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

   Для нахождения минимума $Q(w)$ найдем её производную по w и приравняем её к 0:

$$\frac{\partial Q}{\partial w} = 0$$
$$\frac{\partial}{\partial w}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - wx_i)^2\right] = 0$$
$$\frac{\partial}{\partial w}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}y_i^2 - 2wx_iy_i + w^2x_i^2\right] = 0$$
$$-\frac{2}{\ell}\sum_{i=1}^{\ell}y_ix_i - wx_i^2 = 0$$

   Выразим w:

$$-\frac{2}{\ell}\left[\sum_{i=1}^{\ell}y_ix_i - w\sum_{i=1}^{\ell}x_i^2\right] = 0$$
$$\frac{2}{\ell}\sum_{i=1}^{\ell}y_ix_i = w\sum_{i=1}^{\ell}x_i^2$$
$$w = \frac{\sum_{i=1}^{\ell}y_ix_i}{\sum_{i=1}^{\ell}x_i^2}$$

   Это значение является оптимальной оценкой параметра w, которое минимизирует квадратичную функцию потерь в модели линейной регрессии без bias.

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

   Проведем аналогичные преобразования:

Для нахождения минимума $Q(w)$ найдем её производную по w и приравняем её к 0:

$$\frac{\partial Q}{\partial w} = 0$$

$$\frac{\partial}{\partial w}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - wx_i)^2 + \lambda w^2\right] = 0$$

$$\frac{\partial}{\partial w}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i^2 - 2wx_iy_i + w^2x_i^2) + \lambda w^2\right] = 0$$

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

$$-\frac{2}{\ell}\left[\sum_{i=1}^{\ell} y_ix_i - w\sum_{i=1}^{\ell} x_i\right] + 2\lambda w = 0$$

$$w\left[\frac{2}{\ell}\sum_{i=1}^{\ell} x_i + 2\lambda\right] - \frac{2}{\ell}\sum_{i=1}^{\ell} y_ix_i = 0$$

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

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

   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.

   Особенность этой задачи в том, что функция потерь включает модуль $|w|$, который не является дифференцируемым в точке $w = 0$. Поэтому рассмотрим случаи $w > 0$ и $w < 0$ отдельно:

   - Случай, когда $w > 0$:

   $$\frac{\partial Q}{\partial w} = 0$$

   $$-\frac{2}{\ell}\sum_{i=1}^{\ell}(y_ix_i - wx_i^2) + \lambda = 0$$

   $$w = \frac{\sum_{i=1}^{\ell} y_ix_i - \frac{\ell\lambda}{2}}{\sum_{i=1}^{\ell} x_i^2}$$

   - Случай, когда $w < 0$:

   $$\frac{\partial Q}{\partial w} = 0$$

   $$-\frac{2}{\ell}\sum_{i=1}^{\ell}(y_ix_i - wx_i^2) - \lambda = 0$$

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

   - Случай, когда $w = 0$: $Q(w)$ недифференцируема

Итого:

$$w = \begin{cases} \frac{\sum_{i=1}^{\ell} y_i x_i - \frac{\ell\lambda}{2}}{\sum_{i=1}^{\ell} x_i^2}, & \text{если } \sum_{i=1}^{\ell} y_i x_i > \frac{\ell\lambda}{2}, \\ 0, & \text{если } \left|\sum_{i=1}^{\ell} y_i x_i\right| \le \frac{\ell\lambda}{2}, \\ \frac{\sum_{i=1}^{\ell} y_i x_i + \frac{\ell\lambda}{2}}{\sum_{i=1}^{\ell} x_i^2}, & \text{если } \sum_{i=1}^{\ell} y_i x_i < -\frac{\ell\lambda}{2}. \end{cases}$$

Это решение демонстрирует важное свойство LASSO-регрессии: она может давать точно нулевые значения параметров w (в отличие от Ridge-регрессии), что делает её полезной для отбора признаков.
Если $y_i x_i$ (вклад данных) слишком мал по сравнению с регуляризацией λ, то значения соответствующих коэффициентов w_i будут равны 0.

