Н. Ю. Золотых

Задачи по машинному обучению

Версия: 15 сентября 2020 г.

Нижний Новгород
2013, 2018, 2019, 2020

1 Матричное дифференцирование

1. Пусть $g: \mathbf{R}^n \to \mathbf{R}^m$, $x \in \mathbf{R}^n$. *Матрицей Якоби* называется матрица

$$\frac{\partial g}{\partial x} = \begin{pmatrix} \dfrac{\partial g_1}{\partial x_1} & \dfrac{\partial g_1}{\partial x_2} & \dots & \dfrac{\partial g_1}{\partial x_n} \\ \dfrac{\partial g_2}{\partial x_1} & \dfrac{\partial g_2}{\partial x_2} & \dots & \dfrac{\partial g_2}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \dfrac{\partial g_m}{\partial x_1} & \dfrac{\partial g_m}{\partial x_2} & \dots & \dfrac{\partial g_m}{\partial x_n} \end{pmatrix}.$$

В частности, если $m = 1$ (т.е. $g(x)$ — скалярная функция векторного аргумента x), то $\dfrac{\partial g}{\partial x}$ — градиент функции g.

Доказать, что

1) если $a \in \mathbf{R}^n$, $x \in \mathbf{R}^n$, то $\dfrac{\partial(a^\top x)}{\partial x} = a$;

2) если $A \in \mathbf{R}^{m\times n}$, $x \in \mathbf{R}^n$, то $\dfrac{\partial(Ax)}{\partial x} = A$;

3) если $A \in \mathbf{R}^{n\times n}$, $x \in \mathbf{R}^n$, то $\dfrac{\partial(x^\top Ax)}{\partial x} = (A + A^\top)x$; в частности, если $A^\top = A$, то $\dfrac{\partial(x^\top Ax)}{\partial x} = 2Ax$;

4) если $x \in \mathbf{R}^n$, то $\dfrac{\partial\|x\|^2}{\partial x} = 2x$;

5) если g — скалярная функция и под $g(x)$ понимается применение функции g к каждой компоненте вектора $x \in \mathbf{R}^n$, то

$$\frac{\partial g(x)}{\partial x} = \mathrm{diag}(g'(x)),$$

где $\mathrm{diag}(a)$ — диагональная матрица с диагональю a;

6) если $h: \mathbf{R}^n \to \mathbf{R}^m$, $g: \mathbf{R}^m \to \mathbf{R}^p$, $x \in \mathbf{R}^n$, то

$$\frac{\partial g(h(x))}{\partial x} = \frac{\partial g(h(x))}{\partial h}\frac{\partial h(x)}{\partial x}.$$

2 Линейная регрессия

Пример 1. По обучающей выборке

x	-1	0	0	1	2
y	1	-2	1	7	8

методом наименьших квадратов построить модель вида $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$.

Решение: Составляем матрицу $\mathbf{X}$ и вектор $\mathbf{y}$:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} 1 \\ -2 \\ 1 \\ 7 \\ 8 \end{pmatrix}.$$

Имеем

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 5 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{pmatrix}, \qquad \mathbf{X}^\top\mathbf{y} = \begin{pmatrix} 15 \\ 22 \\ 40 \end{pmatrix}.$$

Решая систему нормальных уравнений $\mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top \mathbf{y}$, находим

$$\beta = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

Таким образом, нашли модель $1 + 2x + x^2$ (синяя линия на графике).

Пример 2. Построить модель того же вида методом ридж-регрессии с параметром регуляризации $\lambda = 2$.

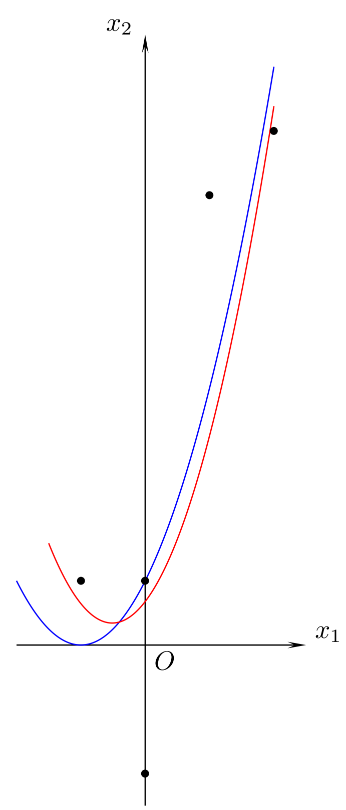
Решение: Для $\lambda = 2$ получаем

$$\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} = \begin{pmatrix} 7 & 2 & 6 \\ 2 & 8 & 8 \\ 6 & 8 & 20 \end{pmatrix}.$$

