# 1 Линейная регрессия

**1.1** Рассмотрим задачу линейной регресии

$$Q(w) = (y - Xw)^T (y - Xw) \to \min_w .$$

1. Найдите $dQ(w)$ и $d^2Q(w)$.
2. Выведите формулу для оптимального $w$.
3. Выведите формулу для матрицы-шляпницы (hat-matrix), связывающей вектор фактических $y$ и вектор прогнозов $\hat{y} = H \cdot y$.

**1.2** Рассмотрим задачу регрессии с одним признаком и без константы, $\hat{y}_i = w \cdot x_i$. Решите в явном виде задачи МНК со штрафом:

1. $Q(w) = (y - \hat{y})^T (y - \hat{y}) + \lambda w^2$;
2. $Q(w) = (y - \hat{y})^T (y - \hat{y}) + \lambda |w|$;

**1.3** Храбрая и торопливая исследовательница Мишель хочет решить задачу линейной регрессии по $n$ наблюдениям с вектором $y$ и матрицей признаков $X$. Сначала исследовательница Мишель так торопилась, что совсем забыла последнее наблюдение и оценила задачу с более коротким вектором $y^-$ и матрицей $X^-$, где не хватает последней строки. Затем Мишель взяла правильную матрицу $X$, но неправильный вектор $y^*$, в котором она вместо фактического последнего наблюдения вектора $y$ вписала его прогноз, полученный с помощью регрессии с $y^{-1}$ и $X^-$.

1. Как связаны $\hat{y}_n^-$ и $\hat{y}_n^*$ (прогнозы для последнего наблюдения полученные по модели без последнего наблюдения и модели с неверным последним наблюдением)?
2. Как выглядит вектор, равный разнице $y - y^*$?
3. Какие величины находятся в векторе $H \cdot (y - y^*)$? Чему равна последняя, $n$-ая, компонента этого вектора? Выразите её через $H_{nn}$ и ошибку прогноза последнего наблюдения по модели без последнего наблюдения, $y_n - \hat{y}_n^-$.
4. Как связаны между собой ошибка прогноза $n$-го наблюдения по полной модели, ошибка прогноза $n$-го наблюдения по модели без последнего наблюдения и $H_{nn}$?
5. Как быстро провести кросс-валидацию с выкидыванием одного наблюдения для задачи линейной регрессии?

# 2 Линейные классификаторы

**2.1** Рассмотрим плоскость в $\mathbb{R}^3$, задаваемую уравнением $5x_1 + 6x_2 - 7x_3 + 10 = 0$ и две точки, $A = (2, 1, 4)$ и $B = (4, 0, 4)$.

1. Найдите любой вектор, перпендикулярный плоскости.
2. Правда ли, что отрезок $AB$ пересекает плоскость?
3. Найдите длину отрезка $AB$;
4. Не находя расстояние от точек до плоскости, определите, во сколько раз точка $A$ дальше от плоскости, чем точка $B$;
5. Найдите расстояние от точки $A$ до плоскости.

**2.2** Рассмотрим простейший персептрон с константой, единственным входом $x_1$ и пороговой функцией активации. Подберите веса так, чтобы персептрон реализовывал логическое отрицание (в ответ на 0 выдавал 1, и наоборот).

**2.3** Рассмотрим простейший персептрон с константой, двумя входами $x_1$, $x_2$ и пороговой функцией активации.

Здесь ассистенты нарисуют в tikz картинку, достойную стоять вместо Джоконды в Лувре

1. Подберите веса так, чтобы персептрон реализовывал логическое ИЛИ (OR).
2. Подберите веса так, чтобы персептрон реализовывал логическое И (AND).
3. Докажите, что веса невозможно подобрать так, чтобы персептрон реализовывал исключающее логическое ИЛИ (XOR).
4. Добавьте персептрону вход $x_3 = x_1 \cdot x_2$ и подберите веса так, чтобы персептрон реализовывал XOR.
5. Реализуйте XOR с помощью трёх персептронов с двумя входами и константой. Укажите веса и схему их взаимосвязей.

**2.4** В коробке завалялось три персептрона, у каждого два входа с константой и пороговая функция активации. Реализуйте с их помощью функцию

$$y = \begin{cases} 1, & \text{если } x_2 \geqslant |x_1 - 3| + 2; \\ 0, & \text{иначе} \end{cases}.$$

**2.5** Рассмотрим следующий набор данных:

| $x_i$ | $z_i$ | $y_i$ |
|---|---|---|
| -1 | -1 | 0 |
| 1 | -1 | 0 |
| -1 | 1 | 0 |
| 1 | 1 | 0 |
| 0 | 2 | 1 |
| 2 | 0 | 1 |
| 0 | -2 | 1 |
| -2 | 0 | 1 |

1. Существует ли перспетрон с константой, двумя входами и пороговой функцией активации, способный идеально классифицировать $y_i$ на данной выборке? А хватит ли двух таких персептронов? А может хватит трёх?
2. Введите такое преобразование исходных признаков $h_i = h(x_i, z_i)$, при котором с идеальной классификацией $y_i$ справился бы даже персептрон с одним входом, константой и пороговой функцией активации.

