1. Рассмотрим задачу восстановления регрессии при квадратичной функции потерь. Доказать, что если $f^*(x) = \operatorname*{argmin}_c \mathsf{E}\left((Y-c)^2 \,|\, x\right)$, то $f^*(x) = \mathsf{E}(Y \,|\, x)$ (*регрессионная функция*). Чему тогда равен средний риск $R(f^*)$?

2. Рассмотрим задачу восстановления регрессии с функцией потерь $L(y' \,|\, y) = |y' - y|$. Доказать, что минимум среднему риску доставляет при этом условная медиана $f(x) = \operatorname{median}(Y \,|\, x)$.

3. Рассмотрим задачу восстановления регрессии, в которой $\mathbf{y}$ распределен согласно нормальному закону $N(\mathbf{X},\, \sigma^2\mathbf{I})$, а β имеет априорное распределение $N(0, \tau\mathbf{I})$. Найти апостериорное распределение для β. Доказать, что β^{ridge} есть его математическое ожидание. Найти связь между параметром регуляризации λ и дисперсиями τ, σ^2.

4. Показать, что процедура гребневой регрессии эквивалентна обычному методу наименьших квадратов, примененному к расширенным данным: к центрированной матрице $\mathbf{X}$ дописывается матрица $\sqrt{\lambda}\mathbf{I}$, к вектору $\mathbf{y}$ приписывается p нулей.

5. Показать, как (и объяснить почему) задачу квадратичного программирования в методе лассо

$$\widehat{\beta}^{\text{lasso}} = \operatorname*{argmin}_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2 \right\},$$

при условии

$$\sum_{j=1}^{d} |\beta_j| \le s$$

можно свести к задаче квадратичного программирования с $2d+1$ неизвестными и $2d+1$ линейными ограничениями.

6. Пусть N точек распределены случайно равномерно в единичной d-мерной гиперсфере. Доказать, что медианное расстояние от центра сферы до ближайшей точки равно

$$\rho(d, N) = \sqrt[d]{1 - \sqrt[N]{1/2}}.$$

Найти предел $\rho(d, N)$ при $N \to \infty$, $d \sim N$. Какой вывод из этого можно сделать применительно к методу ближайшего соседа при больших d?

7. *Bias-variance trade-off.* Рассмотрим задачу восстановления зависимости $Y = f^*(X)$, где X — случайная величина, а f^* — неизвестная *детерминированная* функция. Пусть $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ суть независимые реализации величины X. В качестве модельной зависимости возьмем функцию $f(x, D)$, где $D = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right\}$. Разложить $\mathsf{E}_D\big(f(x, D) - f^*(x)\big)^2$ в сумму квадрата математического ожидания смещения (bias) и дисперсии (variance).

8. *Метод использования линейной регрессии в задаче классификации* заключается в следующем. Сопоставим каждому классу k вектор $(y_1, y_2, \ldots, y_K)$, в котором $y_k = 1$, а $y_i = 0$ при $i \neq k$. Собрав вместе индикаторные векторы объектов обучающей выборки, получим матрицу $\mathbf{Y}$ размера $N \times K$. Пусть $\mathbf{X}$ — матрица размера $N \times (d+1)$, первый столбец которой состоит из единиц, а последующие представляют собой векторы из обучающей выборки. Применяя метод наименьших квадратов одновременно к каждому столбцу матрицы $\mathbf{Y}$, получаем значения

$$\widehat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Для каждого столбца $\mathbf{y}_k$ матрицы $\mathbf{Y}$ получим свой столбец коэффициентов $\widehat{\beta}_k$. Соберем их в матрицу $\widehat{\mathbf{B}}$ размера $(d+1) \times K$. Имеем

$$\widehat{\mathbf{B}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Объект x будем классифицировать согласно следующему правилу: Вычислим вектор-строку длины K

$$g(x) = (1, x)\,\widehat{\mathbf{B}}.$$

Отнесем x к классу

$$f(x) = \operatorname*{argmax}_k g_k(x).$$

Доказать, что

$$\sum_{k=1}^{K} g_k(x) = 1.$$

Доказать, что в случае $K = 2$ данный метод эквивалентен решению одной задачи восстановления регрессии. Какой?