Решая регуляризованную систему нормальных уравнений $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\beta = \mathbf{X}^\top \mathbf{y}$, находим

$$\beta = \begin{pmatrix} {}^{21}\!/_{31} \\ {}^{81}\!/_{62} \\ {}^{79}\!/_{62} \end{pmatrix}.$$

Таким образом, нашли модель $\dfrac{21}{31} + \dfrac{81}{62}x + \dfrac{79}{62}x^2 = 0.6774 + 1.3065x + 1.2742x^2$ (красная линия на графике).



Задачи для самостоятельного решения

2. Пользуясь № 1, найдите градиент $\dfrac{\partial g(\beta)}{\partial \beta}$ и гессиан $\dfrac{\partial^2 g(\beta)}{\partial \beta^\top \partial \beta}$ функции $g(\beta) = \|\mathbf{X}\beta - \mathbf{y}\|^2$. Выведите отсюда, что решение линейной задачи наименьших квадратов $\widehat{\beta} = \operatorname{argmin} \|\mathbf{X}\beta - \mathbf{y}\|^2$ является решением нормальной системы линейных уравнений $\mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top \mathbf{y}$.

3. Дана обучающая выборка

x	1	1	0	0	−1
y	4	4	0	2	6

1) изобразить точки;
2) методом наименьших квадратов построить модель вида $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$; построить график этой функции;
3) построить модель того же вида методом ридж-регрессии с параметром регуляризации $\lambda = 1$; построить график этой функции.

   Замечание: при ручных вычислениях по методу наименьших квадратов рекомендуется составить систему $\mathbf{X}^\top\mathbf{X}\beta = \mathbf{X}^\top\mathbf{y}$ и решить ее. Регуляризованная система: $(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})\beta = \mathbf{X}^\top\mathbf{y}$, где $\mathbf{I}$ — единичная матрица.

4. Рассмотрим задачу восстановления регрессии, в которой $\mathbf{y}$ распределен согласно нормальному закону $N(\mathbf{X}\beta,\, \sigma^2\mathbf{I})$, 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.

5. 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 d нулей.

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

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

при условии

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

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

7. *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

$$f(x) = \underset{k}{\operatorname{argmax}}\, g_k(x).$$

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?

3 Дискриминантный анализ

Пример 3. Дана обучающая выборка:

x_1	0	2	1	1	1	2	4	4	4	6
x_2	4	4	3	5	4	1	0	2	1	1
y	0	0	0	0	0	1	1	1	1	1

1) Изобразить объекты обучающей выборки в пространстве признаков x_1, x_2;
2) с помощью линейного дискриминантного анализа для каждого класса построить дискриминантные функции и записать уравнение разделяющей поверхности; изобразить поверхность;
3) с помощью квадратичного дискриминантного анализа для каждого класса построить дискриминантные функци.

Решение: Оцениваем вероятности классов:

$$\widehat{\mathsf{Pr}}\,\{Y=0\} = \frac{1}{2}, \qquad \widehat{\mathsf{Pr}}\,\{Y=1\} = \frac{1}{2}.$$

Оцениваем средние для классов:

$$\widehat{\mu}_0 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}, \qquad \widehat{\mu}_1 = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

Выборочные матрицы ковариации для каждого класса:

$$\widehat{\mathbf{\Sigma}}_0 = \frac{1}{N_0-1}\sum_{y^{(i)}=0}(x^{(i)}-\widehat{\mu}_0)(x^{(i)}-\widehat{\mu}_0)^\top = \frac{1}{4}\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

$$\widehat{\mathbf{\Sigma}}_1 = \frac{1}{N_1-1}\sum_{y^{(i)}=1}(x^{(i)}-\widehat{\mu}_1)(x^{(i)}-\widehat{\mu}_1)^\top = \frac{1}{4}\begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix}.$$

Оцениваем матрицу ковариации:

$$\widehat{\mathbf{\Sigma}} = \frac{1}{N-K}\sum_k\sum_{y^{(i)}=k}(x^{(i)}-\widehat{\mu}_k)(x^{(i)}-\widehat{\mu}_k)^\top = \frac{1}{8}\left(\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix}\right) = \frac{1}{8}\begin{pmatrix} 10 & 0 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} 5/4 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