**2.6** Бандерлог из Лога[1] ведёт блог, любит считать логарифмы и оценивать логистические регрессии. С помощью нового алгоритма Бандерлог решил задачу классификации по трём наблюдениям и получил $b_i = \hat{\mathbb{P}}(y_i = 1|x_i)$.

| $y_i$ | $b_i$ |
|---|---|
| 1 | 0.7 |
| -1 | 0.2 |
| -1 | 0.3 |

1. Постройте ROC-кривую.
2. Найдите площадь под ROC-кривой и индекс Джини.
3. Постройте PR-кривую (кривая точность-полнота).

[1] деревня в Кадуйском районе Вологодской области

4. Найдите площадь под PR-кривой.
5. Как по-английски будет «бревно»?

**2.7** Классификатор Бандерлога имеет вид

$$a_i = \begin{cases} 1, & \text{если } b_i > t; \\ -1, & \text{иначе.} \end{cases}$$

Докажите, что площадь под ROC-кривой равна вероятности того, случайно выбранный положительный объект окажется позже случайно выбранного отрицательного объекта, если объекты ранжированы по возрастанию величины $b_i$.

**2.8** Все средние издалека выглядят одинаково, среднее $= f^{-1}(0.5f(x_1) + 0.5f(x_2))$. Например, у среднего арифметического $f(t) = t$, у среднего гармонического $f(t) = 1/t$.

1. Какая $f$ используется для среднего геометрического?

Для измерения качества бинарной классификации Ара использует среднее арифметическое точности и полноты, Гена — среднее геометрическое, а Гарик — среднее гармоническое.

2. У кого будут выходить самые «качественные» и самые «некачественные» прогнозы?

**2.9** Бандерлог начинает все определения со слов «это доля правильных ответов»:

1. accuracy — это доля правильных ответов...
2. точность (precision) — это доля правильных ответов...
3. полнота (recall) — это доля правильных ответов...
4. TPR — это доля правильных ответов...

Закончите определения Бандерлога так, чтобы они были, хм, правильными.

**2.10** Алгоритм бинарной классификации, придуманный Бандерлогом, выдаёт оценки вероятности $b_i = \hat{\mathbb{P}}(y_i = 1|x_i)$. Всего у Бандерлога 10000 наблюдений. Если ранжировать их по возрастанию $b_i$, то окажется что наблюдения с $y_i = 1$ занимают ровно места с 5501 по 5600.

Найдите площадь по ROC-кривой и площадь под PR-кривой.

**2.11** Бандерлог собрал выборку из 900 муравьёв и 100 китов. Переменная $y_i$ равна 1 для китов. Бандерлог хочет, чтобы его алгоритм классификации выдавал для каждого наблюдения число $b_i = f(x_i) \in [0; 1]$, оценку вероятности того, что наблюдение является китом. В качестве признака Бандерлог использует количество глаз, не задумавшись о том, что оно равно двум и для муравьёв, и для китов.

Решите задачу минимизации эмпирической функции риска и найдите все $b_i$ для функций потерь:

1. $L(y_i, b_i) = (y_i - b_i)^2$, если для муравьёв $y_i = 0$;
2. $L(y_i, b_i) = |y_i - b_i|$, если для муравьёв $y_i = 0$;
3. $L(y_i, b_i) = \begin{cases} -\log b_i, & \text{если } y_i = 1 \\ -\log(1 - b_i), & \text{иначе.} \end{cases}$;
4. $L(y_i, b_i) = \begin{cases} 1/b_i, & \text{если } y_i = 1 \\ 1/(1 - b_i), & \text{иначе.} \end{cases}$;

**2.12** Бандерлог утверждает, что открыл новую верхнюю границу для пороговой функции потерь, $\tilde{L}(M_i) = 1 + \frac{1}{\pi} \cdot \arctan(-x_i)$, где $M_i = y_i \cdot \langle w, x_i \rangle$. Прав ли бандерлог?

