模式识别

HW4

201300035 方盛俊 人工智能学院

习题一

(a)

使用 Python 进行调用, 代码为:

```
from liblinear.liblinearutil import *

mnist = svm_read_problem('mnist')
mnist_t = svm_read_problem('mnist.t')

# 使用默认参数训练模型
default_model = train(mnist[0], mnist[1])
# 在测试集上的准确率
print("default_model:")
p_label, p_acc, p_val = predict(mnist_t[0], mnist_t[1], default_model)

# 对每个训练和测试样例的特征值进行开根变换
def sqrt_data(data):
    sqrt_data = ([*data[0]], [*data[1]])
    for i in range(len(sqrt_data[1])):
        sqrt_data[1][i] = {key: sqrt_data[1][i][key] ** 0.5 for key in
sqrt_data[1][i]}
    return sqrt_data

sqrt_mnist = sqrt_data(mnist)
sqrt_mnist_t = sqrt_data(mnist_t)
sqrt_model = train(sqrt_mnist[0], sqrt_mnist[1])
print("sqrt_model:")
sqrt_p_label, sqrt_p_acc, sqrt_p_val = predict(sqrt_mnist_t[0],
sqrt_mnist_t[1], sqrt_model)
```

(b)

使用默认参数训练与测试得到的准确率为:

```
default_model:
Accuracy = 91.53% (9153/10000) (classification)
```

(c)

使用 $x \leftarrow \sqrt{x}$ 变换得到的数据进行训练与测试得到的准确率为:

```
sqrt_model:
Accuracy = 91.37% (9137/10000) (classification)
```

(d)

得到的准确率和后一问的准确率几乎相同, 且为极高的 91.53% 与 91.37%. 据猜测, 可能是软件版本更新导致默认参数发生了变化.

按照正常想法来说, 开根变换后准确率应该会有所上升. 这是因为未经过缩放的原 mnist 数据集的特征取值范围较大, 且不同特征之间的取值范围又不一致, 导致数值较大的特征对最终结果影响较大. 经过开根变换后, 近似相当于将取值范围进行了缩小, 变得接近于经过缩放的数据 (scaled mnist), 因此最终准确率会有所上升.

习题二

(a)

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

(b)

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

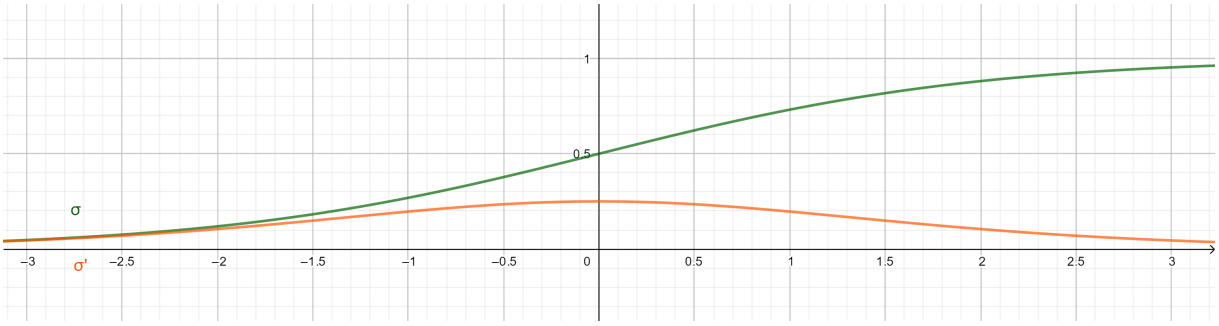

图 1: 绿色线为 $\sigma(x)$ 的曲线, 橙色线为 $\sigma'(x)$ 的曲线.

(c)

第 i 层网络可以表达为 $\boldsymbol{y}^{(i)}=\sigma\left(\boldsymbol{z}^{(i)}\right)=\sigma\left(f\left(\boldsymbol{x}^{(i)},\boldsymbol{\theta}^{(i)}\right)\right)$, 其中 $\boldsymbol{x}^{(i)}$ 是第 i 层的输入, 而 $\boldsymbol{z}^{(i)}=f\left(\boldsymbol{x}^{(i)},\boldsymbol{\theta}^{(i)}\right)$ 是第 i 层网络激活函数 $\sigma(\cdot)$ 前对输入的处理.

