模式识别第四次作业

201300086 史浩男 人工智能学院

一、教材习题

9.6 在本题中我们将使用 LIBLINEAR 软件并且尝试一种特定的数据变换.

(a) 下载 LIBLINEAR 软件 (https://www.csie.ntu.edu.tw/~cjlin/liblinear/) 并且学习如何使用它. 你也可以用 Matlab/Octave 绑定并且用 Matlab/Octave 来调用.

(b) 从此网址 https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#mnist 下载 MNIST 数据集. 使用非缩放 (non-scaled) 的版本, 这会包括一个训练集和一个测试集. 使用 LIBLINEAR 的默认参数, 准确率怎么样?

(c) 对每个特征值 (包括训练和测试样例), 使用如下数据变换

$$x \leftarrow \sqrt{x}.$$

在这个数据变换之后准确率怎么样?

(d) 为什么开根变换会这样影响准确率?

(a)

```python
from liblinear.python.liblinear import problem
from liblinear.python.commonutil import *
import math


def process(filename):
    y, x = svm_read_problem(filename)
    for line in x:
        for key in line.keys():
            line[key] = line[key] ** 0.5
    file = open(filename + "_pre", "a")
    for i in range(0, len(x)):
        output = ""
        output += str(int(y[i]))
        line = x[i]
        for key in line.keys():
            output += " "
            output += str(key)
            output += ":"
            output += str(line[key])
        output += "\n"
        file.write(output)
    file.close()

process("mnist.scale")
```

(b)

使用默认参数准确率：

```
Accuracy = 86.58%
```

(c)

使用数据变换后准确率变成

```
Accuracy = 86.37%
```

(d)

正常来讲，准确率应该有提升。因为数值较大的特征回怼结果有影响，开根后这种影响应该减小，相当于数据缩放。

但准确率下降了，可能是版本问题，导致默认参数变化

二、教材习题

9.7 (sigmoid 函数) 对数几率 sigmoid 函数在机器学习和模式识别特别是在神经网络领域中被广泛使用. 令

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

表示对数几率 sigmoid 函数. 我们将在本题中研究这个函数.

(a) 证明 $1-\sigma(x) = \sigma(-x)$.

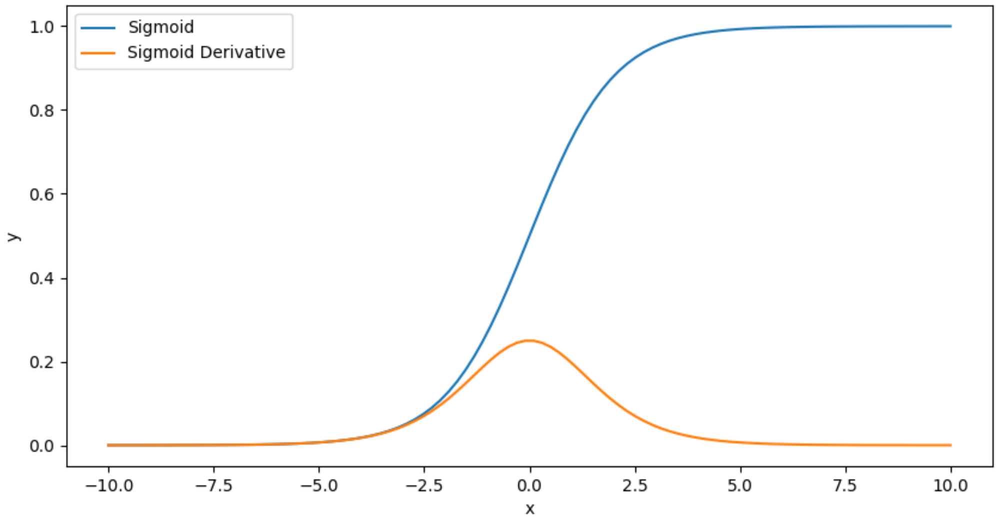
(b) 证明 $\sigma'(x) = \sigma(x)(1-\sigma(x))$, 其中 $\sigma'(x)$ 表示 $\frac{\mathrm{d}}{\mathrm{d}x}\sigma(x)$. 画图同时展示 $\sigma(x)$ 和 $\sigma'(x)$ 的函数曲线.