**2.13** Бандерлог из Лога оценил логистическую регрессию по четырём наблюдениям и одному признаку с константой, получил $b_i = \hat{\mathbb{P}}(y_i = 1|x_i)$, но потерял последнее наблюдение:

| $y_i$ | $b_i$ |
|---|---|
| 1 | 0.7 |
| -1 | 0.2 |
| -1 | 0.3 |
| ? | ? |

1. Выпишите функцию потерь для задачи логистической регрессии.
2. Выпишите условие первого порядка по коэффициенту перед константой.
3. Помогите бандерлогу восстановить пропущенные значения!

**2.14** У Бандерлога три наблюдения, первое наблюдение — кит, остальные — муравьи. Киты кодируются $y_i = 1$, муравьи — $y_i = -1$. На этот раз Бандерлог, чтобы быть уверенным, что $x_i$ различаются, сам лично определил $x_i = i$. После этого Бандерлог оценивает логистическую регрессию с константой.

1. Выпишите эмпирическую функцию риска, которую минимизирует Бандерлог;
2. При каких оценках коэффициентов логистической регрессии эта функция достигает своего минимума?

**2.15** Рассмотрим целевую функцию логистической регрессии с константой

$$Q(w) = \frac{1}{\ell} \sum L(y_i, b_i),$$

где $b_i = 1/(1 + \exp(-\langle w, x_i \rangle))$ и $L(y_i, b_i) = \begin{cases} -\log b_i, & \text{если } y_i = 1 \\ -\log(1 - b_i), & \text{иначе.} \end{cases}$ .

1. Найдите $dQ(w)$ и $d^2Q(w)$;
2. Найдите $dQ(0)$ и $d^2Q(0)$;
3. Выпишите квадратичную аппроксимацию для $Q(w)$ в окрестности $w = 0$;
4. С какой задачей совпадает задача минимизации квадратичной аппроксимации?

**2.16** Винни-Пух знает, что мёд бывает правильный, $honey_i = 1$, и неправильный, $honey_i = 0$. Пчёлы также бывают правильные, $bee_i = 1$, и неправильные, $bee_i = 0$. По 100 своим попыткам добыть мёд Винни-Пух составил таблицу сопряженности:

| | $honey_i = 1$ | $honey_i = 0$ |
|---|---|---|
| $bee_i = 1$ | 12 | 36 |
| $bee_i = 0$ | 32 | 20 |

Винни-Пух использует логистическую регрессию с константой для прогнозирования правильности мёда с помощью правильности пчёл.

1. Какие оценки коэффициентов получит Винни-Пух?
2. Какой прогноз вероятности правильности мёда при встрече с неправильными пчёлами даёт логистическая модель? Как это число можно посчитать без рассчитывания коэффициентов?

**2.17** Винни-Пух оценил логистическую регрессию для прогнозирования правильности мёда от высоты дерева (м) $x_i$ и удалённости от дома (км) $z_i$: $\ln odds_i = 2 + 0.3x_i - 0.5z_i$.

1. Оцените вероятность того, что $y_i = 1$ для $x = 15$, $z = 3.5$.
2. Оцените предельный эффект увеличения $x$ на единицу на вероятность того, что $y_i = 1$ для $x = 15$, $z = 3.5$.
3. При каком значении $x$ предельный эффект увеличения $x$ на единицу в точке $z = 3.5$ будет максимальным?

## 3 Матрицы

**3.1** Известна матрица $X$,

$$X = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ -1 & 0 \end{pmatrix};$$

1. Найдите QR-разложение матрицы $X'X$;
2. Найдите QR-разложение матрицы $XX'$;
3. Найдите спектральное разложение матрицы $X'X$;
4. Найдите спектральное разложение матрицы $XX'$;
5. Найдите сингулярное разложение (SVD) матрицы $X$;

**3.2** Объясните геометрический смысл QR, SVD и спектрального разложений.

**3.3** Бандрелог выполнил SVD-разложение матрицы регрессоров $X$. Помогите Бандерлогу поскорее найти формулу для матрицы-шляпницы $H$, которая проецирует $y$ на пространство столбцов матрицы $X$, $\hat{y} = Hy$.

**3.4** Бандрелог выполнил QR-разложение матрицы регрессоров $X$. Помогите Бандерлогу поскорее найти формулу для матрицы-шляпницы $H$, которая проецирует $y$ на пространство столбцов матрицы $X$, $\hat{y} = Hy$.

## 4 Метод опорных векторов

**4.1** На плоскости имеются точки двух цветов. Красные: $(1,1)$, $(1,-1)$ и синие: $(-1,1)$, $(-1,-1)$.

1. Найдите разделяющую гиперплоскость методом опорных векторов при разных $C$.
2. Укажите опорные вектора.

