现代智能方法: 配套课后题答案解析

己亥年己巳月辛酉日之研究生部讨论班习题解答

作者：罗敏中

组织：中国原子能科学研究院信息中心

时间：May 27, 2019

版本：1.02

中国原子能科学研究院
CNNC CHINA INSTITUTE OF ATOMIC ENERGY

科学是一种生活方式，它只在人们具有信仰自由的时候才能繁荣起来; — 控制论之父罗伯特 · 维纳

目　　录

附录 课后习题答案解析

A.1 练习题

练习 A.1 证明：关于高斯分布的方差的极大似然估计是有偏的。

注意 有同学可能会问这样的问题：为何它会是有偏的，大部分是因为混淆一致性与无偏性。这是两个不同的概念。是否一致非常重要，是否无偏完全不重要。常见的很多（也许可以说是大部分）统计量就是有偏的，比如标准差的点估计，不管分母是 $\sqrt{n}$ 还是 $\sqrt{n-1}$，都是有偏的（后者平方估计方差才是无偏的）。所以有偏不需要问为什么，无偏才需要问为什么。

解 先求方差的最大似然估计：假设样本集 $\mathcal{D}$ 中有 n 个样本：$x_1, x_2, \ldots x_n$；待估计参数为 θ，由于这些样本是独立抽取的，所以有下式成立：

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

为简化计算，使用对数似然函数：

$$l(\theta) = \ln(p(\mathcal{D}|\theta)) = \sum_{k=1}^{n} \ln(p(x_k|\theta)$$

要求其极大值，对其求梯度，梯度为零的地方就是可能的极大值处：

$$\nabla_\theta = \sum_{k=1}^{n} \nabla_\theta \ln(p(x_k|\theta))$$

对于一维的正态分布，有:

$$\ln p(x) = -\frac{1}{2}2\pi\sigma - \frac{1}{2\sigma}(x-\mu)^2$$

假设 μ 已知，估计 σ：

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)^2$$

下证 σ 的有偏性：

$$
\begin{aligned}
E(\hat{\sigma}^2) &= E\{\frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)^2)\} \\
&= E\{\frac{1}{n}\sum_{k=1}^{n}x_k-\mu^2)\} \\
&= \frac{1}{n}\sum_{k=1}^{n}E(x_k^2)-E(\mu^2) \\
&= (\sigma^2+\mu^2)-E\{(\frac{x_1+x_2+\ldots+x_n}{n})^2\} \\
&= (\sigma^2+\mu^2)-\frac{1}{n^2}E(Y^2) \\
&= (\sigma^2+\mu^2)-\frac{1}{n^2}(n^2\mu^2+n\sigma^2) \\
&= \frac{n-1}{n}\sigma^2 \neq \sigma^2
\end{aligned}
$$

得证；

✍ **练习 A.2 (2015 年春, 中科院自动化所考博真题)** 关于神经网络：（1）针对多层前馈神经网络，请给出反向传播算法的工作原理和训练步骤；（2）请分析“在前馈神经网络中，隐含层数越多对分类预测可能产生的影响”。

解 反向传播算法最早出现于 1986 年，用于解决多层神经网络的训练问题，由 Rumelhart 和 Hinton 等人提出，这篇论文当时发表在 «Nature>> 上：

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by back-propagating errors. Nature, 323(99): 533-536, 1986.

代价函数可以表示为：

$$
\begin{aligned}
E_{(i)} &= \frac{1}{2}\left\|\boldsymbol{y}^{(i)}-\boldsymbol{o}^{(i)}\right\| \\
&= \frac{1}{2}\sum_{k=1}^{n_L}\left(y_k^{(i)}-o_k^{(i)}\right)^2
\end{aligned}
$$

权重的更新算法表示为：

$$
\begin{aligned}
W^{(l)} &= W^{(l)}-\mu\frac{\partial E_{total}}{\partial W^{(l)}} \\
&= W^{(l)}-\frac{\mu}{N}\sum_{i=1}^{N}\frac{\partial E_{(i)}}{\partial W^{(l)}}
\end{aligned}
$$