(a)

$$1-\sigma(x) = 1-\frac{1}{1+e^{-x}} = \frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{x}} = \sigma(-x) \tag{1}$$

(b)

$$\sigma'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\sigma(x) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{1+e^{-x}} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\cdot\frac{1}{1+e^{x}} = \sigma(x)(1-\sigma(x)) \tag{2}$$



(c)

首先 $\boldsymbol{z}^{(i)}=f\left(\boldsymbol{x}^{(i)},\boldsymbol{\theta}^{(i)}\right)$ 是第 i 层网络激活函数前对第 i 层网络输入$\boldsymbol{x}^{(i)}$的处理

$\sigma(\cdot)$是激活函数

第 i 层网络可以表达为：

$$\boldsymbol{y}^{(i)}=\sigma\left(\boldsymbol{z}^{(i)}\right)=\sigma\left(f\left(\boldsymbol{x}^{(i)},\boldsymbol{\theta}^{(i)}\right)\right) \tag{3}$$

由链式法则有

$$\frac{\partial \ell}{\partial\left(\boldsymbol{\theta}^{(i)}\right)^{T}}=\frac{\partial \ell}{\partial\left(\boldsymbol{y}^{(i)}\right)^{T}} \frac{\partial \boldsymbol{y}^{(i)}}{\partial\left(\boldsymbol{z}^{(i)}\right)^{T}} \frac{\partial \boldsymbol{z}^{(i)}}{\partial\left(\boldsymbol{\theta}^{(i)}\right)^{T}} \tag{4}$$

由于sigmoid函数的特点:

$$\begin{gathered}\text{在 }\left|z_j^{(i)}\right| \rightarrow \infty \text{时} \sigma\left(z_j^{(i)}\right) \rightarrow 0 \\ \sigma^{\prime}\left(z_j^{(i)}\right)<=0.25\end{gathered} \tag{5}$$

因此，对于这一层的每个元素j:

$$\left[\frac{\partial \boldsymbol{y}^{(i)}}{\partial\left(\boldsymbol{z}^{(i)}\right)^{T}}\right]_j=\sigma^{\prime}\left(z_j^{(i)}\right)<=0.25 \tag{6}$$

于是连乘会导致 $\left\|\frac{\partial \ell}{\partial\left(\left(\theta^{(i)}\right)\right)^{T}}\right\|$ 就越趋近于 0 , 即梯度消失困难

(c) 一个神经网络经常是一个逐层处理的机器. 例如, 输入 $\boldsymbol{x}$ 对应的输出 $\boldsymbol{y}$ 可以通过

$$\boldsymbol{y}=f^{(L)}\left(f^{(L-1)}\left(\cdots f^{(2)}\left(f^{(1)}(\boldsymbol{x})\right)\right)\right),$$

产生, 其中 $f^{(i)}$ 是一个描述第 i 层处理过程的数学函数. 当处理层数 L 很大时, 它通常被称为深度神经网络 (参见第 15 章).

随机梯度下降 (stochastic gradient descent) 通常被用于优化神经网络. 令 $\boldsymbol{\theta}$ 为网络中所有参数的当前值, $\boldsymbol{g}$ 是损失函数对 $\boldsymbol{\theta}$ 的梯度, 那么 $\boldsymbol{\theta}$ 的更新方式是

$$\boldsymbol{\theta}^{new}\leftarrow\boldsymbol{\theta}-\lambda\boldsymbol{g}, \tag{9.63}$$

其中 λ 是一个正的学习率 (learning rate).

