Комментарии по результатам контрольной работы

В данном тексте собраны наиболее типичные ошибки, которые были допущены при выполнении контрольной работы, а также дан ряд методических указаний, каким образом этих ошибок можно было избежать.

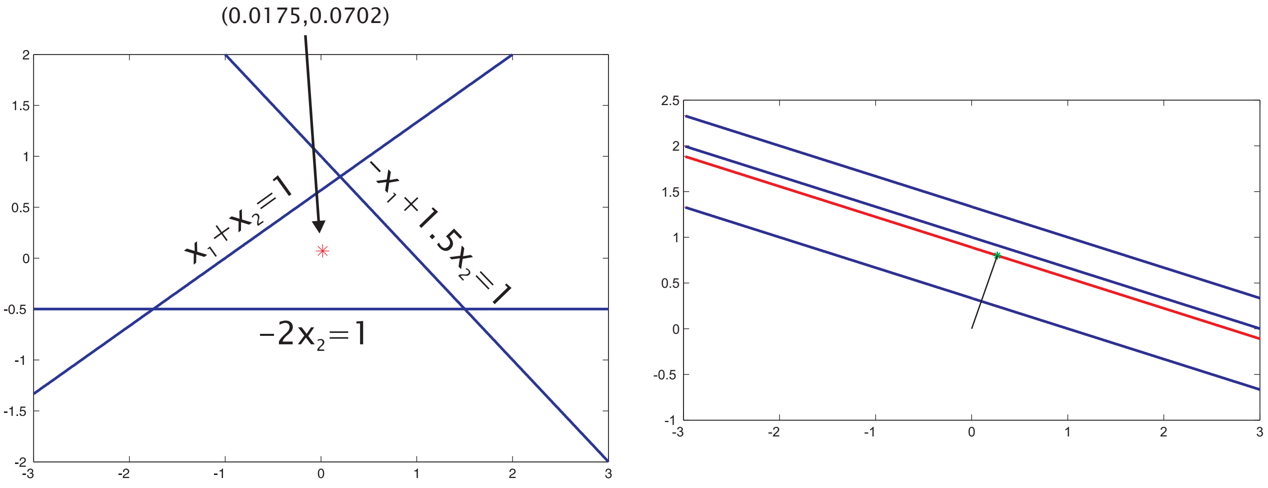
1. При приведении примеров практических задач у многих обнаружилась путаница в понимании разницы между задачами классификации и идентификации. В задаче классификации целевой переменной является метка класса из конечного множества $\{1, \dots, L\}$. Например, для изображений лиц людей классы могут быть связаны с пигментацией волос: «шатен», «брюнет», «русый» и т.д. При этом предполагается, что 1) для каждого класса в наличии имеется значительная выборка данных для обучения и 2) распознаваемые объекты обязательно принадлежат к одному из классов, которые были представлены в обучающей выборке. В задаче идентификации целевой переменной является **бинарная характеристика**, которая показывает, обладает ли объект заданным свойством. При этом предполагается, что для объектов, обладающих соответствующим свойством, можно собрать представительную обучающую выборку данных, а для объектов, не обладающих данным свойством, такую представительную выборку собрать нельзя, т.к. невозможно охарактеризовать класс «все остальное». Например, задача «определить, является ли данная фотография фотографией Иванова Ивана Ивановича» является задачей идентификации, т.к. мы можем относительно легко собрать набор фотографий конкретного человека, а собрать представительную выборку фотографий всех остальных людей не представляется возможным. С другой стороны, задача «определить, является ли данная фотография фотографией шатена» не является задачей идентификации, т.к. если мы предполагаем, что мы можем собрать представительную выборку фотографий шатенов, то мы также можем собрать и представительную выборку брюнетов, блондинов и т.д. (всего классов совсем немного).

