

Список всех вопросов-задач

Баллы в билетах определяют затраты времени на написание ответа (с решением). 1 балл = 6 минут. Самое главное в задачах — написать алгоритм решения, который при реализации на python привел бы к получению правильного ответа (все параметры должны быть однозначно определены!). Если останется время, можно в простых задачах попытаться получить числовой ответ, что будет учтено при оценивании решения.

Кроме перечисленных ниже вопросов-задач будут вопросы (в основном однобальные) на понимание основных понятий машинного обучения и особенностей работы алгоритмов.

Пример 1. Какой из критериев информативности лучше и почему: точный тест Фишера или энтропийный?

Пример 2. В задаче бинарной классификации точки обоих классов сильно перемешаны друг с другом (как в наших лабах 1 и 2). Какой из линейных алгоритмов классификации лучше применять: SVM или логистическую регрессию? Ответ обоснуйте

Пример 3. Что лучше: обычная постановка задачи машинного обучения или вероятностная? Ответ обоснуйте.

Единого правильного ответа на "философские" вопросы не существует. Будет оцениваться логика рассуждений.

1. Для заданного объекта $x = 0.73$ натренированный алгоритм рассчитал следующие вероятности ответа y :

y	-1	0	1
$p(y x)$	0.7	0.2	0.1

Какой ответ y нужно выбрать в качестве наилучшего предсказания, если в случае отличия правильного ответа от неверного предсказания на 1 потери составят 1 тыс руб, а в случае отличия на 2 — составят 14 тыс. руб. Обоснуйте решение математически. (3 балла)

2. Для заданного объекта $x = 0.3$ натренированный алгоритм рассчитал следующие вероятности ответа $y \in [0; 1]$:

$$p(y|x) = \begin{cases} 0.7, & 0 < y < 0.5 \\ 1.3, & 0.5 < y < 1 \end{cases}$$

Какой из двух ответов: $a(x) = 0.5$ или $a(x) = 0.6$, — приводит к меньшему среднему риску, если потери заданы формулой: $\mathcal{L}(a, y) = |a - y|$. Обоснуйте решение математически. (3 балла)

3. На обучающей выборке натренированы три модели M_1 , M_2 , M_3 вероятностного распределения бинарной случайной величины $x \in \{-1, 1\}$:

x	-1	1
$p_1(x)$	0.3	0.7

x	-1	1
$p_2(x)$	0.5	0.5

x	-1	1
$p_3(x)$	0.4	0.6

Для заданной проверочной выборки из трех точек: $\{-1, 1, 1\}$ выясните, какая (или какие) из этих моделей недообучены и какая (какие) переобучены, если их правдоподобия на обучающей выборке равны соответственно: 0.16, 0.15 и 0.19. (3 балла)

4. Обучающая выборка состоит из 5 объектов. Значения признака f_1 суть $\mathfrak{F}_1 = \{0.38, 0.04, 0.79, 0.04, 0.01\}$. Метод обучения построил нам плотность распределения вероятностей $p(x)$ признака f_1 , которая всюду равна нулю кроме маленьких ε -окрестностей точек из \mathfrak{F}_1 , в которых $p(x) = 0.1\varepsilon^{-1}$. Такая модель имеет очень большое правдоподобие для выборки \mathfrak{F}_1 . Причем правдоподобие тем больше, чем меньше ε . Согласитесь ли вы с ней? Если нет, то чем она плоха? Как называются такие модели в машинном обучении? (1 балл)
5. Для заданного объекта $x = 0.09$ натренированный алгоритм рассчитал следующие вероятности ответа $y \in [0; 1]$:

$$p(y|x) = \begin{cases} 1.1, & 0 < y < 0.5 \\ 0.9, & 0.5 < y < 1 \end{cases}$$

Какой из двух ответов: $a(x) = 0.3$ или $a(x) = 0.5$, — приводит к меньшему среднему риску, если потери заданы формулой: $\mathcal{L}(a, y) = (a - y)^2$. Обоснуйте решение математически. (3 балла)

