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

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.

Задача 1 (1 балл).

Предположим, что мы хотим предсказывать вероятности принадлежности объекта классу в задаче с двумя классами $0, 1$ (обратите внимание, что не $-1, +1$) с помощью линейного классификатора:

$$Q(w, X^l) = \sum_{i=1}^{l} \mathcal{L}(y_i, \langle w, x_i \rangle) \to \min_w,$$

где $\mathcal{L}(y, z)$ – некоторая гладкая функция потерь. Какую функцию потерь было бы логично использовать в этом случае и почему? Чтобы ответить на этот вопрос, проделайте следующее:

- Запишите логарифм правдоподобия выборки для модели, предсказывающей вероятность $p(x_i)$ принадлежности объекта x_i положительному классу.
- Добейтесь того, чтобы классификатор возвращал числа из отрезка $[0, 1]$, положив

  $$p(x_i) = \sigma\left(\langle w, x_i \rangle\right) = \frac{1}{1 + \exp\left(-\langle w, x_i \rangle\right)}$$

  Сигмоидная функция монотонно невозрастает, поэтому чем больше скалярное призведение, тем большая вероятность положительного класса будет предсказана объекту.
- Подставьте трансформированный ответ линейной модели в логарифм правдоподобия. К какой известной функции потерь мы пришли? (Обычно она записана для обозначений классов $-1, 1$.)

Задача 2 (2 балла). Докажите, что в случае линейно разделимой выборки не существует вектора параметров (весов), который бы максимизировал правдоподобие вероятностной модели логистической регрессии. Предложите, как можно модифицировать вероятностную модель, чтобы оптимум достигался.

Выпишите формулы пересчета значений параметров при оптимизации методом градиентного спуска для обычной модели логистической регрессии и предложенной модификации.

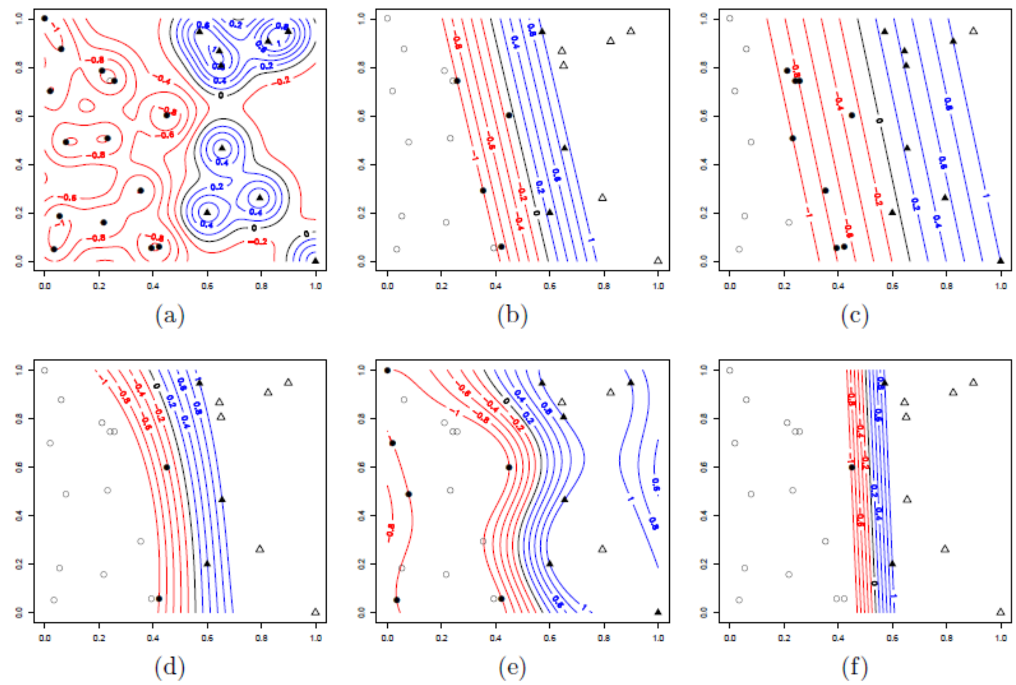

Рис. 1: Иллюстрация к задаче 3.