2. В решении тестового варианта определение задачи регрессии было дано не четко. В результате, многие приводили некорректные примеры задач регрессии. Задача регрессии ничем не отличается от задачи классификации за одним исключением: целевой переменной в задаче регрессии является вещественнозначная величина (не функция, не зависимость). Например, предсказание абсолютной величины возраста человека по его фотографии является задачей регрессии. Исходом задачи регрессии является не «непрерывная вещественно-значная функция зависимости», а значение вещественнозначной величины (например, возраста человека). При этом функция прогноза этой величины по значениям признаков объекта (функция регрессии) не обязана быть непрерывной функцией.

3. Много ошибок было допущено при решении задачи на поиск нормального псевдорешения системы линейных уравнений. Пусть дана система линейных уравнений в матричном виде:

$$A\boldsymbol{x} = \boldsymbol{b}. \tag{1}$$

   Здесь матрица A не обязана быть квадратной. *Решением* системы (1) называется вектор $\boldsymbol{x}$: $A\boldsymbol{x} = \boldsymbol{b}$. Решение системы не всегда существует. В том случае, когда решения нет, рассматривают т.н. псевдорешение. *Псевдорешением* системы (1) называется вектор $\boldsymbol{x}_{ps} = \arg\min_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2$. Псевдорешение системы всегда существует, но может быть не единственным. В последнем случае рассматривают также нормальное



(a) (b)

Рис. 1: Иллюстрация псевдорешения и нормального псевдорешения. В случае (a) — псевдорешение единственно, в случае (b) псевдорешений бесконечно много (красная линия) и нормальное псевдорешение соответствует основанию перпендикуляра к прямой псевдорешений.

псевдорешение — псевдорешение с минимальной нормой). Обозначим через X_{ps} множество всех псевдорешений системы (1). Тогда *нормальным псевдорешением* системы (1) называется вектор $\boldsymbol{x}_{ps.norm}$: $\|\boldsymbol{x}_{ps.norm}\| = \arg\min_{\boldsymbol{x}\in X_{ps}} \|\boldsymbol{x}\|$.

Рассмотрим методы поиска псевдорешений и нормальных псевдорешений. Легко показать (подробнее см. лекции), что множество псевдорешений X_{ps} системы (1) совпадает с множеством решений системы

$$A^T A\boldsymbol{x} = A^T\boldsymbol{b}. \tag{2}$$

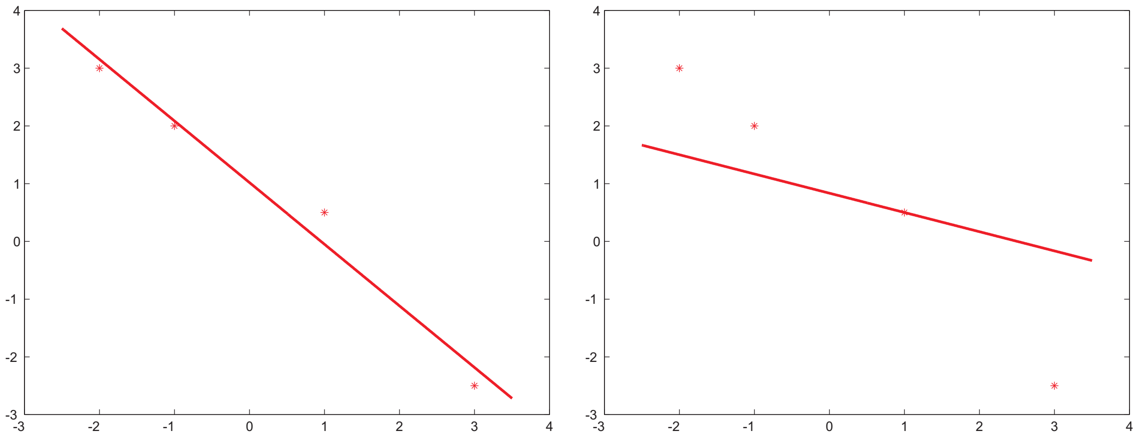
В том случае, если матрица $A^T A$ является вырожденной, то псевдорешений бесконечно много, и имеет смысл искать нормальное псевдорешение. Для этого можно найти общее решение системы (2) (ненулевое частное решение плюс линейная комбинация векторов базиса ядра матрицы $A^T A$), а затем найти среди общих решений решение с минимальной нормой. Другой способ поиска нормального псевдорешения:

$$\boldsymbol{x}_{ps.norm} = \lim_{\lambda\to 0+}(A^T A + \lambda I)^{-1}A^T\boldsymbol{b}.$$