Находим

$$\widehat{\mathbf{\Sigma}}_0^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \widehat{\mathbf{\Sigma}}_1^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \widehat{\mathbf{\Sigma}}^{-1} = \begin{pmatrix} 4/5 & 0 \\ 0 & 2 \end{pmatrix}.$$

Линейные дискриминантные функции:

$$\delta_0(x) = x^\top\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mu}_0 - \frac{1}{2}\widehat{\mu}_0^\top\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mu}_0 + \ln\widehat{\mathsf{Pr}}\,\{Y=0\} = \frac{4}{5}x_1 + 8x_2 - \frac{82}{5} - \ln 2,$$

$$\delta_1(x) = x^\top\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mu}_1 - \frac{1}{2}\widehat{\mu}_1^\top\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mu}_1 + \ln\widehat{\mathsf{Pr}}\,\{Y=1\} = \frac{16}{5}x_1 + 2x_2 - \frac{37}{5} - \ln 2.$$

Разделяющая поверхность — прямая с уравнением $\delta_0(x) = \delta_1(x)$ (красная прямая на графике):

$$4x_1 - 10x_2 + 15 = 0.$$

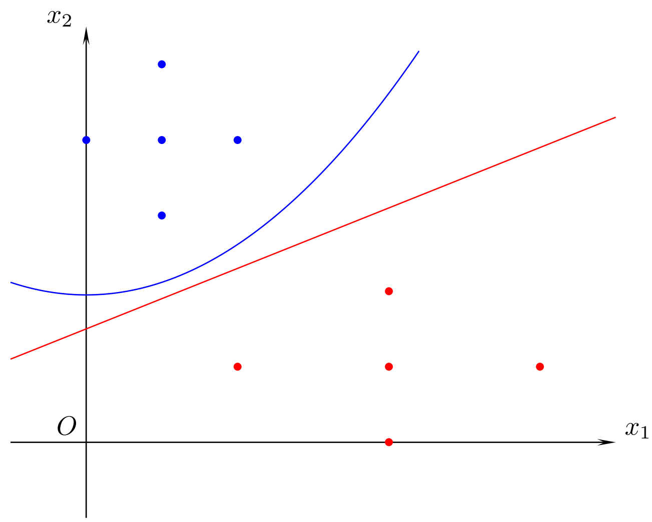
Квадратичные дискриминантные функции:

$$\delta_0(x) = -\frac{1}{2}\ln\det\widehat{\mathbf{\Sigma}}_0 - \frac{1}{2}(x-\widehat{\mu}_0)^\top\widehat{\mathbf{\Sigma}}_0^{-1}(x-\widehat{\mu}_0) + \ln\widehat{\mathsf{Pr}}\,\{Y=0\} = -x_1^2 + 2x_1 - x_2^2 + 8x_2 - 17,$$

$$\delta_1(x) = -\frac{1}{2}\ln\det\widehat{\mathbf{\Sigma}}_1 - \frac{1}{2}(x-\widehat{\mu}_1)^\top\widehat{\mathbf{\Sigma}}_1^{-1}(x-\widehat{\mu}_1) + \ln\widehat{\mathsf{Pr}}\,\{Y=1\} = -\frac{1}{4}x_1^2 + 2x_1 - x_2^2 + 2x_2 - 5 - \ln 2.$$

Разделяющая поверхность — парабола с уравнением (синяя кривая на графике)

$$\frac{3}{4}x_1^2 - 6x_2 + 12 - \ln 2 = 0.$$



Задачи для самостоятельного решения

8.

9. Дана обучающая выборка

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

1) Методом линейного дискриминантного анализа для каждого класса построить дискриминантную функцию и записать уравнение разделяющей поверхности.
2) Методом квадратичного дискриминантного анализа построить дискриминантные функции. Изобразить точки и разделяющие поверхности (кривые).

