模式识别第一次作业

201300086 史浩男 人工智能学院

一、教材习题1.1

1.1 以下公式是我在网上一个娱乐短视频剪辑中看到的:

$$\sqrt[3]{a+\frac{a+1}{3}\sqrt{\frac{8a-1}{3}}}+\sqrt[3]{a-\frac{a+1}{3}\sqrt{\frac{8a-1}{3}}}. \tag{1.1}$$

你觉得该式会等于什么? 请注意我们只考虑实数的情况 (即本问题中不会出现复数). 在网上娱乐视频中, 这样一条复杂的公式或许并不常见. 当然, 该式与模式识别或者机器学习也几乎毫不相关. 然而, 正如本题即将说明的那样, 在求解过程中我们能够收获一些有用的思维模式. 这些思维模式在机器学习和模式识别的学习过程中同样是至关重要的. 因此, 让我们来仔细看看这条公式.

(a) **对输入的要求.** 在一个模式识别或机器学习问题中, 我们必须对输入数据进行一些强制约束. 这些约束可能通过预处理技术得以实现, 也可能作为数据采集过程中的限制, 或者通过其他一些方式体现.
在上述式子中, 输入的要求是以 a 的形式给出的. 那么, 我们该如何强制要求变量 a 呢?

(b) **观察数据与问题.** 解决模式识别或机器学习问题的第一步通常是对数据进行观察或可视化, 换言之, 是获得一些关于手头问题的直觉. 在对数据进行观察或可视化的过程中, 有两种数据是流行的选择: 具有代表性的数据 (可以观察到一些共有属性), 以及那些具有特殊性质的数据 (可以观察到一些极端情况).
关于公式 (1.1) 的特殊数据的一个样例是 $a=\frac{1}{8}$. a 的这一取值具有的特殊性在于它将极大地简化该式. 那么, 当 a 取该值时, 公式 (1.1) 的值是多少?

(c) **提出你的想法.** 在对数据进行观察之后, 你可能会想到一些关于解决该问题的直觉或想法. 如果该想法是合乎情理的, 那便值得去探索.
你能找到 a 的其他特殊样例吗? 在那种情况下, 公式 (1.1) 的取值是多少? 基于以上的观察, 你对公式 (1.1) 有什么想法吗?

(d) **对你的想法进行合理性检验.** 如何能确信你的想法是合理的呢? 一种常用的方法是在简单的情形上进行测试, 或者写一套简单的原型系统来验证你的想法.
对于公式 (1.1), 我们可以写一条 Matlab/Octave 命令来对其进行求值. 例如, 使用 `a=3/4` 来对 a 进行赋值, 我们可以使用下式来计算公式 (1.1):

```
f = ( a + (a+1)/3 * sqrt((8*a-1)/3) )^(1/3) + ...
    ( a - (a+1)/3 * sqrt((8*a-1)/3) )^(1/3)
```

这条命令的返回值是多少?

(e) **避免编程中的陷阱.** 当然, 这条命令的返回值是错误的——我们知道其结果应该是实数. 导致该问题的原因是什么? 其实这正是由于编程过程中一个小小的陷阱导致的. 我们应该对原型系统的编程细节给予足够的重视, 以保证它能够正确地实现我们的想法.
阅读 Matlab 的在线手册并尝试修复这一问题. 正确的结果应该是多少? 如果你使用了正确的代码并对 a 的许多不同取值 ($a \geqslant 0.125$) 计算公式 (1.1), 它能否支持你的想法?
你或许一开始就提出了一个好想法并得到了代码的支持. 如果是这样的话, 你可以进入下一部分. 否则, 请仔细观察数据, 并提出一个比原来更好的想法, 测试它, 直到它通过了你的合理性检验实验.

(f) **形式化且严谨的证明.** 不可避免地, 你需要在一些任务上形式化地证明你的结论. 证明过程需要正确性与严谨性. 有效证明的第一步或许是定义你的符号与标记, 这样你才能用数学语言来准确地描述你的问题和想法.
定义你的符号并以精准的数学表述写下你的想法. 然后, 严谨地证明它.