Матрица $A^T A + \lambda I$ всегда невырождена для $\forall\lambda > 0$.

Рассмотрим ситуацию, когда матрица A имеет размер 3×2, т.е. мы имеем дело с тремя линейными уравнениями на плоскости. Тогда понятие псевдорешения и нормального псевдорешения легко проиллюстрировать на картинке. На рис. 1a показана ситуация отсутствия решения и наличия единственного псевдорешения (примерно равноудаленного от всех прямых), а на рис. 1b показана ситуация, когда псевдорешений бесконечно много и среди них можно выбрать нормальное псевдорешение как вектор, находящийся ближе всех к началу координат.

4. При решении задач на построение линейной регрессии, поиске нормального псевдорешения и уменьшения размерности с помощью метода главных компонент было допущено множество арифметических ошибок. Во всех этих задачах решение имеет геометрическую интерпретацию. Поэтому после проведения всех вычислений полученный результат можно легко проверить на здравый смысл, нарисовав соответствующую картинку. Например, при восстановлении линейной регрессии полученная функция регрессии должна проходить максимально близко ко всем точкам выборки (см. рис. 2a). Картинка, показанная на рис. 2b, говорит о том, что функция регрессии восстановлена неверно.



(a) (b)

Рис. 2: Пример восстановления линейной регрессии по 4 точкам. В случае (a) регрессия восстановлена верно, в случае (b) — неверно.

Аналогично, при восстановлении оптимальной гиперплоскости проектирования в методе главных компонент данная гиперплоскость должна проходить через центр масс выборки так, чтобы точки выборки были максимально близки к данной гиперплоскости. На рис. 3 показан пример корректного восстановления такой гиперплоскости (синей звездочкой обозначен центр масс выборки).

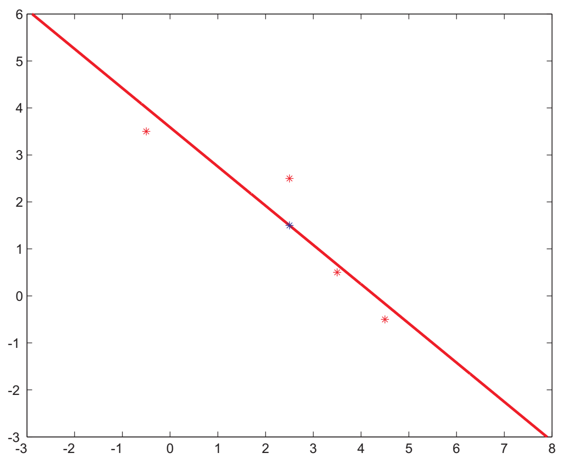


Рис. 3: Пример корректного восстановления оптимальной гиперплоскости в методе главных компонент.

5. В тестовом варианте в задаче на условную оптимизацию было только одно условие. В вариантах контрольной работы условий было несколько, что привело к тому, что многие неверно записали функцию Лагранжа и, как следствие, эквивалентную безусловную задачу оптимизации. Кроме того, даже записав верную безусловную задачу оптимизации, многие побоялись приступить к ее решению ввиду нелинейности полученной системы. Поэтому здесь разбирается полное решение одной из предложенных задач на условную оптимизацию.

