Домашнее задание по курсу «Введение в машинное обучение»

Задача 1. Рассмотрим функцию, равную косинусу угла между двумя векторами $x, z \in \mathbb{R}^d$:

$$K(x, z) = \cos(\widehat{x, z}).$$

Покажите, что она является ядром.

Задача 2. Рассмотрим следующую функцию на пространстве вещественных чисел:

$$K(x, z) = \frac{1}{1 + e^{-xz}}.$$

Покажите, что она не является ядром.

Задача 3. Рассмотрим постановку оптимизационной задачи одноклассового метода опорных векторов:

$$\begin{cases} \frac{1}{2}\|w\|^2 + \sum\limits_{i=1}^{\ell} (\xi_i - \nu\rho) \to \min\limits_{w,\xi,\rho}, \quad \xi = (\xi_i)_{i=1}^{\ell}, \\ \langle w, x_i \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i = \overline{1, \ell}, \end{cases} \tag{0.1}$$

где $\nu \in (0, 1]$. Модель одноклассового метода опорных векторов строит решающее правило $a(x) = \operatorname{sign}(\langle w, x \rangle - \rho)$, принимающее значение 1 на области как можно меньшего объёма, содержащей как можно больше объектов выборки. Модель $a(x)$ разделяет выборку от начала координат с максимальным отступом, такого рода модели применяются в задачах детектирования аномалий, когда требуется провести *фильтрацию* данных и отсеять выбросы или выделить наиболее нетипичные объекты относительно имеющейся выборки.

Докажите, что задача (0.1) является задачей выпуклого программирования. Для задачи (0.1) выпишите функцию Лагранжа и по ней постройте двойственную задачу, покажите, что двойственная задача является также задачей выпуклого программирования. Какие объекты в двойственной задаче являются опорными? Обобщите двойственную задачу на случай ядра $K(x_i, x_j) \stackrel{\text{def}}{=} \langle x_i,\ x_j \rangle,\ i, j = \overline{1, \ell}$. Покажите, что гиперпараметр ν является верхней оценкой на долю аномалий в выборке $\{x_i\}_{i=1}^{\ell}$, то есть что $\frac{1}{\ell}\sum\limits_{i=1}^{\ell}[\langle w,\ x_i \rangle < \rho] \leqslant \nu$, где w — оптимальный вектор весов.

Задача 4. Рассмотрим следующую задачу метода опорных векторов (SVM) с двумя классами -1 и $+1$ и $p_i \in [0;1]$, $i = \overline{1, \ell}$:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell} p_i \left(\max\left\{1 - y_i(\langle w, x_i\rangle + b),\ 0\right\}\right)^2 \to \min_{w,b}, \quad C > 0. \tag{0.2}$$

Скаляры p_i обозначают важность объектов. Запишите задачу безусловной оптимизации (0.2) в виде эквивалентной задачи условной оптимизации, введя вспомогательные переменные. Выведите двойственную к этой задаче, выразив через функцию ядра $K(x_i, x_j) \overset{\text{def}}{=} \langle x_i,\ x_j\rangle,\ i, j = \overline{1, \ell}$. Выразите явно решающее правило $a(x) = \langle w, x\rangle + b$, используя ядра и важность объектов.

Задача 5. Рассмотрим нелинейное преобразование ReLU:

$$g(x) = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases}$$

Требуется построить нейронную сеть с одним скрытым слоем из двух нейронов с активацией ReLU и одним выходным нейроном без нелинейного преобразования для классификации выборки XOR:

$$\begin{aligned} x_1 &= (0, 0), \quad y_1 = 0, \\ x_2 &= (0, 1), \quad y_2 = 1, \\ x_3 &= (1, 0), \quad y_3 = 1, \\ x_4 &= (1, 1), \quad y_4 = 0. \end{aligned}$$

Класс выдаваемый сетью определяется с помощью некоторого порога значения на выходном нейроне. Свободных членов в сети нет. Достаточно найти любое подходящее решение.

Задача 6. Критерий информативности для набора объектов R вычисляется на основе того, насколько хорошо их целевые переменные предсказываются константой (при оптимальном выборе этой константы):

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c),$$

где $L(y, c)$ — некоторая функция потерь. Соответственно, чтобы получить вид критерия при конкретной функции потерь, необходимо аналитически найти оптимальное значение константы и подставить его в формулу для $H(R)$.

Выведите критерии информативности для следующих функций потерь:

1. $L(y, c) = (y - c)^2$;
2. $L(y, c) = \sum_{k=1}^{K} (c_k - [y = k])^2$;
3. $L(y, c) = -\sum_{k=1}^{K} [y = k] \log c_k$.

У вас должны получиться дисперсия, критерий Джини и энтропийный критерий соответственно.

Задача 7. Пусть даны выборка X, состоящая из 8 объектов x_i, и классификатор $b(x)$, предсказывающий оценку принадлежности объекта положительному классу. Предсказания $b(x)$ и реальные метки объектов приведены ниже:

$$\begin{aligned}
&b(x_1) = 0.1, \quad y_1 = +1,\\
&b(x_2) = 0.8, \quad y_2 = +1,\\
&b(x_3) = 0.2, \quad y_3 = -1,\\
&b(x_4) = 0.9, \quad y_4 = -1,\\
&b(x_5) = 0.9, \quad y_5 = +1,\\
&b(x_6) = 0.3, \quad y_6 = +1,\\
&b(x_7) = 0.6, \quad y_7 = -1,\\
&b(x_8) = 0.95, \quad y_8 = +1.
\end{aligned}$$

Постройте ROC-кривую и вычислите AUC-ROC для множества классификаторов $a(x;t)$, порожденных $b(x)$, на выборке X. Множество классификаторов параметризовано с помощью t — порога, при котором в случае $b(x) > t$ для объекта x присваивается метка класса $+1$ и -1 иначе.