(g) **当可行时, 充分利用现有的结果.** 在研究和研发的过程中, 我们必须充分利用现有的资源, 如数学定理、优化方法、软件库以及开发框架. 话虽如此, 为了使用现有的结果、资源和工具, 那便意味着你得听说过它们. 因此, 充分了解自己相关领域中的主要结果和工具是有用的, 即便你对其具体细节不甚了了.
当然, 这些资源和工具包括那些你自己研发的成果. 使用你在本问题中刚刚证明的定理来计算下述表达式:

$$\sqrt[3]{2+\sqrt{5}}+\sqrt[3]{2-\sqrt{5}}.$$

(h) **或许还可以将你的结果扩展到更具有通用性的理论.** 你的一些结果有可能会成为一种更普遍、更有用的理论. 并且, 当出现这种可能性时, 这样做是相当值得的.
上述公式实际上来自于一个更一般的结论: 卡丹 (Cardano) 的三次方程解法. Gerolamo Cardano 是一名意大利数学家, 他证明了方程

$$z^3+pz+q=0$$

的根可通过与公式 (1.1) 相关的表达式来求解. 仔细阅读https://en.wikipedia.org/wiki/Cubic_equation网页上的信息 (尤其是与卡丹方法相关的部分), 并尝试理解这种联系.

(a)

为了使表达式只出现实数，a必须限制范围：

$$8a-1\geq 0, a\geq \frac{1}{8} \tag{1}$$

(b)

$$a=1/8\text{时，表达式}=2a^{\frac{1}{3}}=1 \tag{2}$$

(c)

$a=1/2$也是特殊样例，此时公式值为1

想法：此公式的值恒为1

(d)求这条命令的返回值：

```
>> a=3/4;
>> b=(a+(a+1)/3*sqrt((8*a-1)/3))^(1/3)+(a-(a+1)/3*sqrt((8*a-1)/3))^(1/3)

b =

   1.2182 + 0.1260i
```

(e)

返回值错误原因：matlab对负数开三次根得到三个解，默认输出第一个解而不是默认输出我们需要的实数解

解决办法：更改代码中开根方式

```
>> a=3/4;
>> b=nthroot(a+(a+1)/3*sqrt((8*a-1)/3),3)+nthroot(a-(a+1)/3*sqrt((8*a-1)/3),3)

b =

    1.0000
```

(f)

准确描述想法：

$$\text{当}b=\frac{(a+1)^2(8a-1)}{27}\text{且}a\geq\frac{1}{8}\text{时，}\sqrt[3]{a+\sqrt{b}}+\sqrt[3]{a-\sqrt{b}}=1\text{恒成立} \tag{3}$$

严谨证明：

$$
\begin{aligned}
&\text{令} x=\sqrt[3]{a+\sqrt{b}}, y=\sqrt[3]{a-\sqrt{b}}, x+y=z \\
&\text{则} x^3+y^3=2a \\
&xy=\sqrt[3]{a^2-b}=\sqrt[3]{a^2-\frac{(a+1)^2(8a-1)}{27}}=\frac{-2a+1}{3}
\end{aligned} \tag{4}
$$

$$
\begin{aligned}
x^3+y^3&=(x+y)((x+y)^2-3xy) \\
&=z^3-3xyz
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
&\text{得到方程} z^3+(2a-1)z-2a=0 \\
&\text{方程有唯一实数解} z=1\text{，即} x+y\text{有唯一实数取值}1\text{，恒为}1\text{成立}
\end{aligned} \tag{6}
$$

(g)

对应上述定理中$a=2, b=5$，因此表达式值=1

(h)理解与原公式与卡丹方法的联系

卡丹方法：

$$
\begin{aligned}
&\text{若} \Delta=\frac{q^2}{4}+p^3 27>0\text{，则方程} z^3+pz+q=0\text{有实根：} \\
&\sqrt[3]{-\frac{q}{2}+\sqrt{\Delta}}+\sqrt[3]{-\frac{q}{2}-\sqrt{\Delta}}
\end{aligned} \tag{7}
$$

在本题原公式中，$q=-2a, \Delta=b$，因此原公式$\sqrt[3]{a+\sqrt{b}}+\sqrt[3]{a-\sqrt{b}}=1$就是方程$z^3+(2a-1)z-2a=0$的实根

二、正态分布估计

若 $X \sim \mathcal{N}(0,1)$, 证明以下不等式:

(a). 对于任意 $\epsilon>0$, 有

$$
P(X \geq \epsilon) \leq \frac{1}{2} e^{-\epsilon^2/2}.
$$

$$
\begin{aligned}
P(X \geqslant \epsilon) &= \int_{\epsilon}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\
&= \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(x+\epsilon)^2/2} dx \\
&= e^{-\epsilon^2/2} \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{-x\varepsilon} dx \\
&\leqslant e^{-\epsilon^2/2} \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\
&= \frac{1}{2} e^{-\epsilon^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \text{（正态分布的规范性）} \\
&= \frac{1}{2} e^{-\epsilon^2/2}
\end{aligned} \tag{8}
$$

Mill不等式

(b). 对于任意 $\epsilon > 0$, 有

$$
P(|X| \geq \epsilon) \leq \min\left\{1, \sqrt{\frac{2}{\pi}} \frac{e^{-\epsilon^2/2}}{\epsilon}\right\}.
$$