6. Примените метод ближайшего соседа для следующей обучающей выборки и скользящим контролем (leave-one-out) вычислите процент ошибочных классификаций:

x	0.39	-0.14	0.45	-0.35	0.14	0.35	-0.12	0.49	-0.38	-0.47
y	-1	+1	-1	+1	-1	+1	-1	+1	-1	-1

(2 балла)

7. Вычислите выступ для объекта $x = 2$ заданной обучающей выборки для метрического алгоритма классификации с окном Парзена ширины $h = 2$ для треугольного ядра: $K(r) = \max(1 - |r|, 0)$

x	-0.1	0.2	1	1.3	2
y	-1	+1	+1	-1	-1

По рассчитанному выступу определите тип объекта (2 балла)

8. Выполните одну итерацию алгоритма STOLP для заданной обучающей выборки, метода одного ближайшего соседа и начального множества эталонов: $\Omega = \{-0.5, 1.7\}$

x	0.9	1.7	-0.5	-1	0.9
y	-1	+1	-1	+1	+1

(2 балла)

9. Примените метод трех ближайших соседей для следующей обучающей выборки и скользящим контролем (leave-one-out) вычислите процент ошибочных классификаций:

x	-0.32	0.27	-0.24	-0.47	0.15	0.44	0.08	-0.03	0.25	-0.41
y	-1	+1	-1	+1	+1	+1	+1	-1	-1	+1

(3 балла)

10. Вычислите выступ для объекта $x = 0.27$ заданной обучающей выборки для метода 5 ближайших соседей для следующей обучающей выборки:

x	-0.05	-0.48	0.01	-0.34	-0.17	0.27	-0.24	0.35	0.18	-0.08
y	-1	-1	-1	+1	-1	-1	-1	-1	-1	-1

По рассчитанному выступу определите тип объекта (2 балла)

11. Примените метод двух ближайших соседей для следующей обучающей выборки и скользящим контролем (leave-one-out) вычислите процент ошибочных классификаций:

x	0	0.19	0.37	0.26	0.1	0.24	-0.29	0.38	0.25	-0.37
y	+1	-1	-1	+1	+1	+1	+1	-1	+1	-1

(3 балла)

12. Вычислите выступ для объекта $x = -0.06$ заданной обучающей выборки для метода 6 ближайших соседей для следующей обучающей выборки:

x	-0.38	0.14	-0.06	-0.45	-0.49	-0.38	0.2	0.24	-0.1	0.08
y	-1	+1	-1	+1	-1	-1	-1	-1	-1	+1

По рассчитанному выступу определите тип объекта (2 балла)

13. Для заданного объекта x и классов $y \in \{-1, +1\}$ алгоритм машинного обучения рассчитал условные вероятности $p(x|y)$: $p(x|+1) = 0.4$, $p(x|-1) = 0.9$. Какой класс нужно предсказать объекту x , чтобы вероятность ошибки была минимальна, если в обучающей выборке 12% объектов имеют класс $y = +1$ и 88% — класс $y = -1$. Если оба решения имеют одинаковые вероятности ошибки, так и напишите. Обоснуйте решение математически. (2 балла)
14. Банку нужно принять решение о выдаче клиенту кредита величиной 1 млн. руб. на 1 год под 24% годовых. В случае отрицательного решения банк рискует потерять сумму, равную 24% от величины кредита. А в случае положительного решения и невозврата кредита клиентом банк рискует потерять сумму, равную величине кредита. Для заданного клиента x и классов $y \in \{\text{вернет, не вернет}\}$ алгоритм машинного обучения рассчитал условные вероятности $p(y|x)$: $p(\text{вернет}|x) = 0.8$, $p(\text{не вернет}|x) = 0.2$. Помогите банку принять верное решение, минимизирующее величину среднего риска. Если оба решения приводят к одинаковому среднему риску, так и напишите. Обоснуйте решение математически. (2.5 балла)
15. Задана выборка одномерной случайной величины: $X^\ell = \{6, 4, 7, 5, 6, 2, 3, 2, 1, 7\}$. Постройте приближение к плотности распределения, используя параметрический подход и нормальное распределение. (2 балла)
16. В обучающей выборке 55% объектов имеют класс $y = +1$ и 45% — класс $y = -1$. Параметрический подход рассчитал приближенные частные плотности распределения $p(x_1|y)$ и $p(x_2|y)$ признаков x_1 и x_2 для каждого класса $y \in \{-1, +1\}$:

$$p(x_1|+1) \sim N(5, 1), \quad p(x_1|-1) \sim N(9, 1), \quad p(x_2|+1) \sim N(-1, 1), \quad p(x_2|-1) \sim N(-4, 1).$$

Здесь $N(a, \sigma^2)$ - нормальное распределение с матожиданием a и дисперсией σ^2 . Используя наивный байесовский подход, напишите, как вы будете рассчитывать вероятности классов $y = \pm 1$ для объекта $(x_1, x_2) = (2, -4)$. Приближенные вычисления экспонент выполнять не нужно, пусть они останутся в формуле ответа как есть. (2 балла)

17. Задана выборка одномерной случайной величины: $X^\ell = \{8, 4, 5, 1, 3, 2, 7, 8, 6, 2\}$. Постройте приближение к плотности распределения, используя гистограмму с 5 столбцами. (1 балл)
18. Задана выборка двумерной случайной величины:

x_1	-1	0	0	0	0	0	-1	1	-1	0
x_2	1	1	0	0	1	0	1	1	1	1

Постройте приближение к ее плотности распределения, используя параметрический подход и двумерное нормальное распределение. (3 балла)

19. Задана выборка одномерной случайной величины: $X^\ell = \{2, 9, 1, 3, 4, 5, 3, 5, 7, 4\}$. Вычислите приближенно плотность распределения в точке $x = 7$, используя локальную непараметрическую оценку Парзена-Розенблатта с прямоугольным ядром и $h = 1.5$. (2 балла)
20. Задано подмножество объектов выборки, относящихся к одному и тому же заданному классу y и описываемое двумя признаками:

x_1	-1	-1	1	0	0	-1	0	-1	-1	-1
x_2	1	2	1	1	1	2	0	2	0	0

Постройте приближение к совместной плотности распределения признаков, используя одновременно наивный байесовский подход и нормальные распределения. (2.5 балла)

21. Плотность вероятности $p(x_1, x_2|y)$ распределения объекта $\bar{x} = (x_1, x_2)$ в классах $y = \pm 1$ есть

$$p(x_1, x_2|+1) = \frac{1}{2\pi} e^{-\frac{1}{2}((x_1-3)^2+(x_2-1)^2)}, \quad p(x_1, x_2|-1) = \frac{1}{2\pi} e^{-\frac{1}{2}((x_1-2)^2+(x_2-2)^2)}.$$

Вероятности классов и цены за ошибки равны: $p(y = -1) = p(y = +1)$, $\lambda_+ = \lambda_- = 1$. Проверьте выполнение условий теоремы о логистической регрессии (напишите эту проверку!) и, если они выполняются, найдите уравнение линии, разделяющей классы, и выразите вероятности $p(y|x_1, x_2)$ через логистическую функцию. (3 балла)

22. Задана матрица объектов-признаков:

x_1	x_2	y
нет	дождь	-3
нет	без осадков	-8
да	дождь	-1
да	снег	-3
нет	без осадков	0
да	снег	-1
да	дождь	4
да	дождь	1
да	снег	4

Примените бинаризацию (one hot encoding) к признаку x_2 . Напишите получившуюся матрицу объектов-признаков. (2 балла)