   Рассмотрим следующую задачу условной оптимизации:

$$\begin{aligned} &x_1x_2x_3 \to \min, \\ &2x_1x_2 + x_2x_3 = 12, \\ &2x_1 - x_2 = 8. \end{aligned}$$

   Запишем функцию Лагранжа

$$L(x_1, x_2, x_3, \lambda_1, \lambda_2) = x_1x_2x_3 + \lambda_1(2x_1x_2 + x_2x_3 - 12) + \lambda_2(2x_1 - x_2 - 8).$$

Заметим, что в функции Лагранжа вводится свой коэффициент лагранжа λ_i для каждого условия. Приравнивая к нулю производные функции Лагранжа по x_1, x_2, x_3 и добавляя ограничения из исходной задачи, получаем эквивалентную исходной задаче безусловную задачу оптимизации:

$$\frac{\partial L}{\partial x_1} = x_2 x_3 + 2\lambda_1 x_2 + 2\lambda_2 = 0, \tag{3}$$

$$\frac{\partial L}{\partial x_2} = x_1 x_3 + \lambda_1(2x_1 + x_3) - \lambda_2 = 0, \tag{4}$$

$$\frac{\partial L}{\partial x_3} = x_1 x_2 + \lambda_1 x_2 = 0, \tag{5}$$

$$2x_1 x_2 + x_2 x_3 = 12, \tag{6}$$

$$2x_1 - x_2 = 8. \tag{7}$$

Условие (5) эквивалентно тому, что либо $x_2 = 0$, либо $x_1 = -\lambda_1$. Условие $x_2 = 0$ противоречит условию (6). Подставляя $x_1 = -\lambda_1$ в уравнение (4), получаем:

$$-\lambda_1 x_3 + \lambda_1(-2\lambda_1 + x_3) - \lambda_2 = -2\lambda_1^2 - \lambda_2 = 0 \ \Rightarrow \ \lambda_2 = -2\lambda_1^2.$$

Подставляя условия $x_1 = -\lambda_1$ и $\lambda_2 = -2\lambda_1^2$ в уравнения (3),(6),(7), получаем следующую систему уравнений:

$$\begin{aligned} &x_2 x_3 + 2\lambda_1 x_2 - 4\lambda_1^2 = 0, \\ &-2\lambda_1 x_2 + x_2 x_3 = 12, \\ &-2\lambda_1 - x_2 = 8. \end{aligned}$$

Вычитая из первого уравнения второе, и выражая из третьего уравнения x_2 через λ_1, получаем следующее квадратное уравнение:

$$3\lambda_1^2 + 8\lambda_1 - 3 = 0.$$

Решая данное уравнение и переходя к решениям для x_1, x_2, x_3, получаем два решения, подозрительных на минимум исходной задачи:

$$x_1 = 3, x_2 = -2, x_3 = -12; \quad x_1 = -\frac{1}{3}, x_2 = -\frac{26}{3}, x_3 = -\frac{28}{39}.$$

Вычисляем значение целевой функции в обеих полученных точках и выбираем решение, доставляющее минимум целевой функции:

$$f_{min} = -\frac{56}{27}, \ x_1 = -\frac{1}{3}, x_2 = -\frac{26}{3}, x_3 = -\frac{28}{39}.$$

6. Часто допускались ошибки при выписывании функции правдоподобия. Пусть вероятность исхода i равна q_i и исход i наступил ровно N_i раз. Тогда функция правдоподобия записывается как $\prod_i q_i^{N_i}$. При подсчете комбинаторных вероятностей многие забывали учитывать число сочетаний C_n^k. Например, вероятность решить от двух до трех задач из 6 при вероятности решения каждой задачи q составляет $C_6^2 q^2 (1-q)^4 + C_6^3 q^3 (1-q)^3$.

7. При решении задач на поиск нормального псевдорешения и построения линейной регрессии в процессе вычислений обратной матрицы многие допускали ошибки. Избежать этих ошибок можно с помощью простого мнемонического правила обращения матрицы 2×2:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Таким образом, при обращении матрицы 2×2 с точностью до определителя диагональные элементы меняются местами, а внедиагональные элементы меняют знаки.