Задача 8. Пусть дан классификатор $b(x)$, который возвращает оценку принадлежности объекта x положительному классу. Отсортируем все объекты по неубыванию ответа классификатора: $b(x_{(1)}) \leq \dots \leq b(x_{(\ell)})$. Обозначим истинные ответы на этих объектах через $y_{(1)}, \dots, y_{(\ell)}$.

Покажите, что AUC-ROC для данной выборки будет равен вероятности того, что случайно выбранный положительный объект окажется в отсортированном списке не раньше случайно выбранного отрицательного объекта.

Задача 9. Пусть дана некоторая выборка X и классификатор $b(x)$, возвращающий в качестве оценки принадлежности объекта x положительному классу 0 или 1, а не некоторое вещественное число.

1. Постройте ROC-кривую для классификатора $b(x)$ на выборке X.

2. Покажите, что AUC-ROC классификатора $b(x)$ на выборке X может быть выражен через долю правильных ответов и полноту классификатора $a(x;t)$, получающегося при выборе некоторого порога $t \in (0;1)$. Помимо указанных величин в формулу могут входить только величины ℓ_-, ℓ_+, ℓ (количество отрицательных, положительных и общее количество объектов в выборке X соответственно).

3. Покажите, что в случае сбалансированной выборки ($\ell_- = \ell_+$) AUC-ROC классификатора $b(x)$ на выборке X совпадает с долей правильных ответов классификатора при выборе некоторого порога $t \in (0;1)$.

Задача 10. Пусть целевая переменная имеет отрицательные биномиальное распределение с фиксированным параметром r:

$$p(y \mid \mu(x)) = C^y_{y+r-1}\mu(x)^y(1-\mu(x))^r,$$

где $\mu(x) \overset{\text{def}}{=} \sigma(\langle w,\ x\rangle),\ w$ — веса, $\sigma(\cdot)$ — функция сигмоиды. Предложенное распределение позволяет моделировать количественные события на целых неотрицательных числах, например, сроки госпитализации пациентов.

Запишите оптимизационную задачу поиска вектора весов модели w для соответствующей обобщенной линейной модели (для метода максимального прадоподобия), выразите значение параметра $\mu(x)$ через оптимальное значение w^* и найдите матожидание ответов y, обусловленное натуральным параметром θ через дифференцирование $c(\theta)$. Совпадает ли оно с $\mu(x)$ и почему? Как будет выглядеть итоговый алгоритм прогнозирования $y = a(x)$? Напоминаем, обобщённая линейная модель выглядит следующим образом:

$$p(y|\theta,\varphi) = \exp\left(\frac{y\theta - c(\theta)}{\varphi} + h(y,\varphi)\right),$$

где θ, φ — параметры распределения, $c(\theta), h(y,\varphi)$ — параметры–функции.

Задача 11. Рассмотрим метод восстановления плотности распределения с помощью гистограмм. Разобьем все пространство на непересекающиеся области δ_i. Каждому δ_i ставится в соответствие вероятность h_i. По заданной выборке $\{x_i\}_{i=1}^{\ell}$, найдите оптимальные значения h_i с помощью метода максимального правдоподобия.

Задача 12. В задаче классификации имеется частично размеченная выборка. Для восстановления плотностей распределения классов вводится смесь гауссиан с явным указанием гауссианы для размеченных объектов:

$$\sum_{i=1}^{\ell} \log \pi_{y_i} p(x_i \mid y_i, \theta) + \sum_{i=\ell+1}^{\ell+k} \log \sum_{y\in\mathbb{Y}} \pi_y p(x_i \mid y, \theta) \to \max_{\theta,\pi} \quad \sum_{y\in\mathbb{Y}} \pi_y = 1,$$

где $p(x \mid y, \theta) = \mathcal{N}(x \mid \mu_y, \Sigma_y)$. Введите скрытые переменные и запишите EM-алгоритм сначала в общем виде для смеси произвольных вероятностных распределений, затем для смеси гауссиан.

Задача 13. Наблюдается упорядоченная выборка (последовательность, временной ряд) бинарных значений $y = (y_1, \dots, y_n)$, $y_i \in \{0, 1\}$. Элементы выборки генерируются независимо, но известно, что в некоторый момент z меняется вероятность единиц, т.е. для всех $i < z$ $P(y_i = 1) = \theta_1$, а для всех $i \geqslant z$ $P(y_i = 1) = \theta_2$. Выведите EM-алгоритм для восстановления смеси этих двух распределений и определения момента разладки z. Что является скрытой переменной, а что — параметрами распределений? *Задачи о разладке* (change point detection) встречаются при моделировании активности (частоты кликов) пользователей, при обнаружении предаварийных режимов работы оборудования, при выявлении заражённых участков по данным счётчика Гейгера.

Задача 14. В алгоритме K-Means минимизируется среднее внутрикластерное расстояние:

$$\sum_{k=1}^{K}\sum_{i=1}^{\ell}[a(x_i) = k]\|x_i - c_k\|^2.$$

Покажите, что

- при фиксированных центрах c_k каждый объект оптимально приписывать к кластеру, центр которого является ближайшим;
- при фиксированном распределении объектов по кластерам $a(x_i)$ оптимальное значения для центра вычисляется как средняя точка в кластере.