23. Заданы значения количественного признака x для различных объектов обучающей выборки: $x = \{18, 28, 65, 38, 72, 30, 74, 71, 41, 60, 66, 21, 84, 90, 61, 37, 80\}$. С помощью дискретизации на 4 группы (выберите их самостоятельно) превратите этот признак в номинальный. Затем выполните бинаризацию (one hot encoding) полученного номинального признака. Укажите в ответе значения номинального признака и запишите итоговую бинарную матрицу объектов-признаков. (3 балла)
24. Нарисуйте график рассеяния значений одного признака: $x = \{0, 16, -1, 13, 15, 0, 14, 5, 16, 1\}$. Про- моделируйте распределение данного признака смесью из двух нормальных распределений. По графику рассеяния определите, какие точки относятся к какой компоненте, и вычислите начальные приближения ко всем параметрам смеси. (3 балла)
25. Можно ли применить логистическую регрессию к следующей матрице из номинальных объектов-признаков:

x_1	x_2	y
красный	плохо	+1
зеленый	хорошо	-1
красный	хорошо	-1
красный	хорошо	-1
синий	хорошо	+1
красный	плохо	-1
синий	хорошо	+1
зеленый	хорошо	-1
красный	плохо	-1

Если да, то каким образом это сделать лучше всего? Если вы будете что-то менять в матрице, напишите как вы будете это делать и укажите новую матрицу. (2 балла)

26. Выполните шаг Expectation ЕМ-алгоритма (расчет скрытых переменных g_{ij}) для следующей выборки одного признака $x = \{4, 8, 5, 4, 11, 4, 5, 2, 8, 1\}$, если в качестве компонент выбраны два равномерных распределения: первое — на отрезке $[-4; 6]$, а второе — на отрезке $[3; 12]$. Веса компонент суть: $w_1 = 0.4$, $w_2 = 0.6$. (3 балла)
27. Является ли выражение $a(x) = \text{sign}(8x_2 - 1 + 7x_1)$ линейной моделью нейрона МакКаллока-Питтса? Если да, определите параметры: вектор w и скаляр w_0 . (0.5 балла)
28. Выполните один шаг метода стохастического градиента, при котором изменится вектор w , для обучающей выборки:

$x^{(1)}$	-1	-1	0	0	1	0	-1	-1	-1	0
$x^{(2)}$	1	0	1	0	1	1	0	1	0	1
$x^{(3)}$	1	1	1	1	1	1	1	1	1	1
y	+1	+1	-1	-1	+1	+1	-1	+1	-1	-1

Начальное значение вектора $w = (1, -1, -1)$, функция потерь — экспоненциальная: $\mathcal{L}(a, y) = e^{-(x_i, w)y_i}$, $x_i = (x_i^{(1)}, x_i^{(2)}, x_i^{(3)})$. Приближенные вычисления экспонент выполнять не нужно, пусть они останутся в формуле ответа как есть. (2 балла)

29. Вычислите выступ линейного классификатора $a(x) = \text{sign}(7x_1 - 8x_2)$ для объекта $x = (0.8, 0.1)$ с правильным ответом $y = -1$. (1 балл)
30. Вычислите AUC алгоритма $a(x)$ классификации на два класса: «-» и «+», если известны следующие результаты его работы:

x	1	2	3	4	5	6	7	8	9	10
$a(x)$	0.42	0.87	0.75	0.04	0.41	0.11	0.83	0.51	0.71	0.64
y	-	+	+	-	-	-	-	+	+	+

Здесь y - правильное значение класса. (3 балла)

31. Постройте ROC-кривую алгоритма $a(x)$ классификации для двух классов «-» и «+», если известны следующие результаты его работы:

x	1	2	3	4	5	6	7	8	9	10
$a(x)$	0.27	0.22	0.15	0.24	0.41	0.53	0.67	0.28	0.79	0.47
y	-	+	+	+	-	-	+	+	-	+

Здесь y - правильное значение класса. (3 балла)