**4.2** На плоскости имеются точки двух цветов. Красные: $(1,1)$, $(1,-1)$ и синие: $(-1,1)$, $(-1,-1)$ и $(2,0)$.

1. Найдите разделяющую гиперплоскость методом опорных векторов при разных $C$.
2. Укажите опорные вектора.

**4.3** Эконометресса Авдотья решила использовать метод опорных векторов с гауссовским ядром с параметром $\sigma = 1$ и штрафным коэффициентом $C = 1$. Соответственно, она минимизировала целевую функцию

$$\frac{w'w}{2} + C\sum_{i=1}^{n} \xi_i,$$

где разделяющая плоскость задаётся $w'x - w_0 = 0$, а $\xi_i$ — размеры «заступа» за разделяющую полосу.

Затем Автдотья подумала, что неплохо бы выбрать наилучшие $C$ и $\sigma$. Ей лень было использовать кросс-валидацию, поэтому Авдотья минимизировала данную функцию по $C \geqslant 0$ и $\sigma \geqslant 0$. Какие значения она получила?

**4.4** Задан вектор $w = (2,3)$ и число $w_0 = 7$.

1. Нарисуйте прямые $\langle w, x\rangle = w_0$, $\langle w, x\rangle = w_0 + 1$, $\langle w, x\rangle = w_0 - 1$.
2. Найдите ширину полосы между $\langle w, x\rangle = w_0 + 1$ и $\langle w, x\rangle = w_0 - 1$.
3. Найдите расстояние от точки $(5,6)$ до прямой $\langle w, x\rangle = w_0 - 1$.

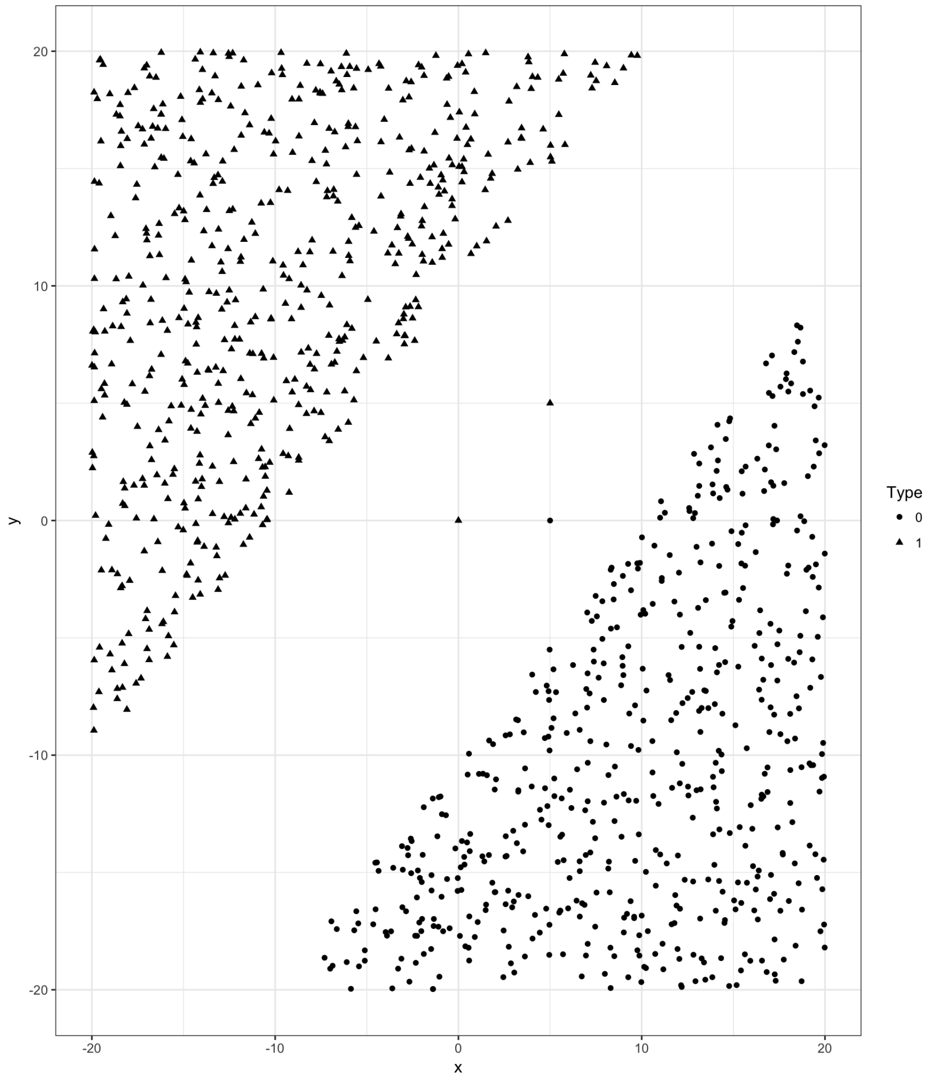
**4.5** Заданы две прямые, $l_0$: $x^{(1)} + 3x^{(2)} = 9$ и $l_1$: $x^{(1)} + 3x^{(2)} = 13$. Найдите подходящий вектор $w$ и число $w_0$ так, чтобы прямая $l_0$ записывалась как $\langle w, x\rangle = w_0 - 1$, а прямая $l_1$ как $\langle w, x\rangle = w_0 + 1$.