10. Задача Фишера сводится к максимизации отношения Рэлея

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

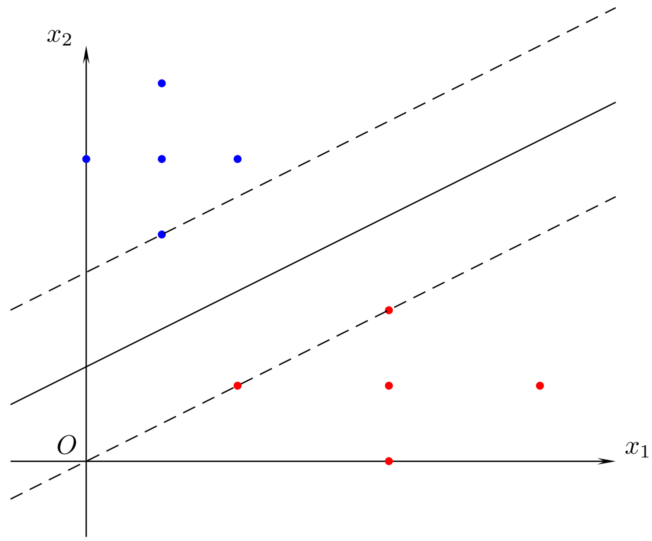
Показать, как эта задача сводится к обобщенной задаче на собственные значения

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

4 Машина опорных векторов

Пример 4. Для данных из предыдущего примера найти уравнение оптимальной разделяющей гиперплоскости, указать опорные точки.

Решение: В качестве двух возможных наборов кандидатур в опорные точки подходят $(1,3)$, $(2,1)$, $(4,2)$ или $(1,3)$, $(2,4)$, $(2,1)$. Среди этих двух наборов нужно выбрать тот, для которого ширина «разделяющего коридора» больше. Для первого набора этот коридор заключен между прямыми (они изображены на рисунке — штриховая малиновая линия) $x_1 - 2x_2 = 0$ и $x_1 - 2x_2 = -5$. Расстояние между прямыми, т. е. ширина коридора, есть $\sqrt{5}$. Для второго набора точек получаем коридор, ограниченный прямыми $x_1 - x_2 = 1$, $x_1 - x_2 = -2$. Его ширина равна $3/\sqrt{2}$. Первый коридор шире, поэтому прямая $2x_1 - 4x_2 = -5$, проходящая через его центр, — оптимальная разделяющая, а точки $(1,3)$, $(2,1)$, $(4,2)$ — опорные.



Задачи для самостоятельного решения

11. 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.

12. 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 $\alpha = 1/(2\gamma)$.

13. *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

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

14. 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.

5 Наивный байесовский классификатор

Пример 5. Дана обучающая выборка:

x_1	1	0	0	0	1	0	1	1	1	0
x_2	1	1	1	1	0	0	0	0	0	0
y	0	0	0	0	0	1	1	1	1	1

С помощью наивного байесова классификатора оценить апостериорные вероятности $\mathsf{Pr}(y|x)$, если (a) $x_1 = 1$, $x_2 = 0$; (b) $x_1 = 0$, $x_2 = 1$.

Оцениваем априорные вероятности:

$$\widehat{\mathsf{Pr}}\{Y=0\} = \frac{1}{2}, \qquad \widehat{\mathsf{Pr}}\{Y=1\} = \frac{1}{2}.$$

Оцениваем условные вероятности:

$$\widehat{\mathsf{Pr}}\{X_1=0|Y=0\} = \frac{3}{5}, \quad \widehat{\mathsf{Pr}}\{X_1=1|Y=0\} = \frac{2}{5}, \qquad \widehat{\mathsf{Pr}}\{X_1=0|Y=1\} = \frac{2}{5}, \quad \widehat{\mathsf{Pr}}\{X_1=1|Y=1\} = \frac{3}{5},$$

$$\widehat{\mathsf{Pr}}\{X_2=0|Y=0\} = \frac{1}{5}, \quad \widehat{\mathsf{Pr}}\{X_2=1|Y=0\} = \frac{4}{5}, \qquad \widehat{\mathsf{Pr}}\{X_2=0|Y=1\} = 1, \quad \widehat{\mathsf{Pr}}\{X_2=1|Y=1\} = 0.$$