9. Задача Фишера сводится к максимизации отношения Рэлея

$$\max_a \frac{a^\top\mathbf{B}a}{a^\top\mathbf{W}a}.$$

Показать, как эта задача сводится к обобщенной задаче на собственные значения

$$\mathbf{B}a = \lambda\mathbf{W}.$$

10. Показать, что оптимальная гиперплоскость, разделяющая два множества, является плоскостью, проходящей через середину отрезка, соединяющего пару ближайших точек из выпуклой оболочки каждого из классов, и перпендикулярно ему. Указание: рассмотреть задачу, двойственную к задаче определения оптимальной гиперплоскости.

11. Показать, что в алгоритме *SVM* задача

$$\min_{\beta,\,\beta_0,\,\xi_i} \frac{1}{2}\|\beta\|^2 + \gamma\sum_{i=1}^{N}\xi_i,$$

при ограничениях

$$y_i(x_i^\top \beta + \beta_0) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (i = 1, 2, \ldots, N),$$

эквивалентна задаче

$$\min_{\beta,\, \beta_0} \sum_{i=1}^{N} \Big[1 - y_i \left(x_i^\top \beta + \beta_0\right)\Big]_+ + \alpha\|\beta\|^2,$$

где $[\cdot]_+$ означает положительную часть, и $\alpha = 1/(2\gamma)$.

12. *SVM и задача восстановления регрессии.* Для восстановления β, β_0 в модели $f(x) = x^\top \beta + \beta_0$. рассмотрим задачу минимизации функции

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V\left(y_i - f(x_i)\right) + \frac{\alpha}{2}\|\beta\|^2,$$

где

$$V(t) = V_\varepsilon(t) = \begin{cases} 0, & \text{если } |t| < \varepsilon, \\ |t| - \varepsilon & \text{в противном случае.} \end{cases}$$

Доказать, что решение $\widehat{\beta}$, $\widehat{\beta}_0$, минимизирующее функцию $H(\beta, \beta_0)$, можно представить в виде

$$\widehat{\beta} = \sum_{i=1}^{N} (\widehat{\alpha}_i^* - \widehat{\alpha}_i) x_i, \quad \widehat{f}(x) = \sum_{i=1}^{N} (\widehat{\alpha}_i^* - \widehat{\alpha}_i) \langle x, x_i \rangle + \beta_0,$$

где $\widehat{\alpha}_i$ и $\widehat{\alpha}_i^*$ являются решением следующей задачи квадратичного программирования:

$$\min_{\alpha_i,\, \alpha_i^*} \left(\varepsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) + \frac{1}{2} \sum_{i,j=1}^{N} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \langle x_i, x_j \rangle \right)$$

при ограничениях

$$0 \leq \alpha_i, \alpha_i^* \leq \frac{1}{\lambda}, \quad \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) = 0, \quad \alpha_i \alpha_i^* = 0.$$

13. Пусть $Z_1, Z_2, \ldots, Z_N$ — независимые одинаково распределенные случайные величины.

$$\mathsf{Pr}(Z_i = 1) = \theta, \qquad \mathsf{Pr}(Z_i = 0) = 1 - \theta$$

(схема Бернулли). Доказать, что

$$\mathsf{Pr}(|\widehat{\theta} - \theta| > \gamma) \leq 2e^{-2\gamma^2 N},$$

где

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N} Z_i.$$

14. Доказать, что если $\mathcal{F}$ конечно, то $\mathrm{VC}\,\mathcal{F} \le \log_2 |\mathcal{F}|$. Для каждого d построить пример $\mathcal{F}$, на котором эта оценка достигается.