梯度 $\boldsymbol{g}$ 用链式法则 (chain rule) 计算. 令 $\boldsymbol{\theta}^{(i)}$ 为第 i 层的参数, $\boldsymbol{y}^{(i)}$ 是经过前 i 层计算后的输出. 那么

$$\frac{\partial\ell}{\partial(\boldsymbol{\theta}^{(i)})^T}=\frac{\partial\ell}{\partial(\boldsymbol{y}^{(i)})^T}\frac{\partial\boldsymbol{y}^{(i)}}{\partial(\boldsymbol{\theta}^{(i)})^T}, \tag{9.64}$$

其中 ℓ 是需要被最小化的损失. 这个计算叫作*误差反向传播*(error backpropagation), 因为 ℓ 的误差从最后一层反向传递至第一层.

然而, 这个学习策略经常会遭遇*梯度消失*(diminishing gradient) 问题, 其含义是对有的 i, 当前层的梯度 $\frac{\partial\ell}{\partial(\boldsymbol{\theta}^{(i)})^T}$ 变得非常小, 或者当 i 由 L 变为 1 时, 很快有 $\left\|\frac{\partial\ell}{\partial(\boldsymbol{\theta}^{(i)})^T}\right\|\to 0$. sigmoid 函数 $\sigma(x)$ 在神经网络中很流行. 许多层 $f^{(i)}$ 对其输入的每个元素分别运用 sigmoid 函数.

说明 sigmoid 函数容易导致梯度消失困难. (提示: 你可以只看梯度中的单个元素. 看看你在上一个子问题中画的图.)

三、

10.6 令 X 是一个连续随机变量, 其概率密度函数是 $q(x)$. 假设当 $x\geqslant 0$ 时 $q(x)>0$; 当 $x<0$ 时 $q(x)=0$. 进一步地, 假设 X 的均值是 $\mu>0$, 并且 X 的熵存在.

证明参数为 $\lambda=\frac{1}{\mu}$ 的指数分布是在这样约束条件的最大熵分布 (maximum entropy distribution).

不妨设总体分布为 p(x) , 优化目标如下:

$$\begin{gathered}\max -\int_X p(x)\log p(x)\mathrm{d}x\\ \text{s.t. } \int_X p(x)\mathrm{d}x=1\\ \int_X p(x)\mathrm{d}x=\mu\end{gathered}\tag{7}$$

可以列出拉格朗日函数:

$$
\begin{aligned}
\mathcal{L}\left(p(x), \lambda_0, \lambda_1\right) &= -\int_X p(x)\log p(x)dx + \lambda_0\left(\int_X p(x)dx - 1\right) + \lambda_1\left(\int_X x\cdot p(x)dX - \mu\right) \\
\text{and } \frac{\partial \mathcal{L}}{\partial p(x)} &= 0 = -\log p(x) - 1 + \lambda_0 + \lambda_1 x \\
\frac{\partial \mathcal{L}}{\partial \lambda_0} &= 0 = \int_X p(x)dx - 1 \\
\frac{\partial \mathcal{L}}{\partial \lambda_1} &= 0 = \int_X x\cdot p(x)dx - \mu \\
\Longrightarrow p(x) &= \exp\left(-1 + \lambda_0 + \lambda_1 x\right) \text{ and } \int_X p(x)dx = 1, \int_X x\cdot p(x)dx = \mu
\end{aligned} \tag{8}
$$

结合概率密度函数 $q(x) = 0, x < 0$，代入得:

$p(x) = \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)$ where $x \in [0, \infty)$

所以参数 $\lambda = \frac{1}{\mu}$的指数分布是在这样约束条件的最大熵分布.