提示: 对于 $\mathcal{N}(0,1)$ 的概率密度函数 $f(x)$, 有 $f'(x) = -xf(x)$.

$$
\begin{aligned}
P(|X| \geqslant \epsilon) &= 2\int_{\epsilon}^{+\infty} f(t)dt = 2\int_{\epsilon}^{+\infty} \frac{tf(t)}{t} dt \\
&\leqslant 2\int_{\epsilon}^{+\infty} \frac{tf(t)}{\epsilon} dt = -2\int_{\epsilon}^{+\infty} \frac{f'(t)}{\epsilon} dt \\
&= -\frac{2}{\epsilon}[f(t)]_{\epsilon}^{+\infty} \\
&= \frac{2}{\epsilon} \frac{1}{\sqrt{2\pi}} e^{-\epsilon^2/2} = \sqrt{\frac{2}{\pi}} \frac{e^{-\epsilon^2/2}}{\varepsilon}
\end{aligned} \tag{9}
$$

三、

一个函数 f 的共轭函数 (conjugate function) 定义为

$$
f^*(y) = \sup_{x \in \mathrm{dom}\, f} \left(y^T x - f(x)\right).
$$

(a). 证明 $\inf_x f(x) = -f^*(0)$.

(b). 证明对任意 x, y, $f(x) + f^*(y) \geq x^T y$ (对于 $x \notin \mathrm{dom}(f)$ 令 $f(x) = \infty$).

(c). 证明对任意 x, $f^{**}(x) \leq f(x)$, 其中 $f^{**}(x)$ 为 f^* 的共轭函数.

(a)

$$
-f^*(0) = -sup_x(-f(x)) = inf_x f(x) \tag{10}
$$

(b)

对于$x \in \mathrm{dom}(f)$:

$$f^*(y) = sup_x(y^T x - f(x)) \geq x^T y - f(x) \tag{11}$$

对于$x \notin \mathrm{dom}(f)$:

$$f(x) = \infty \geq x^T y - f^*(y) \tag{12}$$

(c)

$$\begin{aligned} f^{**}(x) &= sup_y(x^T y - f^*(y)) \\ &= sup_y(x^T y - sup_z(z^T y - f^*(z))) \\ &\leq sup_y(x^T y - (x^T y - f(x))) \\ &= sup_y(f(x)) = f(x) \end{aligned} \tag{13}$$

四、

在教材第三章中，我们了解到细节问题（p43）对设计一个模式识别系统的影响。现在我们将探讨如何解决以下细节问题（以教材中的人脸识别为例）

a) 假设存储在设备中的人脸图像是 100×100 的分辨率，即 $\boldsymbol{x} \in \mathbb{R}^{10000}$，而设备将你的照片拍成 400×400。请写出两种不同的预处理方式，使得你的照片能和设备中的照片正常匹配。

(a)两种预处理方式

1、maxpooling最大池化，把400x400划分为100x100个4x4，对每个4x4做步长为4的最大池化

2、平均池化：把400x400划分为100x100个4x4，对每个4x4做平均池化

b) 我们假设一共有 n 张照片，且将每张存储的照片看作一个 100×100 的矩阵。已知两两不相交的 2×2 的像素格内都具有相似的像素值，如下矩阵示意：

$$\begin{bmatrix} 1 & 1 & 155 & 156 & \dots \\ 1 & 1 & 154 & 155 & \dots \\ 50 & 51 & 254 & 253 & \dots \\ 49 & 50 & 255 & 255 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{1}$$

你有什么办法能降低存储照片的容量开销吗？存储开销能降低多少？

(b)降低存储开销

对每个2x2做步长为2的平均池化，能把存储开销降低至1/4

c) 教材中提到了不平衡二分类问题（p46）。我们假设训练集中：A 类有 9900 个样本，B 类有 100 个样本。测试集中：A 类有 5000 个样本，B 类有 5000 个样本。如果我们学习到一个映射 $f(\cdot)$，它将所有输入的样本都预测为 A 类，那么我们在训练集上的准确率 acc_{train} 是多少？在测试集上的准确率 acc_{test} 是多少？

(c)准确率

$acc_{train} = 99\%, acc_{test} = 50\%$

d) 你可能已经知道计算准确率有两种不同的计算方法：micro 和 macro。请简要描述评价指标计算方法中 micro 和 macro 两种计算方式的区别？在 c) 中我们计算准确率用到的是 micro 还是 macro 的计算方式？如果不了解这两者的区别，请搜索网上资源，自行了解他们的区别。