因此由链式法则有

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

其中由于我们知道 $\boldsymbol{y}^{(i)}=\sigma\left(\boldsymbol{z}^{(i)}\right)$ 是一个逐元素操作, 因此有 $y_j^{(i)}=\sigma\left(z_j^{(i)}\right)$, 则我们分析 $\frac{\partial \boldsymbol{y}^{(i)}}{\partial\left(\boldsymbol{z}^{(i)}\right)^T}$ 的单个元素, 则有

$$\left[\frac{\partial \boldsymbol{y}^{(i)}}{\partial\left(\boldsymbol{z}^{(i)}\right)^T}\right]_j=\sigma'\left(z_j^{(i)}\right)$$

而由 图 1 我们又知道 $\sigma'\left(z_j^{(i)}\right)$ 是最大值为 0.25 的函数, 且在 $\left|z_j^{(i)}\right| \to \infty$ 时 $\sigma\left(z_j^{(i)}\right) \to 0$.

因此, 在乘上了 $\sigma'\left(z_j^{(i)}\right)$ 的时候, 尤其是当 i 由 L 变为 1 时, 从 $i = L$ 开始到 $i = 1$ 乘上了多个 $\sigma'\left(z_j^{(i)}\right)$, 得到的结果 $\left\|\frac{\partial \ell}{\partial((\boldsymbol{\theta}^{(i)}))^T}\right\|$ 就越趋近于 0, 即梯度消失困难.

习题三

假设有随机变量 Y 满足分布 $p(x)$, 且分布 $p(x)$ 时满足题目条件的分布, 即有 $x \geq 0$ 时 $p(x) > 0$; 当 $x < 0$ 时 $p(x) = 0$. 以及 Y 的均值为 $E[Y] = \int_0^\infty xp(x) = \mu > 0$.

由于 X 是参数为 $\lambda = \frac{1}{\mu}$ 的指数分布, 即有

$$q(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

由不等式 $\mathrm{KL}(p \parallel q) \geq 0$ 可知

$$\begin{aligned}
0 &\leq \mathrm{KL}(p \parallel q) \\
&= \int p(x) \ln \frac{p(x)}{q(x)} \,\mathrm{d}x \\
&= \int p(x) \ln p(x) \,\mathrm{d}x - \int p(x) \ln q(x) \,\mathrm{d}x \\
&= -h(Y) - \int p(x) \ln q(x) \,\mathrm{d}x \\
&= -h(Y) - \int_0^\infty p(x) \ln \lambda e^{-\lambda x} \,\mathrm{d}x \\
&= -h(Y) - \ln \lambda \int_0^\infty p(x) \,\mathrm{d}x + \lambda \int_0^\infty xp(x) \,\mathrm{d}x \\
&= -h(Y) - \ln \lambda + \lambda\mu \\
&= -h(Y) - \ln \lambda \int_0^\infty q(x) \,\mathrm{d}x + \lambda \int_0^\infty xq(x) \,\mathrm{d}x \\
&= -h(Y) - \int_0^\infty q(x) \ln \lambda e^{-\lambda x} \,\mathrm{d}x \\
&= -h(Y) - \int q(x) \ln q(x) \,\mathrm{d}x \\
&= -h(Y) + h(X)
\end{aligned}$$

即有

$$h(X) \geq h(Y)$$

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

习题四

我们先将收缩阈值操作符 (标量情况下) 化为可读性更好的形式, 其中 $\lambda > 0$:

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

为了在 $\lambda > 0$ 的条件下求解优化问题

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

我们令 $F(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{x}\|_1$, 则由范数的定义可知

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

我们又令 $f_i(x_i) = (x_i - y_i)^2 + \lambda |x_i|, i = 1, 2, ..., n$, 由 $f_i(x_i)$ 之间的独立性可知, 只要我们分别求解

$$\arg\min_{x_i} f_i(x_i) = (x_i - y_i)^2 + \lambda |x_i|$$

即可得原优化问题的解.