四、

11.1 (软阈值, soft thresholding) 令 $\lambda > 0$. 证明

$$
\arg\min_{\boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{x}\|_1 \tag{11.24}
$$

的解是将带收缩参数 $\frac{\lambda}{2}$ 的软阈值策略应用到 $\boldsymbol{y}$ 各维的结果. 即

$$
\boldsymbol{x}^\star = \operatorname{sign}(\boldsymbol{y})\left(|\boldsymbol{y}| - \frac{\lambda}{2}\right)_+, \tag{11.25}
$$

其中符号函数 sign、绝对值函数、取负数操作、$(\cdot)_+$ 阈值函数和乘法都是逐元素进行的. 如果我们记收缩 - 阈值操作符为

$$
\mathcal{T}_\lambda(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{x})\left(|\boldsymbol{x}| - \lambda\right)_+, \tag{11.26}
$$

解可以表示为 $\boldsymbol{x}^\star = \mathcal{T}_{\frac{\lambda}{2}}(\boldsymbol{y})$.

由于λ>0，可以更清楚的表示收缩-阈值操作符：

$$
\mathcal{T}_\lambda(x) = \operatorname{sign}(x)(|x| - \lambda)_+ = \begin{cases} x + \lambda, & x < -\lambda \\ 0, & -\lambda \le x \le \lambda \\ x - \lambda, & x > \lambda \end{cases} \tag{9}
$$

这种表示也暗示我们，在求解时也要分成三段

1、问题分解

$F(x) = \left((x_1 - y_1)^2 + \lambda\left|x_1\right|\right) + \cdots + \left((x_n - y_1)^2 + \lambda\left|x_n\right|\right)$

我们又令 $f_i\left(x_i\right) = \left(x_i - y_i\right)^2 + \lambda\left|x_i\right|, i = 1, 2, \ldots, n$，

所以$f_i\left(x_i\right)$ 之间有独立性

对函数 $f(x) = (x - y)^2 + \lambda|x|$求导可得$f'(x) = 2(x - y) + \lambda\operatorname{sign}(x), x \neq 0$

令导数等于零可得$x = y - \frac{\lambda}{2}\operatorname{sign}(x), x \neq 0$

2、当 $y<-\frac{\lambda}{2}$

x只能有 $x=y+\frac{\lambda}{2}$ 与 x=0 这两种可能取值，因为假设 x>0 则有
$x=y-\frac{\lambda}{2}\operatorname{sign}(x)=y-\frac{\lambda}{2}<-\lambda<0$ 假设不成立.

带入 $f(x)=(x-y)^2+\lambda|x|$ 有

$$f(0)-f\left(y+\frac{\lambda}{2}\right)=y^2-\left[\left(\frac{\lambda}{2}\right)^2-\lambda\left(y+\frac{\lambda}{2}\right)\right]=\left(y+\frac{\lambda}{2}\right)^2>0 \tag{10}$$

即有$f\left(y+\frac{\lambda}{2}\right)<f(0), x=y+\frac{\lambda}{2}$.

3、当 $y>\frac{\lambda}{2}$

x只能有 $x=y-\frac{\lambda}{2}$ 与 x=0 这两种可能取值，因为假设 x<0 则有
$x=y-\frac{\lambda}{2}\operatorname{sign}(x)=y+\frac{\lambda}{2}>\lambda>0$ 假设不成立.

带入$f(x)=(x-y)^2+\lambda|x|$ 有

$$f(0)-f\left(y-\frac{\lambda}{2}\right)=y^2-\left[\left(\frac{\lambda}{2}\right)^2+\lambda\left(y-\frac{\lambda}{2}\right)\right]=\left(y-\frac{\lambda}{2}\right)^2>0 \tag{11}$$

即 $f\left(y-\frac{\lambda}{2}\right)<f(0)$,此时 $x=y-\frac{\lambda}{2}$

4、当$-\frac{\lambda}{2}\leq y\leq\frac{\lambda}{2}$

假设 x<0 则有$x=y-\frac{\lambda}{2}\operatorname{sign}(x)=y+\frac{\lambda}{2}\geq 0$ 假设不成立.
假设 x>0 则有 $x=y-\frac{\lambda}{2}\operatorname{sign}(x)=y-\frac{\lambda}{2}\leq 0$ 假设不成立.
只需证明 $f(\Delta x)>f(0)$ 对于$\Delta x\neq 0$ 成立即可

$$f(\Delta x)=(\Delta x-y)^2+\lambda|\Delta x|=(\Delta x)^2-2\Delta xy+\lambda|\Delta x|+f(0) \tag{12}$$