**4.6** Даны наблюдения

| $x^{(1)}$ | $x^{(2)}$ | $y$ |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 0 | 0 |
| 0 | 3 | 1 |
| 0 | 4 | 1 |

1. Нарисуйте разделяющую полосу наибольшей ширины.
2. Решите задачу оптимизации
$$\min_{w,w_0} \frac{1}{2}\langle w, w\rangle$$
при ограничении: для $y_i = 1$ выполнено условие $\langle w, x\rangle \geqslant w_0 + 1$, а для $y_i = 0$ выполнено условие $\langle w, x\rangle \leqslant w_0 - 1$.
3. Для точки $x = (x^{(1)}, x^{(2)}) = (1, 1)$ найдите значение $\langle w, x\rangle - w_0$ и постройте прогноз $\hat{y}$.

**4.7** По картинке качественно решите задачу разделения точек:




Целевая функция имеет вид:
$$\min_{w,w_0} \frac{1}{2}w'w + C\sum_{i=1}^{n} \xi_i$$

Уравнение разделяющей поверхности — $w'x = w_0$, уравнения краёв полосы: $w'x = w_0 + 1$ и $w'x = w_0 - 1$. Нарушителями считаются наблюдения, которые попали на нейтральную полосу или на чужую территорию. Здесь $\xi_i = |w| \cdot d_i$, где $d_i$ — длина «заступ» наблюдения за черту «своих».

1. Как пройдёт разделяющая полоса при $C = 1$? Найдите $w$, $w_0$, и величины штрафов $\xi_i$.
2. Как пройдёт разделяющая полоса при $C = +\infty$? Найдите $w$, $w_0$, и величины штрафов $\xi_i$.

**4.8** ююю

# 5 Ядра к бою!

**5.1** Ядерная функция, скалярное произведение в расширяющем пространстве, имеет вид $K(a, b) = \exp(-|a - b|^2)$.

Имеются вектора $a = (1, 1, 1)$ и $b = (1, 2, 0)$.

Найдите длину векторов и косинус угла между ними в исходном и расширяющем пространстве.

**5.2** Рассмотрим два вектора, $v_1 = (1, 1, 2)$ и $v_2 = (1, 1, 1)$. Переход в спрямляющее пространство осуществляется с помощью гауссовской ядерной функции с параметром $\gamma$, $k(v, v') = \exp(-\gamma|v - v'|^2)$.

1. Как от $\gamma$ зависят длины векторов в спрямляющем пространстве?
2. Как от $\gamma$ зависит угол между векторами в спрямляющем пространстве?

**5.3** Имеются три наблюдения $A$, $B$ и $C$:

| | $x$ | $y$ |
|---|---|---|
| $A$ | 1 | -2 |
| $B$ | 2 | 1 |
| $C$ | 3 | 0 |

1. Найдите расстояние $AB$ и косинус угла $ABC$.
2. Найдите расстояние $AB$ и косинус угла $ABC$ в расширенном пространстве с помощью гауссовского ядра с $K(x, x') = \exp(-|x - x'|^2)$.
3. Найдите расстояние $AB$ и косинус угла $ABC$ в расширенном пространстве с помощью полиномиального ядра второй степени.

**5.4** Переход из двумерного пространства в расширяющее задан функцией

$$f : (x_1, x_2) \to (1, x_1, x_2, 3x_1x_2, 2x_1^2, 4x_2^2).$$

Найдите соответствующую ядерную функцию.

**5.5** Ядерная функция имеет вид

$$K(x, y) = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2.$$

Как может выглядеть функция $f : \mathbb{R}^2 \to \mathbb{R}^3$ переводящие исходные векторы в расширенное пространство?

**5.6** Является ли функция $K(x, z)$ ядром?

1. $K(x,z) = \begin{cases} 1, & \text{if } x = z; \\ 0, & \text{otherwise} \end{cases}$;
2. $K(x,z) = \begin{cases} 0, & \text{if } x = z; \\ 1, & \text{otherwise} \end{cases}$;
3. $K(x,z) = \sin(x^T z)$;
4. $K(x,z) = \cos(x^T x)\sin(z^T z)$;

**5.7** Пусть $x$ и $z$ — строки символов, возможно разной длины. Рассмотрим две функции. Функция $K_1(x, z)$ равна единице, если строки $x$ и $z$ совпадают. Функция $K_2(x, z)$ — число совпадающих подстрок. Функция $K_3$ — произведение количеств букв «а» в обеих словах.