Используя основное предположение наивного байесова классификатора, получаем

$$\mathsf{Pr}\{Y=0|X_1=1,X_2=0\} = \frac{\mathsf{Pr}\{X_1=1|Y=0\}\,\mathsf{Pr}\{X_2=0|Y=0\}\,\mathsf{Pr}\{Y=0\}}{\mathsf{Pr}\{X_1=1,\,X_2=0\}} \approx \frac{\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{1}{2}}{\frac{1}{25}+\frac{3}{10}} = \frac{\frac{1}{25}}{\frac{17}{50}} = \frac{2}{17},$$

$$\mathsf{Pr}\{Y=1|X_1=1,X_2=0\} = \frac{\mathsf{Pr}\{X_1=1|Y=1\}\,\mathsf{Pr}\{X_2=0|Y=1\}\,\mathsf{Pr}\{Y=1\}}{\mathsf{Pr}\{X_1=1,\,X_2=0\}} \approx \frac{\frac{3}{5}\cdot 1\cdot\frac{1}{2}}{\frac{1}{25}+\frac{3}{10}} = \frac{\frac{3}{10}}{\frac{17}{50}} = \frac{15}{17}$$

(знаменатели в двух формулах выше равны сумме всех числителей и, следовательно окончательные оценки апостериорных вероятностей получаются после вычисления всех числителей).

$$\mathsf{Pr}\{Y=0|X_1=0,X_2=1\} = \frac{\mathsf{Pr}\{X_1=0|Y=0\}\,\mathsf{Pr}\{X_2=1|Y=0\}\,\mathsf{Pr}\{Y=0\}}{\mathsf{Pr}\{X_1=0,\,X_2=1\}} \approx \frac{\frac{3}{5}\cdot\frac{4}{5}\cdot\frac{1}{2}}{\frac{6}{25}+0} = \frac{\frac{6}{25}}{\frac{6}{25}} = 1,$$

$$\mathsf{Pr}\{Y=1|X_1=0,X_2=1\} = \frac{\mathsf{Pr}\{X_1=0|Y=1\}\,\mathsf{Pr}\{X_2=1|Y=1\}\,\mathsf{Pr}\{Y=1\}}{\mathsf{Pr}\{X_1=0,\,X_2=1\}} \approx \frac{\frac{2}{5}\cdot 0\cdot\frac{1}{2}}{\frac{6}{25}+0} = \frac{0}{\frac{6}{25}} = 0.$$

Задачи для самостоятельного решения

15. Дана обучающая выборка

x_1	0	0	1	1	0	0	1	1	1	0
x_2	0	1	0	1	1	1	1	1	1	1
y	0	0	0	0	0	1	1	1	1	1

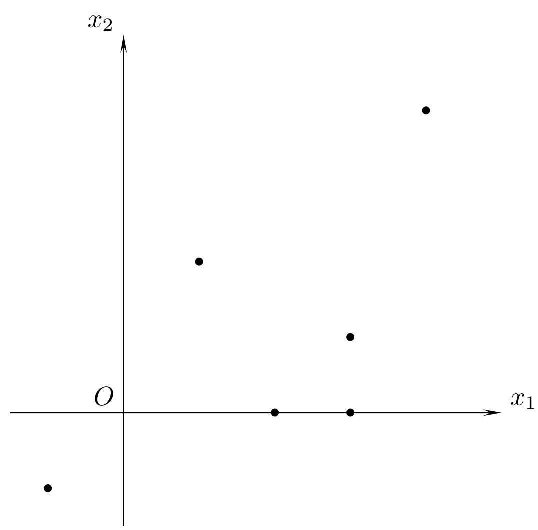
С помощью наивного байесова классификатора оценить вероятности $\mathsf{Pr}(Y = 0 \,|\, X_1 = 1,\ X_2 = 1)$; $\mathsf{Pr}(Y = 1 \,|\, X_1 = 1,\ X_2 = 1)$.

6 Метод главных компонент

Пример 6. Дана выборка:

x_1	4	−1	3	3	2	1
x_2	4	−1	0	1	0	2

Найти главные направления и объясненные дисперсии по главным компонентам.



Решение: Имеем

$$\mathbf{X} = \begin{pmatrix} 4 & 4 \\ -1 & -1 \\ 3 & 0 \\ 3 & 1 \\ 2 & 0 \\ 1 & 2 \end{pmatrix},$$

1 шаг. Центрируем данные. Находим $\overline{x} = (2, 1)$. Вычитая компоненты $\overline{x}$ из первого и второго столбца матрицы $\mathbf{X}$ соответственно, получаем

$$\mathbf{X}_{\text{c}} = \begin{pmatrix} 2 & 3 \\ -3 & -2 \\ 1 & -1 \\ 1 & 0 \\ 0 & -1 \\ -1 & 1 \end{pmatrix}.$$

2 шаг. Находим

$$\mathbf{C} = \mathbf{X}_{\text{c}}^{\top}\mathbf{X}_{\text{c}} = \begin{pmatrix} 16 & 10 \\ 10 & 16 \end{pmatrix}.$$

Матрица

$$\frac{1}{N-1}\mathbf{C} = \begin{pmatrix} {}^{16}\!/\!{}_{5} & 2 \\ 2 & {}^{16}\!/\!{}_{5} \end{pmatrix}.$$

— это выборочная матрица ковариации.