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

$$x = y - \frac{\lambda}{2} \operatorname{sign}(x), x \neq 0$$

下面我们进行分类讨论:

当 $y < -\frac{\lambda}{2}$ 时,

假设 $x < 0$ 则有 $x = y - \frac{\lambda}{2} \operatorname{sign}(x) = y + \frac{\lambda}{2} < 0$ 假设成立.

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.

因此有 $x = y + \frac{\lambda}{2}$ 与 $x = 0$ 这两种可能取值, 我们带入 $f(x) = (x - y)^2 + \lambda |x|$ 有

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

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

因此此时 $x=y+\frac{\lambda}{2}$.

当 $y>\frac{\lambda}{2}$ 时,

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.

假设 $x>0$ 则有 $x=y-\frac{\lambda}{2}\operatorname{sign}(x)=y-\frac{\lambda}{2}>0$ 假设成立.

因此有 $x=y-\frac{\lambda}{2}$ 与 $x=0$ 这两种可能取值, 我们带入 $f(x)=(x-y)^2+\lambda|x|$ 有

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

即有 $f\left(y-\frac{\lambda}{2}\right)<f(0)$,

因此此时 $x=y-\frac{\lambda}{2}$.

当 $-\frac{\lambda}{2}\leq y\leq\frac{\lambda}{2}$ 时,

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.

我们只需证明 $f(\Delta x)>f(0)$ 对于 $\Delta x\neq 0$ 时成立即可. 由于

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

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

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

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

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

因此此时 $x = 0$ 可得极小值.

综上可得

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

即有

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

习题五

(a)

由于情形 1.1 可以得 $p(A,C)=p(A)p(C|A)$ 与 $p(A,B,C)=p(A)p(C|A)p(B|C)$, 因此有

$$p(A,B|C)=\frac{p(A,B,C)}{p(C)}=\frac{p(A)p(C|A)p(B|C)}{p(C)}$$

$$p(A|C)p(B|C)=\frac{p(A,C)}{p(C)}p(B|C)=\frac{p(A)p(C|A)p(B|C)}{p(C)}$$

因此两条等式相等, 即 $p(A,B|C)=p(A|C)p(B|C)$, 说明有

$$A\perp B\mid C$$

(b)

由于情形 1.2 可以得 $p(B,C)=p(B)p(C|B)$ 与 $p(A,B,C)=p(B)p(C|B)p(A|C)$, 因此有

$$p(A,B|C)=\frac{p(A,B,C)}{p(C)}=\frac{p(B)p(C|B)p(A|C)}{p(C)}$$

$$p(A|C)p(B|C)=\frac{p(B,C)}{p(C)}p(A|C)=\frac{p(B)p(C|B)p(A|C)}{p(C)}$$

因此两条等式相等, 即 $p(A,B|C)=p(A|C)p(B|C)$, 说明有

$$A\perp B\mid C$$

(c)

由于情形 2 可以得 $p(A,B,C)=p(C)p(A|C)p(B|C)$, 因此有

$$p(A,B|C)=\frac{p(A,B,C)}{p(C)}=\frac{p(C)p(A|C)p(B|C)}{p(C)}=p(A|C)p(B|C)$$

即有

$$A\perp B\mid C$$

(d)

由于情形 2 可以得 $p(A,B,C)=p(C|A,B)p(A)p(B)$.

当 C 没有被观测到时有

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

即 A 和 B 独立.

我们可以找到一些简单的例子, 例如令 A 和 B 独立地遵循 $p=0.5$ 的伯努利分布, 而令 $C = A \oplus B$, 即 C 为 A 与 B 的异或.

在没有观测到 C 时, A 和 B 是独立的, 均遵循 $p=0.5$ 的伯努利分布, 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 $A \neq B$, 这时候可以看出, A 和 B 不再是独立的了.

(e)

在我们给定 F 的情况下, 由于 F 是 C 的后代, 不独立, 一般则有 $p(F|C) \neq p(F)$.

因此由贝叶斯公式可得

$$
p(C|F) = \frac{p(C,F)}{p(F)} = \frac{p(C)p(F|C)}{p(F)} = p(C)\frac{p(F|C)}{p(F)} \neq p(C)
$$

即给定 F 的情况下 C 的取值会受到影响, 再由上一问的结果则可知 A 和 B 不再是独立的, 而是存在依赖关系.