Задача 3 (1 балл). На одной и той же выборке шесть раз обучили SVM в разных конфигурациях и с различными наборами параметров. Обучение происходило путем решения оптимизационной задачи:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}(1 - M_i(\boldsymbol{w}, b))_+ \to \min_{\boldsymbol{w}, b}, \quad M_i(\boldsymbol{w}, b) = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b).$$

Трижды обучался линейный SVM:

(1) с параметром $C = 10$

(2) с параметром $C = 1$

(3) с параметром $C = 0.1$

Трижды обучался SVM с Гауссовским RBF-ядром $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$:

(4) с параметрами $\gamma = 1$, $C = 3$

(5) с параметрами $\gamma = 10$, $C = 1$

(6) с параметрами $\gamma = 0.1$, $C = 15$

На рис. 1 представлены результаты обучения: разделяющая полоса и линии уровня $f(\boldsymbol{x}) = c$, $c \in \{-1, -0.8, \ldots, 0.8, 1\}$. Соотнесите между собой рисунки и наборы параметров.

Задача 4 (2 балла) Классический линейный метод опорных векторов приводит к задаче квадратичного программирования:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n \to \min_{\boldsymbol{w},\boldsymbol{\xi}}, \quad \text{s.t.} \quad y_n \boldsymbol{w}^T \boldsymbol{x}_n \geq 1 - \xi_n, \quad \xi_n \geq 0 \quad \forall n.$$

Какую функцию потерь оптимизирует линейный метод опорных векторов? Постройте задачу квадратичного программирования аналогичную SVM, но для оптимизации функции потерь:

$$\mathcal{L}(M) = \begin{cases} (M-1)^2, & M < 1 \\ 0, & M \geq 1. \end{cases}$$

Выпишите двойственную к ней задачу.

Задача 5 (2 балла) Процесс обучения линейной регрессии решает следующую оптимизационную задачу:

$$\|Xw - y\|^2 \to \min_w . \tag{1}$$

Её решение можно найти аналитически:

$$w^* = X^\dagger_{\text{left}} y, \quad X^\dagger_{\text{left}} = (X^T X)^{-1} X^T.$$

Здесь $X^\dagger_{\text{left}}$ называется левой псевдообратной матрицей[1]. Практика показывает, что для избежания переобучения можно использовать регуляризацию, которая дополнительно штрафует евклидову норму вектора весов $\|w\|$ в задаче оптимизации 1.

Положим, что система уравнений $Xw = y$ является неопределенной[2], т. е. существует бесконечно много w, которые доставляют точное решение этой системы. Кроме того, будем считать, что матрица X является прямоугольной матрицей полного ранга. Интуиция регуляризации, описанная выше, говорит нам о том, что среди всех точных решений необходимо выбрать такое решение w^*, которое будет иметь наименьшую норму. Докажите, что решением с наименьшей нормой является:

$$w^* = X^\dagger_{\text{right}} y, \quad X^\dagger_{\text{right}} = X^T (X X^T)^{-1}.$$

Матрица $X^\dagger_{\text{right}}$ называется правой псевдообратной.

Подсказка: воспользуйтесь доказательством от противного.

Задача 6 (2 балла). Покажите, что LASSO Тибширани

$$\|Xw - y\|^2 \to \min_w, \quad \text{s.t.} \quad \sum_k |w_k| < \kappa$$

эквивалентен $l1$-регуляризации

$$\|Xw - y\|^2 + \tau \sum_k |w_k| \to \min_w$$

для некоторого коэффициента регуляризации τ.

Дайте геометрическое объяснение тому, что $l1$-регуляризация способствует отбору признаков (занулению компонент вектора весов w), а $l2$-регуляризация таким свойством не обладает.

[1] https://en.wikipedia.org/wiki/Moore-Penrose_pseudoinverse
[2] https://en.wikipedia.org/wiki/Underdetermined_system