3 шаг. Находим собственные числа и собственные векторы матрицы $\mathbf{C}$:

$$\det(\mathbf{C}-\lambda\mathbf{I}) = \begin{vmatrix} 16-\lambda & 10 \\ 10 & 16-\lambda \end{vmatrix} = \lambda^2 - 32\lambda + 156 = (\lambda-26)(\lambda-6).$$

Собственные векторы нормируем:

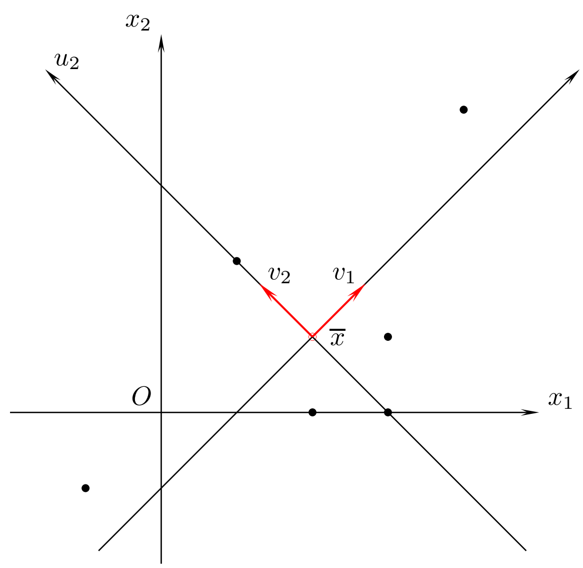
$$\lambda_1 = 26, \quad v_1 = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \lambda_2 = 6, \quad v_2 = \frac{\sqrt{2}}{2}\begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Векторы v_1, v_2 — главные компоненты. Значения

$$\frac{1}{N-1}\lambda_1 = \frac{1}{N-1}\sigma_1^2 = \frac{26}{5} = 5.2, \qquad \frac{1}{N-1}\lambda_2 = \frac{1}{N-1}\sigma_2^2 = \frac{6}{5} = 1.2$$

— дисперсии по главным компонентам. Находим доли объясненной дисперсии:

$$\frac{\lambda_1}{\lambda_1+\lambda_2} = 0.8125, \qquad \frac{\lambda_2}{\lambda_1+\lambda_2} = 0.1875.$$



Найдем сингулярное разложение матрицы $\mathbf{X}_c = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} = \begin{pmatrix} \sqrt{26} & 0 \\ 0 & \sqrt{6} \end{pmatrix}, \qquad \mathbf{V} = (v_1 \mid v_2) = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \approx \begin{pmatrix} 0.7071 & -0.7071 \\ 0.7071 & 0.7071 \end{pmatrix},$$

$$\mathbf{U}\boldsymbol{\Sigma} = \mathbf{X}_c\mathbf{V} = \begin{pmatrix} 5\sqrt{2}/2 & \sqrt{2}/2 \\ -5\sqrt{2}/2 & \sqrt{2}/2 \\ 0 & -\sqrt{2} \\ \sqrt{2}/2 & -\sqrt{2}/2 \\ -\sqrt{2}/2 & -\sqrt{2}/2 \\ 0 & \sqrt{2} \end{pmatrix} \approx \begin{pmatrix} 3.5355 & 0.7071 \\ -3.5355 & 0.7071 \\ 0 & -1.4142 \\ 0.7071 & -0.7071 \\ -0.7071 & -0.7071 \\ 0 & 1.4142 \end{pmatrix},$$

$$\mathbf{U} = \mathbf{X}_c\mathbf{V}\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 5\sqrt{13}/26 & \sqrt{3}/6 \\ -5\sqrt{13}/26 & \sqrt{3}/6 \\ 0 & -\sqrt{3}/3 \\ \sqrt{13}/26 & -\sqrt{3}/6 \\ -\sqrt{13}/26 & -\sqrt{3}/6 \\ 0 & \sqrt{3}/3 \end{pmatrix} \approx \begin{pmatrix} 0.6934 & 0.2887 \\ -0.6934 & 0.2887 \\ 0 & -0.5774 \\ 0.1387 & -0.2887 \\ -0.1387 & -0.2887 \\ 0 & 0.5774 \end{pmatrix}.$$