当$\Delta x>0$ 时利用 $y\leq\frac{\lambda}{2}$

$$\begin{aligned} f(\Delta x) &= (\Delta x)^2-2\Delta xy+\lambda|\Delta x|+f(0) \\ &\geq (\Delta x)^2-2\Delta x\frac{\lambda}{2}+\lambda|\Delta x|+f(0) \\ &= (\Delta x)^2+f(0)>0 \end{aligned} \tag{13}$$

当$\Delta x<0$ 时利用 $y\leq\frac{\lambda}{2}$

$$\begin{aligned} f(\Delta x) &= (\Delta x)^2-2\Delta xy+\lambda|\Delta x|+f(0) \\ &\geq (\Delta x)^2-2\Delta x\frac{\lambda}{2}+\lambda|\Delta x|+f(0) \\ &= (\Delta x)^2+2\lambda|\Delta x|+f(0)>0 \end{aligned} \tag{14}$$

所以 x=0 可得极小值

综上可得

$$x^*=\begin{cases} y+\frac{\lambda}{2}, & y<-\frac{\lambda}{2} \\ 0, & -\frac{\lambda}{2}\leq y\leq\frac{\lambda}{2} \\ y-\frac{\lambda}{2}, & y>\frac{\lambda}{2} \end{cases} \tag{15}$$

$$x^* = \operatorname{sign}(y)\left(|y| - \frac{\lambda}{2}\right)_+ = \mathcal{T}_{\frac{\lambda}{2}}(y) \tag{16}$$

五、

12.3 (条件独立性, conditional independence) 若 $p(A,B|C) = p(A|C)p(B|C)$ 总是成立, 我们称 A 和 B 在给定 C 时条件独立, 记为

$$A \perp B \mid C.$$

A、B 和 C 可以是离散或者连续的, 也可以是单变量或多变量随机变量. 在本题中, 我们使用如图 12.6 所示的各种简单的概率图模型来说明变量间的条件独立性.

在有向图模型 (directed graphical model) 中, 箭头表示直接的依赖关系——一个结点依赖于其亲代结点 (即有箭头指向该结点的那些结点). 例如, 图 12.6a 解读为 C 依赖于 A, B 依赖于 C, 但是 A 不依赖于任何结点, 换言之, 联合密度可以分解为

$$p(A,B,C) = p(A)p(C|A)p(B|C).$$

(a) 对图 12.6a 中的简单情形 1.1, 证明 $A \perp B \mid C$.

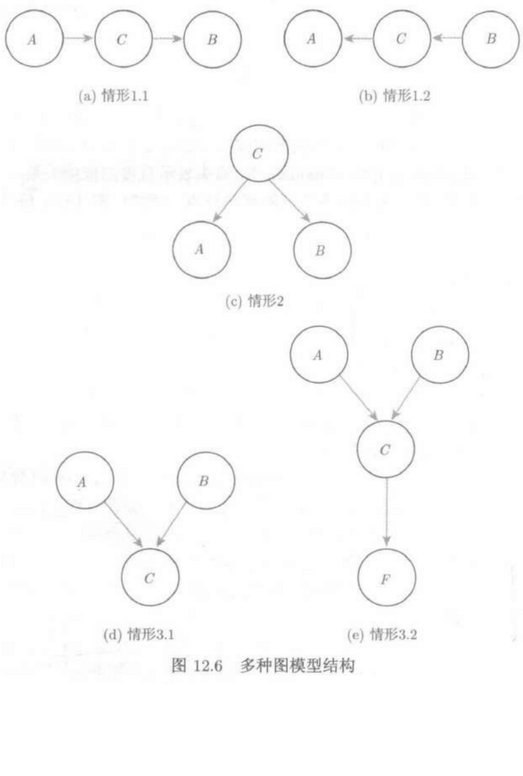
(b) 对图 12.6b 中的简单情形 1.2, 证明 $A \perp B \mid C$.