15. Доказать, что при $h \le N$

$$\binom{N}{0} + \binom{N}{1} + \ldots + \binom{N}{h} < \left(\frac{eN}{h}\right)^h.$$

16. Пусть $\mathcal{X}$ — подмножество в $\mathbf{R}^d$, а $\mathcal{F}$ — некоторое множество функций, отображающих $\mathcal{X}$ в $\{0,1\}$. Введем класс $\mathcal{F}' = \{f \vee g :\ f, g \in \mathcal{F}\}$, состоящий из дизъюнкций каждой пары функций в $\mathcal{F}$. Доказать, что для функции роста класса $\mathcal{F}'$ справедливо неравенство $\Delta(\mathcal{F}', N) \le \Delta(\mathcal{F}, N)^2$. Воспользовавшись леммой Зауэра, доказать, что $\mathrm{VC}\,\mathcal{F}' \le 10\,\mathrm{VC}\,\mathcal{F}$. Что изменится, если вместо лизънкций рассмотреть (а) все конъюнкции, (б) суммы по модулю 2?

17. Функцию $f :\ \mathbf{R}^d \to \{0,1\}$ назовем *ящиком*, если существуют вещественые числа $a_1, a_2, \ldots, a_d$, $b_1, b_2, \ldots, b_d$, такие, что $f(x) = 1$ тогда и только тогда, когда $a_i \le x \le b_i$ $(i = 1, 2, \ldots, d)$. Найти функцию роста и размерность Вапника–Червоненкиса для класса всех ящиков. Проиллюстрировать на этом примере лемму Зауэра.

18. Пусть T_h — множество всех функций $f :\ \mathbf{R}^d \to \{0,1\}$, вычисляемых бинарными деревьями решений, высоты не выше h. Найти функцию роста и размерность Вапника–Червоненкиса для класса T_h. Проиллюстрировать на этом примере лемму Зауэра.

19. Пусть H_d — множество всех булевых функций $f :\ \{0,1\}^d \to \{0,1\}$, представимых ДНФ, в которых каждая элементарная конъюнкция представляет собой одиночный символ, обозначающий переменную (без отрицания). Найти функцию роста и размерность Вапника–Червоненкиса для класса H_d. Проиллюстрировать на этом примере лемму Зауэра.

20. Функцию $f :\ \mathbf{R}^2 \to \{0,1\}$ назовем полигоном (точнее: k-вершинным полигоном), если найдется выпуклый k-угольник M, такой, что $f(x) = 1$ тогда и только тогда, когда x принадлежит M. Пусть P — множество всех полигонов, а P_k — множество всех k-вершинных полигонов. Найти $\mathrm{VC}\,P$ и $\mathrm{VC}\,P_k$.

21. Привести пример бесконечного класса $\mathcal{F}$, для которого $\mathrm{VC}\,\mathcal{F} = 1$.

22. *Размерность Вапника–Червонекиса для задачи восстановления регрессии.* Пусть $\mathcal{F}$ — некоторый класс функций $f:\ \mathcal{X} \to \mathcal{Y}$. Размерностью Вапника–Червоненкиса для класса $\mathcal{F}$ называется $\mathrm{VC}\,\mathcal{F}'$, где

$$\mathcal{F}' = \left\{ I\big(f(x) - y\big) :\ f \in \mathcal{F},\, y \in \mathcal{Y} \right\}.$$

Найти размерность Вапника–Червоненкиса для класса $\{\sin \alpha x :\ \alpha \in \mathbf{R}\}$.

23. *Может ли использование коррелированных переменных улучшить качество предсказания?* Рассмотрим задачу классификации с двумя классами. Пусть пространство признаков двумерное. Объекты каждого класса имеют нормальное распределение с математическим ожиданием $(-1, -1)$ и $(1, 1)$ соответственно и единичной матрицей ковариации каждый. Априорные вероятности классов равны $\frac{1}{2}$.

1. Вычислить коэффициент корреляции для переменных x_1, x_2.
2. Найти байесов классификатор и вычислить байесову ошибку для усеченной задачи, рассматривая только одну переменную x_1.
3. Найти байесов классификатор и вычислить байесову ошибку для исходной задачи.
4. Приводит ли использование второй переменной к уменьшению ошибки?

