant 0 \quad \forall\lambda \geqslant 0 \quad \Rightarrow \lambda = 0$$

3. Найдите оптимальное значение w при λ_{CV}, подобранном на предыдущем шаге.

$$w = \frac{w_1+w_2+w_3}{3} = \frac{8+8+6}{3} = \frac{22}{3}$$

4. Выведите формулу для λ_{CV} при произвольном количестве наблюдений.

   Поскольку у нас все Q_k - гиперболы с осью симметрии в -1, то наименьшее $Q_k \quad \forall k$ будет при $\lambda = 0$.

   $$\lambda_{CV} = 0$$

Задача 3. Убедитесь, что вы знаете ответы на следующие вопросы:

- Что такое гиперпараметр модели и чем он отличается от параметра модели?

  Ответ: гиперпараметр - параметр, который нельзя подобрать по обучающей выборке. Гиперпараметр не настраивается моделью - его задают вручную.

- Почему коэффициент регуляризации нельзя подбирать по обучающей выборке? Как подобрать оптимальное значение для коэффициента регуляризации?

  Ответ: Коэффициент регуляризации нельзя подбирать по обучающей выборке, потому что если посмотреть на функцию ошибки $Q(w) + \alpha R(w)$, то непонятно, как подбирать α. Если мы хотим минимизировать $Q(w) + \alpha R(w)$, то надо брать $\alpha = 0$, а если мы хотим минимизировать $Q(w)$, то опять же надо взять $\alpha = 0$, так как эта добавка $R(w)$ будет мешать правильному обучению модели. Подобрать оптимальное значение для коэффициента регуляризации можно

  - на новых данных (по отложенной выборке)
  - по кросс-валидации

  Стратегии перебора:

  - Grid-search - просто перебор
  - Random-search
  - AutoML (умный способ)

- Почему накладывать регуляризатор на свободный коэффициент w_0 может быть плохой идеей?

  Ответ: во-первых, ошибка возрастёт, и при этом никакой ценной информации это увеличение ошибки не будет нести для модели, ведь она не сможет поменять свободный коэффициент. Во-вторых, если мы будем штрафовать за его величину, то получится, что мы учитываем некие априорные представления о близости целевой переменной к нулю и отсутствии необходимости в учёте её смещения.

- Что такое кросс-валидация, чем она лучше использования отложенной выборки?

  Ответ: это способ обучения модели, заключающийся в том, что данные делятся на n частей, одна из которых тестовая, а остальные тренировочные. Мы можем обучить модель на остальных $n-1$ частях, а потестировать на оставшейся. Причём мы будем делать это для всех частей, то есть брать каждую часть за тестовую, а остальные $n-1$ будут обучающие. В итоге мы как будто обучим

n моделей. Кросс-авлидация хороша тем, что мы можем подбирать гиперпараметры на валидационных данных (которые являются частью тестовых!), а не на тестовых, так как если подбирать гиперпараметры на тестовых данных, мы можем неявно заложить модели информацию о тестовых данных.

- Почему категориальные признаки нельзя закодировать натуральными числами? Что такое one-hot encoding?

  Ответ: потому что мы не знаем, есть ли отношение порядка между категориями. Скорее всего нет, и если закодировать категории числами, то мы добавим несуществующее свойство этим категориям, из-за чего обучение модели может быть испорчено.

  One-hot encoding это создание числовых признаков в количестве, равном числу различных категорий в категориальном признаке. Условно, это признаки-индикаторы.

- Для чего нужно масштабировать матрицу объекты-признаки перед обучением моделей машинного обучения?

  Ответ: Чтобы веса у модели были меньше и как следствие не было переобучения.

- Почему L_1-регуляризация производит отбор признаков?

  Ответ: если в выборке есть признак, который не влияет на ответ, то допустим следующую ситуацию: модель подобрала какие-то коэффициенты w. Но она хочет минимизировать $Q(w) + \alpha||w||_1$, поэтому ей нужно как можно меньше сделать $||w||_1$. Это значит что у не влияющих на ответ признаков можно обнулить коэффициент, уменьшим при этом $||w||_1$.

- Почему MSE чувствительно к выбросам?

  Ответ: так как выброс заставляет модель веса двигать в сторону выброса, при этом разница между правильным ответом и предсказанием возводится в квадрат, а квадрат быстро возрастает и, следовательно, меняет значение MSE