1. Найдите $K_1$(«мама», «ам») и $K_2$(«мама», «ам»), $K_3$(«мама», «ам»)
2. Является ли функция $K_1$ ядром?
3. Является ли функция $K_2$ ядром?
4. Является ли функция $K_3$ ядром?

**5.8** На прямой аллее растёт три дуба. Находятся в точках с координатами $x_1 = 1$, $x_2 = 2$ и $x_3 = -3$. Исследователь Винни-Пух проверил и выяснил, что на втором Дубе водятся правильные пчёлы, а на остальных — неправильные.

1. Являются ли пчёлы линейно разделимыми в пространстве исходной аллеи?
2. Помогите Винни-Пуху выписать прямую задачу метода опорных векторов в пространстве исходной аллеи;
3. Помогите Винни-Пуху выписать двойственную задачу метода опорных векторов в пространстве исходной аллеи;
4. Помогите Винни-Пуху выписать двойственную задачу метода опорных векторов в бесконечномерном пространстве с ядерной функцией $K(x, z) = \exp(-(x - z)^2)$; Являются ли точки в нём линейно разделимыми?
5. Помогите Винни-Пуху выписать прямую и двойственную задачу метода опорных векторов в спрямляющем пространстве с ядерной функцией $K(x, z) = (xz + 1)^2$; Являются ли точки в нём линейно разделимыми?

## 6 Двойственные задачи

**6.1** Выпишите двойственную задачу для минимизации $x_1^2+x_2^2+x_3^2$ при ограничении $2x_1+3x_2+5x_3 = 10$.

**6.2** Выпишите двойственную задачу для $x_1 + 2x_2 + 3x_3 \to \max$ при ограничениях $x_1 + x_2 + x_3 \leqslant 10$, $2x_1 + x_2 + x_3 \leqslant 10$, все $x_i \geqslant 0$.

**6.3** Выпишите двойственную задачу для максимизации $1/x_1 + 2/x_2$ при ограничении $2x_1 + 3x_2 = 10$ и $x_1 \in [1; 10]$, $x_2 \in [2; 6]$.

**6.4** Выпишите двойственную задачу для минимизации $f(x) = \frac{1}{2}x'Hx + g'x$ при ограничении $A'x = b$.

**6.5** Выпишите двойственную задачу для минимизации $f(x) = \frac{1}{2}x'Hx + g'x$ при ограничении $A'x \leqslant b$.

**6.6** Выпишите прямую и двойственную задачу для метода опорных векторов в исходном пространстве.

**6.7** Выпишите прямую и двойственную задачу для метода опорных векторов в спрямляющем пространстве с использованием ядра $K(.,.)$.

# 7 Метод главных компонент

**7.1** Найдите прямую, у которой сумма квадратов расстояний до точек $(0,0)$, $(1,1)$, $(2,1)$ будет минимальной. Чему равна при этом доля объяснённого разброса точек?

**7.2** Есть две переменных, $x = (1,0,0,3)'$, $z = (3,2,0,3)'$. Найдите первую и вторую главные компоненты.

**7.3** Известна матрица выборочных ковариаций трёх переменных. Для удобства будем считать, что переменные уже центрированы.

$$\begin{pmatrix} 4 & 1 & -1 \\ 1 & 5 & 0 \\ -1 & 0 & 9 \end{pmatrix}$$

1. Выразите первую и вторую главные компоненты через три исходных переменных.
2. Выразите первую и вторую главные компоненты, через три исходных переменных, если перед методом главных компонент переменные необходимо стандартизировать.

**7.4** Пионеры, Крокодил Гена и Чебурашка собирали металлолом несколько дней подряд. В распоряжение иностранной шпионки, гражданки Шапокляк, попали ежедневные данные по количеству собранного металлолома: вектор $g$ — для Крокодила Гены, вектор $h$ — для Чебурашки и вектор $x$ — для Пионеров. Гена и Чебурашка собирали вместе, поэтому выборочная корреляция $\mathrm{sCorr}(g,h) = -0.9$. Гена и Чебурашка собирали независимо от Пионеров, поэтому выборочные корреляции $\mathrm{sCorr}(g,x) = 0$, $\mathrm{sCorr}(h,x) = 0$. Если регрессоры $g$, $h$ и $x$ центрировать и нормировать, то получится матрица $\tilde{X}$.