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

$$(a) \lim_{\lambda\to 0} \hat{w}_R, \quad (b) \lim_{\lambda\to\infty} \hat{w}_R, \quad (c) \lim_{\lambda\to 0} \hat{w}_L, \quad (d) \lim_{\lambda\to\infty} \hat{w}_L.$$

(a) При $\lambda \to 0$ мы получаем обычное выражение для линейной регрессии без регуляризации:

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

(b) Когда $\lambda \to \infty$ член $\lambda\ell$ доминирует над $\sum_{i=1}^{\ell} x_i$ и таким образом знаменатель $\to \lambda\ell$

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

При $\lambda \to \infty$: $\frac{1}{\lambda} \to 0$:

$$\lim_{\lambda\to\infty} \frac{\sum_{i=1}^{\ell} y_i x_i}{\sum_{i=1}^{\ell} x_i + \lambda\ell} = 0$$

(c) Найти предел $\lim_{\lambda\to 0} \hat{w}_L$, где:

$$\hat{w}_L = \begin{cases} \frac{\sum_{i=1}^{\ell} y_i x_i - \frac{\ell\lambda}{2}}{\sum_{i=1}^{\ell} x_i^2}, & \text{если } \sum_{i=1}^{\ell} y_i x_i > \frac{\ell\lambda}{2}, \\ 0, & \text{если } \left|\sum_{i=1}^{\ell} y_i x_i\right| \le \frac{\ell\lambda}{2}, \\ \frac{\sum_{i=1}^{\ell} y_i x_i + \frac{\ell\lambda}{2}}{\sum_{i=1}^{\ell} x_i^2}, & \text{если } \sum_{i=1}^{\ell} y_i x_i < -\frac{\ell\lambda}{2}. \end{cases}$$

При $\lambda \to 0$ правая часть неравенства условий $\to 0$, как и регуляризационный член в выражениях для первого и третьего случая, тогда получаем:

$$\lim_{\lambda\to 0} \hat{w}_L = \begin{cases} \frac{\sum_{i=1}^{\ell} y_i x_i}{\sum_{i=1}^{\ell} x_i^2}, & \text{если } \sum_{i=1}^{\ell} y_i x_i \ne 0, \\ 0, & \text{если } \sum_{i=1}^{\ell} y_i x_i = 0. \end{cases}$$

(d) Найти предел $\lim_{\lambda\to\infty} \hat{w}_L$, где:

$$\hat{w}_L = \begin{cases} \frac{\sum_{i=1}^{\ell} y_i x_i - \frac{\ell\lambda}{2}}{\sum_{i=1}^{\ell} x_i^2}, & \text{если } \sum_{i=1}^{\ell} y_i x_i > \frac{\ell\lambda}{2}, \\ 0, & \text{если } \left|\sum_{i=1}^{\ell} y_i x_i\right| \le \frac{\ell\lambda}{2}, \\ \frac{\sum_{i=1}^{\ell} y_i x_i + \frac{\ell\lambda}{2}}{\sum_{i=1}^{\ell} x_i^2}, & \text{если } \sum_{i=1}^{\ell} y_i x_i < -\frac{\ell\lambda}{2}. \end{cases}$$

При $\lambda \to \infty$ правая часть неравенства условий $\to \infty$, соответственно первая и третья строка невозможны, всегда

$$-\infty \leq \left|\sum_{i=1}^{\ell} y_i x_i\right| \leq \infty$$

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

5. Как можно проинтерпретировать гиперпараметр λ?

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

Его можно проинтерпретировать следующим образом:

1. Компромисс между точностью и регуляризацией:
   (a) При $\lambda = 0$ мы минимизируем только среднеквадратичную ошибку, не ограничивая величину w
   (b) При $\lambda \to \infty$ модель будет стремиться установить $w = 0$, полностью игнорируя данные
2. Байесовская интерпретация:
   (a) Ridge-регрессию можно рассматривать как максимизацию апостериорной вероятности при нормальном априорном распределении параметра w. В этом контексте λ связана с дисперсией априорного распределения.
   (b) Большие значения λ соответствуют сильной априорной уверенности, что w должно быть близко к нулю

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.