Напомним, что

- в матрице *нагрузок* (loadings) $\mathbf{V}$ по столбцам записаны координаты векторов главных компонент v_1, v_2;
- в матрице *счетов*, или *результатов*, (scores) $\mathbf{U\Sigma}$ по строкам — координаты проекций точек на главные компоненты;
- в матрице *Z-счетов*, или *Z-результатов*, (z-scores) $\mathbf{U}\sqrt{N}$ по строкам — координаты проекций точек на главные компоненты, нормированные на единичные выборочные дисперсии (whitening).

Задачи для самостоятельного решения

16. Дана выборка:

x_1	4	0	-2	2
x_2	3	1	-3	-1

Найти главные направления и дисперсии по главным компонентам. Изобразить точки и главные направления.

17. Дана выборка:

x_1	4	0	-1	3	4
x_2	2	-3	-2	1	2
x_3	3	2	2	1	-3

Найти главные направления и дисперсии по главным компонентам.

7 Логистическая регрессия. Нейронные сети

18. Пусть $\sigma(z) = \dfrac{1}{1+e^{-z}}$ (сигмоидальная функция). Проверить, что $\sigma' = \sigma(1-\sigma)$.

19. Пусть в задаче классификации на K классов $\{1, 2, \ldots, K\}$ последний слой нейронной сети вычисляет softmax-функцию:

$$g_k(s_1, s_2, \ldots, s_K) = e^{s_k} \Big/ \sum_{\ell=1}^{K} e^{s_\ell} .$$

В качестве штрафа используется кросс-энтропия (logloss-функция):

$$R^{(i)} = -\sum_{k=1}^{K} I(y^{(i)} = k) \ln g_k(s_1, s_2, \ldots, s_K),$$

где $g_k(s_1, s_2, \ldots, s_K)$ — softmax-функция. Доказать, что

1) $\dfrac{\partial g_k}{\partial s_\ell} = g_k \cdot \big(I(k=\ell) - g_\ell\big)$;

2) $\dfrac{\partial R^{(i)}}{\partial g_k} = -\dfrac{I(y^{(i)})}{g_k}$;

3) $\dfrac{\partial R^{(i)}}{\partial s_\ell} = g_\ell - I(\ell = y^{(i)})$.

20. Предположим, что рассматривается K задач двуклассовой классификации, в каждой из которых по $x \in \mathcal{X}$ требуется предсказать $y_k \in \{0,1\}$ $(k = 1, 2, \ldots, K)$ (например, требуется определить, присутствует или отсутствует на изображении x объект k). Обучающая выборка составлена из пар $(x^{(i)}, y^{(i)})$, где $y^{(i)} = (y_1^{(i)}, y_2^{(i)}, \ldots, y_K^{(i)}) \in \{0,1\}^K$ $(i = 1, 2, \ldots, N)$. Для решения такой задачи можно использовать нейронную сеть, последний слой которой вычисляет K сигмоидальных функций

$$g_k(s_k) = \frac{1}{1+e^{-s_k}} \qquad (k = 1, 2, \ldots, K).$$

Пусть в качестве штрафа используется

$$R^{(i)} = -\sum_{k=1}^{K} \Big(y_k^{(i)} \ln g_k + (1 - y_k^{(i)}) \ln(1 - g_k) \Big).$$

Доказать, что

$$\frac{\partial R^{(i)}}{\partial s_k} = g_k - y_k^{(i)}.$$

21. Нейронная сеть с двумя нелинейными слоями вычисляет функцию $\operatorname{softmax}\big(B\big(\sigma(Ax)\big)\big)$. Пусть в качестве функции потерь используется logloss. Таким образом, на объекте $x^{(i)}$ потери равны $\operatorname{logloss}\Big(\operatorname{softmax}\big(B(\sigma(Ax^{(i)}))\big)\Big)$. Пользуясь результатом задачи № 1, выпишите матричные формулы для алгоритма backpropagation.

8 Деревья решений

22. 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.

23. 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.