1. Найдите параметр обусловленности матрицы $(\tilde{X}'\tilde{X})$.
2. Вычислите одну или две главные компоненты (выразите их через вектор-столбцы матрицы. $\tilde{X}$), объясняющие не менее 70% общей выборочной дисперсии регрессоров.
3. Шпионка Шапокляк пытается смоделировать ежедневный выпуск танков, $y$. Выразите оценки коэффициентов регрессии $y = \beta_1 + \beta_2 g + \beta_3 h + \beta_4 x + \varepsilon$ через оценки коэффициентов регрессии на главные компоненты, объясняющие не менее 70% общей выборочной дисперсии.

## 8 На природу! В лес! К деревьям!

**8.1** Для случайных величин $X$ и $Y$ найдите индекс Джини и энтропию

| $x$ | 0 | 1 |
|---|---|---|
| $\mathbb{P}(X = x)$ | 0.2 | 0.8 |

,

| $y$ | 0 | 1 | 5 |
|---|---|---|---|
| $\mathbb{P}(Y = y)$ | 0.2 | 0.3 | 0.5 |

**8.2** Найдите энтропию $X$, спутанность (perplexity) $X$, индекс Джини $X$, если

1. величина $X$ равновероятно принимает значения 1, 7 и 9;
2. величина $X$ равновероятно принимает $k \geqslant 2$ значений;
3. величина $X$ равномерно распределена на отрезке $[0; a]$;
4. величина $X$ нормальна $\mathcal{N}(\mu; \sigma^2)$;

**8.3** У Васи была дискретная случайная величина $X$, принимавшая натуральные значения. Вася решил изменить закон распределения величины $X$. Он увеличил количество возможных значений величины $X$ в два раза, разделив каждое событие $X = k$ на два равновероятных подсобытия: $X = k - 0.1$ и $X = k + 0.1$. Как при этом изменились энтропия, спутанность (perplexity) и индекс Джини?

**8.4** Случайная величина $X$ принимает значение 1 с вероятностью $p$ и значение 0 с вероятностью $1 - p$.

1. Постройте график зависимости индекса Джини и энтропии от $p$.
2. Являются ли функции монотонными? выпуклыми?
3. При каком $p$ энтропия и индекс Джини будут максимальны?

**8.5** Шаман Ыуыуыуыыыыы по прошлым наблюдениям знает, большая охота на мамонта оказывается удачной с вероятностью 0.3. Если племя ждёт от Ыуыуыуыыыыы прогноз охоты, то Ыуыуыуыыыыы поплясав вокруг костра (10 минут) и постуча бубном (16 раз) прогнозирует удачную охоту с вероятностью 0.3 и неудачную с вероятностью 0.7.

1. Какова вероятность того, что Ыуыуыуыыыыы ошибётся?
2. Чему равен индекс Джини для случайной величины равной удаче с вероятностью 0.3 и неудаче с вероятностью 0.7?

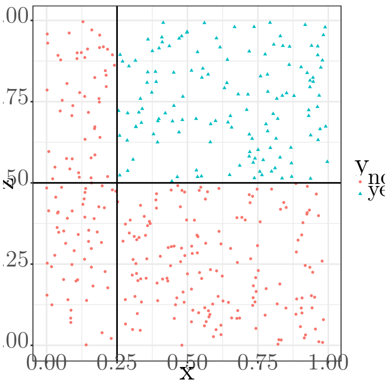
**8.6** Шаман Ыуыуыуыыыыы заметил по прошлым данным, что в дождливые дни большая охота на мамонта удачна с вероятностью 0.7, а в сухие — с вероятностью 0.1. Поэтому в дождливый день Ыуыуыуыыыыы предскажет удачу с вероятностью 0.7, а в сухой — с вероятностью 0.1. Дождливых дней — 20%.

1. Какова вероятность того, что Ыуыуыуыыыыы ошибётся?
2. Чему равен индекс Джини выборки разделённой на две части: в части А шесть бананов и 14 апельсинов, а в части В — восемь бананов и 72 апельсина?

**8.7** Постройте регрессионное дерево для набора данных:

| $y_i$ | $x_i$ |
|---|---|
| 5 | 0 |
| 6 | 1 |
| 4 | 2 |
| 100 | 3 |

Критерий деления узла на два — минимизация $RSS$. Дерево строится до трёх терминальных узлов.




**8.8** Постройте регрессионное дерево для набора данных:

| $y_i$ | $x_i$ |
|---|---|
| 100 | 1 |
| 102 | 2 |
| 103 | 3 |
| 50 | 4 |
| 55 | 5 |
| 61 | 6 |
| 70 | 7 |

Критерий деления узла на два — минимизация $RSS$. Узлы делятся до тех пор, пока в узле остаётся больше двух наблюдений.