   Найдем оптимальное w, для этого приравняем производную функционала по w нулю:

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

$$\frac{\partial Q}{\partial w} = 0$$

$$\frac{\partial}{\partial w}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - w)^2 + \lambda w^2\right] = 0$$

$$\frac{\partial}{\partial w}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i^2 - 2wy_i + w^2) + \lambda w^2\right] = 0$$

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

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

$$w_{opt} = \frac{\frac{1}{\ell}\sum_{i=1}^{\ell} y_i}{1+\lambda}$$

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.

   - Получим значения w_{opt} для каждого "эксперимента"
     1. $y_2 = 6$, $y_3 = 10$, тогда $w_{opt} = \frac{8}{1+\lambda}$
     2. $y_1 = 6$, $y_3 = 10$, тогда $w_{opt} = \frac{8}{1+\lambda}$
     3. $y_1 = 6$, $y_2 = 6$, тогда $w_{opt} = \frac{6}{1+\lambda}$
   - Посчитаем ошибки, допущенные на тесте:

     $$Loss = \left(\frac{8}{1+\lambda} - 6\right)^2 + \left(\frac{8}{1+\lambda} - 6\right)^2 + \left(\frac{6}{1+\lambda} - 12\right)^2$$

     Искомая λ_{opt}:

     $$\lambda_{opt} = \arg\min_{\lambda} Loss$$

     Решая уравнение

     $$\frac{\partial Loss}{\partial \lambda} = 0$$

     Получаем численный ответ:

     $$\lambda_{CV} = \frac{2}{39}$$

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

   $$w = \frac{\frac{1}{\ell}\sum_{i=1}^{\ell} y_i}{1+\lambda} = \frac{\frac{1}{3}(6+6+10)}{1+\frac{2}{39}} = 6\frac{40}{41}$$

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

$$w = \frac{\frac{1}{\ell}\sum_{i=1}^{\ell} y_i}{1+\lambda} = \frac{\bar{y}}{1+\lambda}$$

Пусть:

$$\bar{y}_{-k} = \frac{\sum_{i=1}^{\ell} y_i \cdot [i \neq k]}{\ell - 1}$$

Тогда:

$$w_k = \frac{\bar{y}_{-k}}{1+\lambda}$$

$$L_k = (y_k - w_k)^2 = \left(y_k - \frac{\bar{y}_{-k}}{1+\lambda}\right)^2$$

$$L = \frac{1}{\ell}\sum_{k=1}^{\ell} L_k = \frac{1}{\ell}\sum_{k=1}^{\ell}\left(y_k - \frac{\bar{y}_{-k}}{1+\lambda}\right)^2$$

$$\lambda_{CV} = \arg\min_{\lambda} \frac{1}{\ell}\sum_{k=1}^{\ell}\left(y_k - \frac{\bar{y}_{-k}}{1+\lambda}\right)^2$$

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

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

  Различие между гиперпараметрами и параметрами модели заключается в том, что гиперпараметры задаются вручную до начала обучения и определяют характеристики всего процесса обучения, в то время как параметры модели вычисляются в процессе оптимизации функционала на основе данных.
  Таким образом, в данном примере w является параметром, тогда как λ - гиперпараметром.

- Почему коэффициент регуляризации нельзя подбирать по обучающей выборке?

  В процессе обучения модели происходит оптимизации функционала качества (функции потерь), т.к. λ участвует в расчете функции потерь, подобранное на обучающей выборке значение будет скоре-всего слишком маленьким (или даже 0).

  Параметр λ используется в качестве регуляризации для увеличения обобщающей способности модели и предотвращения переобучения. Подбор λ на обучающей выборке противоречит этим целям.

- Как подобрать оптимальное значение для коэффициента регуляризации?