根据链式法则，误差传递可以表示为：

$$
\begin{aligned}
\frac{\partial E}{\partial w_{ij}^{(l)}} &= \frac{\partial E}{\partial z_i^{(l)}}\frac{\partial z_i^{(l)}}{\partial w_{ij}^{(l)}} \\
&= \delta_i^{(l)}\frac{\partial z_i^{(l)}}{\partial w_{ij}^{(l)}} \\
&= \delta_i^{(l)}a_j^{(l-1)}
\end{aligned}
$$

其对应的矩阵形式为:

$$\begin{aligned}\boldsymbol{\delta}^{(L)} &= -\left(\boldsymbol{y}-\boldsymbol{a}^{(L)}\right)\odot f'\left(\boldsymbol{z}^{(L)}\right)\\ \nabla_{\boldsymbol{W}^{(L)}}E &= \boldsymbol{\delta}^{(L)}\left(\boldsymbol{a}^{(L-1)}\right)^{\top}\end{aligned}$$

增加隐层数可以降低网络误差(也有文献认为不一定能有效降低),提高精度,但也使网络复杂化,从而增加了网络的训练时间和出现“过拟合”的倾向。

✍ **练习 A.3 (2008 年秋, 中科院计算所考博真题)** 样本 x 的类别预测后验概率可以表示为 $\mathbb{P}(\omega_i|x),(i=1...M)$,后验概率最大的类别表示为 ω_{max}:

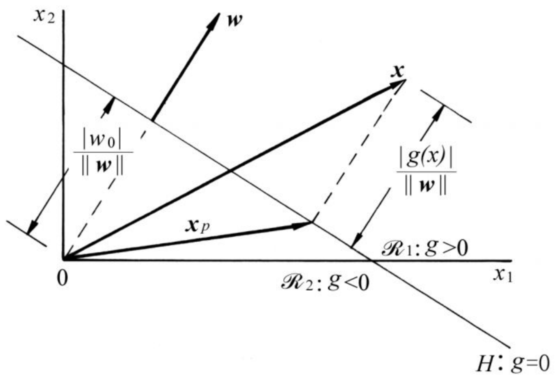
1. 证明 $\mathbb{P}(\omega_{max}|x)\geq 1/M$;
2. 证明最小错误决策的错误率为 $P_E=1-\int\mathbb{P}(\omega_{max}|x)p(x)dx$;

提示:最小错误决策即取后验概率最大;

解 第一小题:用反证法,假设 $\mathbb{P}(\omega_{max}|x)\leq 1/M$,则必存在 $\mathbb{P}(\omega_i|x)\geq 1/M$,则 $\mathbb{P}(\omega_{max}|x)\leq mathbbP(\omega_i|x)$,矛盾;

第二小题:只需证明正确率是 $\int\mathbb{P}(\omega_{max}|x)p(x)dx$,这是显然的;

✍ **练习 A.4** 证明:x 到超平面的投影是 $x_p=x-\frac{g(x)}{\|w\|^2}w$;



解 分情况讨论:

x 在超平面正侧时:$x-x_p=r\frac{w}{\|w\|}=\frac{g(x)}{\|w\|^2}w$,因此 $x_p=x-\frac{g(x)}{\|w\|^2}w$;

x 在超平面负侧时:$x-x_p=-r\frac{w}{\|w\|}=\frac{g(x)}{\|w\|^2}w$,因此 $x_p=x-\frac{g(x)}{\|w\|^2}w$;

A.2 补充习题

✍ **练习 A.5 (神经网络是万能拟合器)** 令 $f:[-1,1]^n\to[-1,1]$ 是 $\rho-$ 利普希茨的,取 $\epsilon>0$,现在构造一个带 sigmoid 激活函数的神经网络 $N:[-1,1]^n\to[-1,1]$,证明:对任意 $\mathbf{x}\in[-1,1]^n$,都有 $|f(\mathbf{x})-N(\mathbf{x})|\leq\epsilon$。

$\rho-$ 利普希茨的定义:

定义 A.1. 利普希茨性

取 $C\subset\mathbb{R}^d$,称函数 $f:\mathbb{R}^d\to\mathbb{R}^k$ 是 $\rho-$ 利普希茨的,若对于任意 $\mathbf{w}_1,\mathbf{w}_2\in C$ 有:

$$\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\| \le \rho \|\mathbf{w}_1 - \mathbf{w}_2\|$$ ♣