**8.9** Дон-Жуан предпочитает брюнеток. Перед Новым Годом он посчитал, что в записной книжке у него 20 блондинок, 40 брюнеток, две рыжих и восемь шатенок. С Нового Года Дон-Жуан решил перенести все сведения в две записные книжки, в одну — брюнеток, во вторую — остальных.

Как изменились индекс Джини и энтропия в результате такого разбиения?

**8.10** Машка пять дней подряд гадала на ромашке, а затем выкладывала очередную фотку «Машка с ромашкой» в инстаграмчик. Результат гадания — переменная $y_i$, количество лайков у фотки — переменная $x_i$. Постройте классификационное дерево.

| $y_i$ | $x_i$ |
|---|---|
| плюнет | 10 |
| поцелует | 11 |
| поцелует | 12 |
| к сердцу прижмёт | 13 |
| к сердцу прижмёт | 14 |

Дерево строится до идеальной классификации. Критерий деления узла на два — максимальное падение индекса Джини.

**8.11** У Винни-Пуха есть 100 песенок (кричалок, вопелок, пыхтелок и сопелок). Каждый день он выбирает и поёт одну из них равновероятно наугад. Одну и ту же песенку он может петь несколько раз. Сколько в среднем песенок оказываются неспетыми за 100 дней?

**8.12** По данной диаграмме рассеяния постройте классификационное дерево для зависимой переменной $y$:

Дерево необходимо построить до идеальной классификации, в качестве критерия деления узла на два используйте минимизацию индекса Джини.

**8.13** Рассмотрим табличку:

| $y_i$ | $x_i$ | $z_i$ |
|---|---|---|
| $y_1$ | 1 | 2 |
| $y_2$ | 1 | 2 |
| $y_3$ | 2 | 2 |
| $y_4$ | 2 | 1 |
| $y_5$ | 2 | 1 |
| $y_6$ | 2 | 1 |
| $y_7$ | 2 | 1 |

Будем называть деревья разными, если они выдают разные прогнозы на обучающей выборке. Сколько существует разных классификационных деревьев для данного набора данных?

**8.14** Исследовательница Мишель строит классификационное дерево для бинарной переменной $y_i$. Может ли при разбиении узла на два расти индекс Джини? Энтропия?

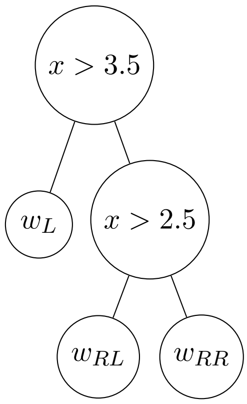
**8.15** Приведите примеры наборов данных, для которых индекс Джини равен 0, 0.5 и 0.999.

**8.16** Рассмотрим задачу построения классификационного дерева для бинарной переменной $y_i$. Приведите пример такого набора данных, что никакое разбиения стартового узла на два не снижает индекс Джини, однако двух разбиений достаточно, чтобы снизить индекс Джини до нуля.

**8.17** Пятачок собрал данные о визитах Винни-Пуха в гости к Кролику. Здесь $x_i$ — количество съеденого мёда в горшках, а $y_i$ — бинарная переменная, отражающая застревание Винни-Пуха при выходе.

Для построения предиктивной модели Пятачок собирается использовать дерево с заданной структурой:

| $y_i$ | $x_i$ |
|---|---|
| 0 | 1 |
| 1 | 4 |
| 1 | 2 |
| 0 | 3 |
| 1 | 3 |
| 0 | 1 |




Пятачок использует квадратичную аппроксимацию для логистической функции потерь:

$$Obj(w) = \sum_{i=1}^{n} \left( loss(y_i, 0) + loss'_w(y_i, 0)(w_i - 0) + \frac{1}{2} loss''_{ww}(y_i, 0)(w_i - 0)^2 \right) + \frac{1}{2}\lambda |w|^2.$$

Помогите Очень Маленькому Существу подобрать оптимальные веса $(w_i)$ при $\lambda = 1$.

**8.18** Нарисовано дерево: деление 1, справа от первого деления — деление 2. Веса равны $w_L$, $w_{RL}$, $w_{LL}$. Дана выборка.

1. Выпишите в явном виде функцию правдоподобия и логистическую функцию потерь.
2. Оцените $w$ методом максимального правдоподобия.
3. Тут другую функцию потерь написать!
4. Разложите функцию потерь в окрестности $w = (0, 0, 1)$ в ряд Тейлора до второго члена и примерно оцените $w$.