(d)micro&macro区别

是多分类问题中不同的取平均方式，在c)中使用的是micro

micro：考虑到了每个类别的数量，所以适用于**数据分布不平衡**的情况。$acc = \frac{\sum_{i=1}^{n} \mathrm{TP}_i}{\sum_{i=1}^{n} \mathrm{TP}_i + \sum_{i=1}^{n} \mathrm{FP}_i}$

macro：没有考虑到数据的数量，所以会平等的看待每一类。$acc = \frac{1}{n} \sum_{i=1}^{n} \mathrm{P}_i$

e) 上述问题实际上描述的是一个长尾识别问题（long-tailed recognition problem）。在这种问题下，我们在训练集上应当采取哪种计算方式来评估准确率？请设计一种针对此问题的训练方法，使得训练集中样本量少的类别 B 能够在测试集上减少误判？此处只需描述主要思路即可，无需提供技术细节。

(e)设计针对性训练方法

应该采用PR或ROC曲线来分析

针对性训练方法：

1. 重采样的方法：对A类样本欠采样，对B类样本过采样。但可能与真实分布不符，导致训练出的模型泛化能力差
2. 错误非均等代价：更改损失函数计算方式，按数据不平衡的程度，加大把小类别数据分类错误的惩罚

五、

我们考虑近邻分类器问题。给定一个包含 8 个样本的训练集 $S = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_8\}$，其中 $\boldsymbol{x}_1 = (0, 0)$，$\boldsymbol{x}_2 = (0, 1)$，$\boldsymbol{x}_3 = (0, -1)$，$\boldsymbol{x}_4 = (-1, 0)$，$\boldsymbol{x}_5 = (1, 0)$，$\boldsymbol{x}_6 = (8, 0)$，$\boldsymbol{x}_7 = (8, 1)$，$\boldsymbol{x}_8 = (9, 0)$。它们的类别分别是 (A, A, A, A, A, B, A, B)

a) 对于两个测试样本 $\boldsymbol{z}_1 = (0, -2)$，$\boldsymbol{z}_2 = (8, 2)$，运用最近邻分类器（1-NN），得到这两个样本的分类结果是什么？

(a)1-NN计算示例

与z_1最近的是x_3，因此分类为A。与z_2最近的是x_7，因此分类为A。

> b) 同样的两个样本 $\boldsymbol{z}_1, \boldsymbol{z}_2$，运用近邻分类器 k-NN，取 k=3。得到的两个样本的分类结果是什么？
>
> c) 分析两次结果不同的原因？
>
> d) $\boldsymbol{x}_7$ 是否可能属于类别 B？在此情况下 k-NN 相比 1-NN 的优势在何处？

(b)3-NN计算示例

与z_1最近的3个是x_1, x_3, x_4，因此分类为A。与z_2最近的3个是x_7, x_8, x_9，因此分类为B。

(c)

结果不同因为最近邻分类中要取邻居中占比更多的那个类别，1-NN只考虑了一个邻居，结果比较片面不准确，在3-NN中有两个邻居都是B，占了大多数，因此最终分类为B

(d)

x_7极有可能属于类别B

k-NN优势在于考虑了更多邻居的信息，可以对抗样本标记中的噪声，得到更可靠的分类结果

六、感想与收获

1、KNN不需要训练，训练过程只是单纯保存数据而已

2、验证集使用后确定了超参数，此时扩大训练集重新训练这一思路给了我启发。我突然理解了在其他项目中遇到的训练集和测试机划分时故意重叠的设置，原来重叠的部分是验证集，设置成重叠也是为了最后训练时更方便。

3、k折的划分使用random.shuffle真是神来之笔！之前写其他算法时，一直苦于没有找到一种简介的划分写法，甚至有时还将k折交叉验证错写成了蒙特卡洛交叉验证，导致偏差增大，方差减小，无偏估计也变成了一致估计。

本题为一道编程题：从零开始构建一个机器学习系统，请参见‘main.ipynb’文件中的提示来完成相关的代码（请自行安装 Jupyter Notebook）。这份工程的功能包括：

1. 常见的机器学习数据集的读取过程（已提供）
2. 训练和验证集的划分（已提供）
3. 实现一个 KNN 分类器（需完成）
4. 实现评估指标-准确率的计算（需完成）
5. 根据验证集进行超参数选择（需完成）
6. 实现 5 折交叉验证并进行超参数选择（需完成）
7. 最终确定超参数之后，完成在测试集上的测试（需完成）
8. 针对不均衡数据集，实现 precision, recall 和 F1 score 的计算 (需完成)

在完成代码后，提交时需要 notebook 文件（包括代码和中间输出结果，notebook 可直接输出成 pdf 或 html），并谈谈你在这次编程的感想（可以包括你遇到的问题、收获等等）。