解

取 $\epsilon > 0$. 根据提示，我们通过不相交的方框子区域覆盖域 $[-1,1]^n$，以便对于位于同一个方框中的每个 x, x'，我们都有 $|f(\mathbf{x}) - f(\mathbf{x}')| \le \epsilon/2$。由于我们的目标是将 f 的近似值定为 ϵ 的准确度，我们可以从每个方框中选择一个任意点。通过适当地选择一组代表点（例如，选择每个框的中心），我们可以不失一般性地假设对于某些 $\beta \in [0,2]$ 和 $d \in \mathbb{N}$（两者都取决于 ρ 和 ϵ），f 是在离散集 $[-1+\beta, -1+2\beta, \ldots, 1]^d$ 上定义的。从这里开始，证明就很简单了。我们的网络应该有两个隐藏层。第一层有 $(2/\beta)^d$ 个节点，这些节点对应于构成我们框区域的间隔。我们可以调整输入和隐藏层之间的权重，这样给定一个输入 $\mathbf{X}$，如果 $\mathbf{X}$ 的相应坐标位于相应的区间内，每个神经元的输出足够接近 1（注意给定一个有限的区间，我们可以使用 sigmoid 函数来近似拟合目标函数）。在下一层，我们为每个子区域构建一个神经元，并添加一个额外的神经元，输出常量 $-1/2$。如果 $\mathbf{X}$ 属于相应的子区域，我们可以调整权重，使每个神经元的输出为 1，否则为 0。最后，我们可以轻松调整第二层和输出层之间的权重，以便获得所需的输出（例如，达到 $\epsilon/2$ 的精度）。

✍ **练习 A.6 (随机梯度下降算法的一个界)** 对于训练集 $\mathbf{v}_1, \ldots, \mathbf{v}_T$，对于初始权重 $\mathbf{w}^{(1)} = 0$ 和更新算法：$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$，有：

$$\sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t \right\rangle \le \frac{\|\mathbf{w}^{\star}\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2$$

解

按内积规则可得：

$$\begin{aligned}
\left\langle \mathbf{w}^{(t)} - \mathbf{w}^{*}, \mathbf{v}_t \right\rangle &= \frac{1}{\eta} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^{*}, \eta \mathbf{v}_t \right\rangle \\
&= \frac{1}{2\eta} \left(-\left\| \mathbf{w}^{(t)} - \mathbf{w}^{*} - \eta \mathbf{v}_t \right\|^2 + \left\| \mathbf{w}^{(t)} - \mathbf{w}^{*} \right\|^2 + \eta^2 \|\mathbf{v}_t\|^2 \right) \\
&= \frac{1}{2\eta} \left(-\left\| \mathbf{w}^{(t+1)} - \mathbf{w}^{*} \right\|^2 + \left\| \mathbf{w}^{(t)} - \mathbf{w}^{*} \right\|^2 \right) + \frac{\eta}{2} \|\mathbf{v}_t\|^2
\end{aligned}$$

按下标 t 累加可得：

$$\sum_{t=1}^{T} \left\langle \mathbf{w}^{(t)} - \mathbf{w}^{*}, \mathbf{v}_t \right\rangle = \frac{1}{2\eta} \sum_{t=1}^{T} \left(-\left\| \mathbf{w}^{(t+1)} - \mathbf{w}^{*} \right\|^2 + \left\| \mathbf{w}^{(t)} - \mathbf{w}^{*} \right\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|\mathbf{v}_t\|^2$$

使用裂项相消的技巧可得：

$$\left\| \mathbf{w}^{(1)} - \mathbf{w}^{*} \right\|^2 - \left\| \mathbf{w}^{(T+1)} - \mathbf{w}^{*} \right\|^2$$

代入即可得到:

$$
\begin{aligned}
\sum_{t=1}^{T}\left\langle \mathbf{w}^{(t)}-\mathbf{w}^{*},\mathbf{v}_t\right\rangle &= \frac{1}{2\eta}\left(\left\|\mathbf{w}^{(1)}-\mathbf{w}^{*}\right\|^2-\left\|\mathbf{w}^{(T+1)}-\mathbf{w}^{*}\right\|^2\right)+\frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2\\
&\le \frac{1}{2\eta}\left\|\mathbf{w}^{(1)}-\mathbf{w}^{*}\right\|^2+\frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2\\
&= \frac{1}{2\eta}\|\mathbf{w}^{*}\|^2+\frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2
\end{aligned}
$$