24. Пусть из обучающая выборки длины N объектов генерируется бутстрэп-выборка (выборка с возвращением) той же длины. Найти математическое ожидание доли объектов, не вошедших в бутстрэп-выборку. Чему равен предел этого математического ожидания при $n \to \infty$. Какой вывод отсюда можно сделать относительно out-of-bag метода оценки качества ансамбля классификаторов?

9 Теория Вапника–Червоненкиса

25. 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

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

где

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

26. 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.

27. Доказать, что при $h \le N$

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

28. 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?

29. 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.

30. 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.

31. Пусть H_d — множество всех булевых функций $f:\ \{0,1\}^d \to \{0,1\}$, представимых ДНФ, в которых каждая элементарная конъюнкиция 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.

32. 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 $\mathrm{VC}\,P_k$.

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

34. *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 $\{\sin \alpha x:\ \alpha \in \mathbf{R}\}$.

10 Еще задачи

35. Рассмотрим задачу восстановления регрессии с квадратичной функцией потерь $L(y', y) = (y' - y)^2$. Доказать, что если $f^*(x) = \underset{c}{\operatorname{argmin}}\, \mathsf{E}\left((Y-c)^2 \,|\, X = x\right)$, то $f^*(x) = \mathsf{E}\,(Y \,|\, X = x)$ (*регрессионная функция*). Чему тогда равен средний риск $R(f^*)$?

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

37. А как должна выглядеть функция потерь, чтобы минимум среднему риску давала условная мода?

38. Пусть в задаче классификации с двумя классами $\{0,1\}$ используется функция потерь $L(y', y)$, такая, что $L(0,0) = L(1,1) = 0$, $L(1,0) = \ell_1$, $L(0,1) = \ell_0$. Докажите, что в этом случае байесов классификатор $f^*(x)$ удовлетворяет условию

$$f(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}}\, \ell_y \operatorname{\mathsf{Pr}}(y \,|\, x).$$

39. Выразить байесов классификатор $f^*(x)$ для задачи классификации с K классами, если функция потерь равна $L(y',\, y) = \ell_{y'y}$ $(y', y = 1, 2, \ldots, K)$.

40. Пусть рассматривается задача бинарной классификации. Доказать, что если известно сколько в выборке представителей каждого из двух классов, то по любым двум показателям из списка TPR, TNR, PPV, NPV определяются остальные два.

41. Пусть рассматривается задача бинарной классификации. Верно ли, что

1) если у двух классификаторов на одной и той же выборке совпадают PPV (Precision) и совпадают TPR (Recall, или Sensitivity), то будут совпадать TNR (Specificity) и NPV;
2) если у двух классификаторов на одной и той же выборке совпадают TNR (Specificity) и совпадают NPV, то будут совпадать PPV (Precision) и TPR (Recall или Sensitivity);
3) совпадение ROC кривых (для двух классификаторов на одной и той же выборке) влечет совпадение Precision-Recall кривых и наоборот?

42. Пусть в задаче классификации на 2 класса $\{0,1\}$ некоторый классификатор (например, наивный байесовский) определяет следующие оценки $g(x)$ апостериорной вероятности принадлежности объекта x к классу 1:

i	1	2	3	4	5	6	7	8	9
$y^{(i)}$	0	0	0	0	0	1	1	1	1
$g(x^{(i)})$	0.75	0.15	0.11	0.23	0.09	0.10	0.66	0.82	0.50

Постройте ROC-кривую. Вычислите AUC. Для классификатора $f(x) = I(g(x) \geq 0.5)$ выпишите матрицу рассогласования и найдите FPR, FNR, TNR, TPR, PPV, *accuracy*, *error*, F1.

43. Видоизмените алгоритм, разобранный на лекции для построения ROC-кривой, так, чтобы он находил ROC AUC.

44. 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 $d \to \infty$, $N = O(d)$. Какой вывод из этого можно сделать применительно к методу ближайшего соседа при больших d?

45. *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 $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$. Разложить $\mathsf{E}_D\big(f(x, D) - f^*(x)\big)^2$ в сумму квадрата математического ожидания смещения (bias) и дисперсии (variance).

46. *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?

47. 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.

48. *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 $x = (0.32, 0)$ окажется ближе (по евклидову расстоянию) к объекту $x^{(1)}$, чем к $x^{(0)}$? (Шум к x не добавляется. Только к $x^{(0)} = 0$ и $x^{(1)} = 1$.)

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

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

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