32. Задана обучающая выборка:

x_1	0	2	1	4	6	3	5	8	1	2
x_2	3	2	0	2	2	-1	2	0	0	3
y	-1	-1	-1	+1	+1	+1	+1	+1	-1	-1

Постройте ее график рассеяния. Найдите линейный алгоритм классификации $a(x) = \text{sign}(w_1x_1 + w_2x_2 + w_0)$ (т.е. найдите коэффициенты w_0, w_1, w_2), который возвращает метод опорных векторов и постройте на графике соответствующую ему разделяющую классы полосу и прямую $a(x) = 0$. (2 балла)

33. Каким образом для обучающей выборки из двух признаков f_1, f_2 заставить линейный метод опорных векторов возвращать в качестве разделяющей линии любую кривую **второго порядка**? Напишите математически (формулы), что нужно сделать. (1 балл)
34. Задана обучающая выборка:

x_1	-1	2	1	0	1	-1	-1	1	3	2
x_2	1	-1	0	4	3	5	4	2	0	1
y	-1	+1	-1	+1	+1	+1	+1	+1	-1	-1

Постройте ее график рассеяния. Метод опорных векторов, примененный к данной выборке, вернул разделяющую полосу шириной **2** (не 4 !!!) с центром на прямой $x_2 = 2$. Нарисуйте ее на том же графике. Выясните, какие из объектов являются периферийными, какие — опорными границными и какие — опорными нарушителями. (2 балла)

35. Примените формулу ядерного сглаживания Надарая-Ватсона с прямоугольным ядром ширины $h = 1$ для решения задачи регрессии в точке $x = 0.5$ для обучающей выборки:

x	-0.2	0.9	0.9	1.3	1.9
y	-5	-2	-2	-4	2

(2 балла)

36. Пользуясь **точным Тестом Фишера**, вычислите информативность предиката $x \leq -0.1$ для следующей обучающей выборки:

x	0.6	0	0.5	-0.9	-0.6	0.8	-0.1	0.5	0.4	0.8
y	+1	+1	-1	+1	-1	-1	-1	+1	+1	-1

Количество сочетаний можно не вычислять. (2 балла)

37. Приведите пример закономерности типа решающий пеня (decision stump) для следующей обучающей выборки с одним признаком:

x	0.8	-0.7	-0.1	-0.9	-0.5	0.8	-0.9	-0.5	0.6	0.3
y	-1	+1	-1	+1	-1	+1	+1	+1	-1	+1

Вычислите для нее параметры p и n . (2 балла)

38. Постройте какой-нибудь (не обязательно оптимальный) решающий список (**возвращающий вероятности**) на основе пороговых закономерностей для следующей обучающей выборки с одним признаком:

x	0.2	1	0.3	-0.9	-0.3	0.8	-1	0.2	-0.1	0.8
y	-1	-1	-1	+1	-1	-1	-1	+1	+1	+1

Выполните предсказание **вероятностей классов** для точки $x = 0.3$. (3 балла)

39. Пользуясь **энтропийным определением**, вычислите информативность предиката $x \leq 0.1$ для следующей обучающей выборки:

x	0.1	0.9	0.2	-0.9	0.5	0.8	1	0.4	-0.9	-0.9
y	-1	+1	+1	+1	-1	+1	-1	-1	-1	+1

Приближенные вычисления логарифмов выполнять не нужно, пусть они останутся в формуле ответа как есть. (2 балла)

40. Постройте какое-нибудь (не обязательно оптимальное) решающее дерево (**возвращающее вероятности**) глубины не менее 3 на основе пороговых закономерностей для следующей обучающей выборки с одним признаком:

x	-0.7	0.5	0.1	-0.9	-0.9	0.8	-0.7	-0.7	0.4	-0.2
y	-1	+1	+1	+1	-1	-1	+1	-1	-1	+1

Выполните предсказание **вероятностей классов** для точки $x = 0.16$. (3 балла)