24. Рассмотрим задачу классификации с двумя классами 0 и 1. Пусть пространство признаков двумерное. Объекты каждого класса имеют нормальное распределение с математическим ожиданием $(0, 0)$ и $(1, 1)$ соответственно и матрицей ковариации

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Априорные вероятности классов равны $\mathsf{Pr}\,\{Y = 0\} = \frac{1}{3}$ и $\mathsf{Pr}\,\{Y = 1\} = \frac{2}{3}$.

1. Найти уравнение разделяющей поверхности байесова классификатора.
2. Найти собственное разложение матрицы $\mathbf{\Sigma}$.
3. Перейти к новым координатам, оси которых совпадают с собственными векторами матрицы $\mathbf{\Sigma}$.
4. Выписать уравнение разделяющей поверхности байесова классификатора в новых координатах.

25. *Влияние шума на качество предсказания.* Пусть пространство признаков одномерное и обучающая выборка состоит из двух объектов $x^{(0)} = 0$, $x^{(1)} = 1$. Добавим к объектам шумовой признак, распределенный равномерно

на отрезке $[-1, 1]$. Какова вероятность, что объект $x = (0.3, 0)$ окажется ближе (по евклидову расстоянию) к объекту $x^{(1)}$, чем к $x^{(0)}$?

26. Выпуклым полиэдром (или выпуклым многогранником) в пространстве $\mathbf{R}^d$ называется пересечение конечного числа полупространств, т. е. множество решений некоторой системы линейных неравенств:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1d}x_d \leq b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2d}x_d \leq b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{md}x_d \leq b_m, \end{cases}$$

Доказать, что область, в которой все точки имеют одинаковых k ближайших соседей (для евклидова расстояния) есть полиэдр.

27. В дереве решений (для некоторой задачи классфикации с K классами) рассмотрим вершину m и соответствующий ящик R_m. Обозначим p_{mk} долю объектов класса k в этом ящике ($k = 1, 2, \ldots, K$). Рассмотрим классификатор, который для объектов, попавших в ящик R_m выбирает класс случайно, причем класс k выбирается с вероятностью, равной p_{mk}. Доказать, что математическое ожидание частоты ошибок этого классификатора на объектах обучающей выборки, попавших в R_m, равно индексу Джини.

28. Пусть при построении дерева решений в задаче классификации с двумя классами в текущую вершину попало 400 объектов из первого класса и столько же из второго. Пусть необходимо сделать выбор между разбиением на две ветви $(300, 100)$ и $(100, 300)$ и разбиением на две ветви $(200, 400)$, $(200, 0)$. Какое из этих разбиений кажется предпочтительнее (объясните)? Какое разбиение выберет критерий на основе минимизации ошибки, энтропийный критерий и критерий Джини? Приведите свой пример, когда все три критерия дают разные разбиения.

29. Дана обучающая выборка:

x_1	-1	-1	1	1
x_2	-1	1	-1	1
y	1	-1	-1	1

Подобрать ядро и указать соответствующий SVM-классификатор, для которого ошибка на обучающей выборке равна 0.

30. *Сингулярное разложение.* Пусть $u_1, u_2, \ldots, u_d$ — собственные векторы матрицы $\mathbf{X}^\top\mathbf{X}$, а $\lambda_1, \lambda_2, \ldots, \lambda_d$ — соответствующие им собственные значения. Выразите ненулевые собственные векторы и собственные значения матрицы $\mathbf{X}\mathbf{X}^\top$ через $u_1, u_2, \ldots, u_d$, $\lambda_1, \lambda_2, \ldots, \lambda_d$. Выразите собственные векторы и собственные значения матрицы $\mathbf{X}^\top\mathbf{X}$ через собственные векторы и собственные значения матрицы $\mathbf{X}\mathbf{X}^\top$. Какой вывод можно сделать относительно алгоритма вычисления сингулярного разложения матрицы $\mathbf{X}$?