(c) 对图 12.6c 中的简单情形 2, 证明 $A \perp B \mid C$.

(d) 图 12.6d 中的情形 3.1 还会更麻烦一些. 证明当 C没有被观测到时有 $p(A,B) = p(A)p(B)$, 即 A 和 B 独立. 然而, 当观测到 C 后 A 和 B不是条件独立的. 试着找到一个解释这个现象的直观例子.

这个现象被称为 *explaining away*. 当两个 (或更多个) 起因 (cause) 都可以产生相同的效果时, 在我们观测到那个效果后这些起因将变得彼此依赖 (dependent).

(e) 情形 3.2 是情形 3.1 的一个变体, 如图 12.6e 所示. 直观解释如下事实: 即使 C 没有被观测到, 当 C 的任意一个后代被观测到时 A 和 B 将仍然存在依赖关系.



(a) 情形1.1　(b) 情形1.2　(c) 情形2　(d) 情形3.1　(e) 情形3.2

图 12.6　多种图模型结构

(a)

已知 $p(A,B,C)=p(A)p(C\mid A)p(B\mid C)$

$$\begin{aligned} p(A,B\mid C)=\frac{p(A,B,C)}{p(C)}=\frac{p(A)p(C|A)p(B|C)}{p(C)} \\ p(A\mid C)p(B\mid C)=\frac{p(A,C)}{p(C)}p(B\mid C)=\frac{p(A)p(C|A)p(B|C)}{p(C)} \end{aligned} \tag{17}$$

因此$p(A,B\mid C)=p(A\mid C)p(B\mid C), A\perp B\mid C$

(b)

已知 $p(A,B,C)=p(B)p(C\mid B)p(A\mid C)$

$$\begin{aligned} p(A,B\mid C)=\frac{p(A,B,C)}{p(C)}=\frac{p(B)p(C|B)p(A|C)}{p(C)} \\ p(A\mid C)p(B\mid C)=\frac{p(B,C)}{p(C)}p(A\mid C)=\frac{p(B)p(C|B)p(A|C)}{p(C)} \end{aligned} \tag{18}$$

因此$p(A,B\mid C)=p(A\mid C)p(B\mid C), A\perp B\mid C$

(c)

已知$p(A,B,C)=p(C)p(A\mid C)p(B\mid C)$

$$p(A,B\mid C)=\frac{p(A,B,C)}{p(C)}=\frac{p(C)p(A\mid C)p(B\mid C)}{p(C)}=p(A\mid C)p(B\mid C) \tag{19}$$

因此$p(A,B\mid C)=p(A\mid C)p(B\mid C), A\perp B\mid C$

(d)

已知 $p(A,B,C)=p(C\mid A,B)p(A)p(B)$.
当 C 没有被观测到时有

$$\begin{aligned} p(A,B) &= \sum_C p(A,B,C) \\ &= \sum_C p(C\mid A,B)p(A)p(B) \\ &= p(A)p(B) \end{aligned} \tag{20}$$

explaining away例子

令 A 和 B 独立地遵循 p=0.5 的伯努利分布，令$C=A\oplus B$
在没有观测到 C 时：

- A 和 B 是独立的，可以看作随机抛两次硬币分别决定 A 和 B 的值

当给定 C=0 时：

- 一定有 A=B；

当给定 C=1 时：

- 一定有 A$\neq$B

(e)

F 是 C 的后代，有 $p(F \mid C) \neq p(F)$.

$$p(C \mid F)=\frac{p(C, F)}{p(F)}=\frac{p(C) p(F \mid C)}{p(F)}=p(C) \frac{p(F \mid C)}{p(F)} \neq p(C) \tag{21}$$

即给定 F 的情况下 C 的取值会受到影响

C受影响时， A 和 B 不再独立