41. Для заданного объекта $x = -0.2$ три разных алгоритма машинного обучения предсказали следующие вероятности $p_i(y|x)$ классов $y = \pm 1$:

$$p_1(+1|x) = 0.2, \quad p_1(-1|x) = 0.8;$$

$$p_2(+1|x) = 0.7, \quad p_2(-1|x) = 0.3;$$

$$p_3(+1|x) = 1, \quad p_3(-1|x) = 0$$

Вычислите, какие вероятности предскажет композиция этих трех алгоритмов, если в качестве корректирующей операции используется взвешенное голосование с весами $\alpha_1 = 0.2$, $\alpha_2 = 0.3$ и $\alpha_3 = 0.5$. (2 балла)

42. Чему равны веса объектов на шаге № T алгоритма AdaBoost? Напишите формулу. (1 балл)
43. Задана обучающая выборка и веса u объектов, рассчитанные на некотором шаге алгоритма AdaBoost:

x	-0.1	0	0.1	0.2	0.5
y	+1	+1	+1	-1	+1
u	0.1	0.2	0.4	0.1	0.2

Пользуясь теоремой Freund-а и Schapire (1996), найдите наилучший алгоритм, оптимизирующий функционал ошибок в AdaBoost, среди всех алгоритмов вида $a(x) = \text{sign}(x - b)$, где $b \in \mathbb{R}$ - произвольный параметр. (3 балла)

44. Постройте PR-кривую алгоритма $a(x)$ классификации для двух классов «-» и «+», если известны следующие результаты его работы:

x	1	2	3	4	5	6	7	8	9	10
$a(x)$	0.14	0.04	0.8	0.92	0.19	0.63	0.74	0.4	0.11	0.74
y	+	+	+	+	-	-	+	+	+	+

Здесь y - правильное значение класса. (3 балла)

45. Вычислите F-меру алгоритма ранжирования документов, если среди 18 найденных им документов 15 релевантны запросу, а 3 — нет. Количество релевантных документов во всей базе данных: 20. (1.5 балла)
46. Вычислите нормированный DCG алгоритма ранжирования документов по убыванию релевантности, если он упорядочил 5 заданных документов следующим образом:

id документа	4	1	2	5	3
истинная релевантность	2	1	2	1	0

(2 балла)

47. Вычислите меру TF-IDF важности слова «экзамен» на форуме мехмата в сообщении, состоящего из 90 слов, если на форуме всего 13 тысяч сообщений, из них 280 сообщений содержат слово «экзамен», а автор в упомянутом сообщении употребил это слово 3 раза. (2 балла)
48. SVD-разложение term-document матрицы для четырех документов D_1, D_2, D_3, D_4 имеет вид:

	D_1	D_2	D_3	D_4
формула	5	4	1	0
вероятность	4	7	0	1
конечно	5	4	4	5
может	4	5	4	4
игра	1	0	5	5

$$=$$

-0.3	0.4	-0.8	-0.3
-0.4	0.6	0.6	0.1
-0.6	-0.2	-0.2	0.8
-0.5	-0.1	0.2	-0.5
-0.3	-0.7	0.1	-0.2

$$\times$$

16	0	0	0
0	8	0	0
0	0	2	0
0	0	0	1

$$\times$$

-0.6	-0.6	-0.4	-0.4
0.3	0.5	-0.6	-0.6
-0.7	0.6	-0.2	0.3
0.3	-0.2	-0.7	0.6

Пользуясь моделью LSA, найдите набор основных тем данной совокупности документов; укажите для каждой темы два доминирующих в ней слова; для каждого документа, укажите с какими коэффициентами в него входят найденные темы. (2 балла)

49. Выполните одну итерацию (Е и М-шаги) алгоритма k-средних для одномерной выборки: $X = \{9, -3, 9, 11, 7, 6, -6, -3, -2, 9\}$, если количество кластеров равно двум и их начальные положения суть $\mu_1 = -5, \mu_2 = 10$. (3 балла)