  1. Отложенная выборка
  2. Кросс-валидация
  3. Grid-Search
  4. Байесовская оптимизация

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

  Смысл w_0:
  w_0 отвечает за базовый уровень предсказаний, когда все признаки равны нулю, он позволяет модели учитывать смещение целевой переменной относительно нуля.
  По сути, w_0 моделирует среднее значение целевой переменной

При регуляризации мы "штрафуем"большие значения w_0, стремясь приблизить его к нулю, но если среднее значение целевой переменной сильно отличается от нуля, такое стремление w_0 к нулю будет вредным. Модель будет вынуждена компенсировать это за счет других коэффициентов, что может привести к переобучению

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

  Кросс-валидация - это метод оценки качества модели, при котором:
  1. Данные разбиваются на k частей
  2. Модель обучается k раз на $(k-1)$ части и тестируется на оставшейся
  3. Полученные оценки качества и метрики усредняются

  Плюсы такого подхода в сравнении с использованием отложенной выборки:

  \+ Более надежная оценка качества:
  Каждое наблюдение участвует в тестировании, а результат меньше зависит от конкретного разбиения данных. Таким образом можно получить не только средние оценки, но и оценить их дисперсию.

  \+ Эффективное использование данных:
  Все наблюдения участвуют в обучении, что особенно важно при малых выборках

  \+ Помогает выявлять нестабильность модели:
  Если качество сильно варьируется между фолдами, это может говорить о проблемах с моделью. Можно обнаружить чувствительность к выбросам.

- Почему категориальные признаки нельзя закодировать натуральными числами?

  Потому что при кодировании категориальных признаков натуральными числами мы неявно задаем отношения порядка на значениях этих признаков. Например, если существовал признак "цвет"и мы закодировали "красный": 1, "синий": 2, "зеленый": 3, то теперь для нашей модели красный $<$ зеленый

- Что такое one-hot encoding?
  Это способ кодирования (обычно, категориальных) признаков объектов обучающей выборки, когда вместо значения признака используется вектор из 0 и 1, где только одна 1 отвечает за то, какое значение признака было у этого объекта до кодирования. Например, признак "цвет"имеющий значения "красный"синий"и "зеленый"для красного объекта будет преобразован в вектор $[1, 0, 0]$, а для зеленого в вектор $[0, 0, 1]$.
  Важно помнить, что при использовании one-hot encoding способом, который описан выше мы получаем признак (набор новых признаков количеством N, где N-количество уникальных значений признака), который в сумме для каждого объекта равен 1 и сталкиваемся с проблемой мультиколлинеарности. Для решения этой проблемы, можно, например, кодировать признак набором новых признаков количеством $N-1$.

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

  - Интерпретируемость весов:
    После масштабирования веса модели можно сравнивать между собой и говорить о том, что чем больше по модулю $|w_i|$ (вес, соответствующий i-ому признаку), тем больший вклад этот признак вносит в предсказание. Это позволяет оценить относительную важность признаков.
  - Влияние на метрики расстояния:
    Многие алгоритмы (kNN, K-means, SVM) используют расстояния между объектами. Признаки большего масштаба будут доминировать при расчёте расстояний, а масштабирование обеспечивает равный вклад всех признаков
  - Без масштабирования регуляризация будет непропорционально влиять на разные признаки. Например, признак "возраст"$(0-100)$ и "доход"$(0-1000000)$ будут штрафоваться по-разному.

- Почему L_1-регуляризация производит отбор признаков?

  Из **Задача 1**, п. 2, п.3:
  L1 (Lasso) регуляризация может давать точно нулевые значения параметров w (в отличие от Ridge-регрессии), что делает её полезной для отбора признаков.
  Если $y_i x_i$ (вклад данных) слишком мал по сравнению с регуляризацией λ, то значения соответствующих коэффициентов w_i будут равны 0.

- Почему MSE чувствительно к выбросам?

  Потому что MSE вычисляет разницу между истинным и предсказанным значением и возводит его в квадрат. Если рассматриваемый объект является выбросом, то такая разница будет очень велика по модулю и внесет большой вклад в Loss.