人工智能导论第二次作业

张立博 2021012487

2023 年 5 月 17 日

1 第一题

(1)

若只划分训练集和测试集，则无法进行有效调参。因为如果使用测试集进行调参，就会导致信息泄露；若使用训练集进行训练加调参，得到的模型可能在训练集上表现很好，但在测试集或新数据上表现很差。因此需要引入验证集进行参数调优。
引入验证集后，验证集用于模型选择和参数调整，而测试集用于模型性能的最终评估，永远不能用来训练模型或给模型调参。

(2)

K 折交叉验证的流程如下：

1. 将数据集分为 K 个子集，其中 K-1 个子集作为训练集，1 个子集作为验证集
2. 在第一轮交叉验证中，使用 K-1 个子集进行训练，使用剩余的 1 个子集进行模型验证，记录模型在验证集上的性能指标
3. 重复步骤 2，将不同的子集作为验证集，其余子集作为训练集，直到每个子集都被用作验证集一次，最终记录所有 K 次验证结果的平均值作为模型的性能指标

(3)

L1 正则化可以得到稀疏解，将权重向量中较小的权重置为 0，可以自动特征选择; 而 L2 得到的权重向量是密集的
另一种正则化方法：Dropout 正则化，在神经网络中应用的一种正则化技术，它随机地在训练过程中关闭一部分神经元，从而减少了神经元之间的依赖关系，降低了过拟合风险。

(4)

在使用 SVM 处理非线性数据时一开始需要引入基函数，但往往难以确定使用哪些基函数，并且使用基函数将原始空间映射到高维空间非常耗时。所以需要引入核函数，其可以在不显示计算的情况下将数据从原始空间映射到高维空间，降低计算复杂度，使 SVM 可以更高效地处理线性不可分问题

(5)

主要使用两种方法:
1. 对数据集进行随机采样 (如自助法) 得到不同样本，用这些样本训练决策树
2. 决策树生长过程中，对于每次分裂在可用的特征中随机选取 k 个特征进行决策

2 第二题

(1)

改进模型或者训练过程的措施

1. 使用引入冲量的随机梯度下降方法 (SGD with Momentum)
   在更新参数的过程中使用冲量 (历史梯度的指数滑动平均) 而不是直接使用梯度

2. 使用学习率衰减技术 (Learning Rate Decay)
   初始时使用较大的学习率来快速收敛，后期逐渐减小学习率以保证训练的稳定性

(2)

1. 局部连接 (Local Connectivity)

2. 参数共享 (Parameter Sharing)

(3)

LeNet 网络共有 7 层，卷积层和全连接层包含可学习的参数
若不考虑偏置参数

第一层，卷积层

可学习参数量 $= 5 \cdot 5 \cdot 1 \cdot 6 = 150$

第二层，池化层

可学习参数量 $= 0$

第三层，卷积层

可学习参数量 $= 5 \cdot 5 \cdot 6 \cdot 16 = 2400$

第四层，池化层

可学习参数量 $= 0$

第五层，全连接层

可学习参数量 $= 5 \cdot 5 \cdot 16 \cdot 120 = 48000$

第六层，全连接层

可学习参数量 $= 120 \cdot 84 = 10080$

第七层，全连接层

可学习参数量 $= 84 \cdot 10 = 840$

(4)

ResNet（Residual Neural Network）在模型架构上提出了残差连接（Residual Connections）的改进。
在残差连接中，输入信号可以直接跨过一个或多个层级，与后续的层级进行相加，使得信息可以更快地传递。这样一来，即使在网络非常深的情况下，梯度可以更容易地向后传播，从而加速了模型的训练过程。

(5)

1. 梯度裁剪 (Gradient Clipping)
2. 层归一化 (Layer Normalization)

3 第三题

(1) 计算另外五个属性的信息增益

根蒂

$$
\begin{aligned}
H(D) &= -(\frac{8}{17}log\frac{8}{17} + \frac{9}{17}log\frac{9}{17}) = 0.998 \\
H(D_1) &= -(\frac{5}{8}log\frac{5}{8} + \frac{3}{8}log\frac{3}{8}) = 0.955 \\
H(D_2) &= -(\frac{3}{7}log\frac{3}{7} + \frac{4}{7}log\frac{4}{7}) = 0.985 \\
H(D_3) &= -(\frac{0}{2}log\frac{0}{2} + \frac{2}{2}log\frac{2}{2}) = 0.000 \\
IG(D) &= H(D) - (\frac{8}{17}H(D_1) + \frac{7}{17}H(D_2) + \frac{2}{17}H(D_3)) = 0.143
\end{aligned}
$$

敲声

$$
\begin{aligned}
H(D) &= -(\frac{8}{17}log\frac{8}{17} + \frac{9}{17}log\frac{9}{17}) = 0.998 \\
H(D_1) &= -(\frac{6}{10}log\frac{6}{10} + \frac{4}{10}log\frac{4}{10}) = 0.971 \\
H(D_2) &= -(\frac{2}{5}log\frac{2}{5} + \frac{3}{5}log\frac{3}{5}) = 0.971 \\
H(D_3) &= -(\frac{0}{2}log\frac{0}{2} + \frac{2}{2}log\frac{2}{2}) = 0.000 \\
IG(D) &= H(D) - (\frac{10}{17}H(D_1) + \frac{5}{17}H(D_2) + \frac{2}{17}H(D_3)) = 0.141
\end{aligned}
$$

纹理

$$
\begin{aligned}
H(D) &= -(\frac{8}{17}log\frac{8}{17} + \frac{9}{17}log\frac{9}{17}) = 0.998 \\
H(D_1) &= -(\frac{7}{9}log\frac{7}{9} + \frac{2}{9}log\frac{2}{9}) = 0.764 \\
H(D_2) &= -(\frac{1}{5}log\frac{1}{5} + \frac{4}{5}log\frac{4}{5}) = 0.722 \\
H(D_3) &= -(\frac{0}{3}log\frac{0}{3} + \frac{3}{3}log\frac{3}{3}) = 0.000 \\
IG(D) &= H(D) - (\frac{9}{17}H(D_1) + \frac{5}{17}H(D_2) + \frac{3}{17}H(D_3)) = 0.381
\end{aligned}
$$

脐部

$$
\begin{aligned}
H(D) &= -(\frac{8}{17}log\frac{8}{17} + \frac{9}{17}log\frac{9}{17}) = 0.998 \\
H(D_1) &= -(\frac{5}{7}log\frac{5}{7} + \frac{2}{7}log\frac{2}{7}) = 0.863 \\
H(D_2) &= -(\frac{3}{6}log\frac{3}{6} + \frac{3}{6}log\frac{3}{6}) = 1.000 \\
H(D_3) &= -(\frac{0}{4}log\frac{0}{4} + \frac{4}{4}log\frac{4}{4}) = 0.000 \\
IG(D) &= H(D) - (\frac{7}{17}H(D_1) + \frac{6}{17}H(D_2) + \frac{4}{17}H(D_3)) = 0.290
\end{aligned}
$$

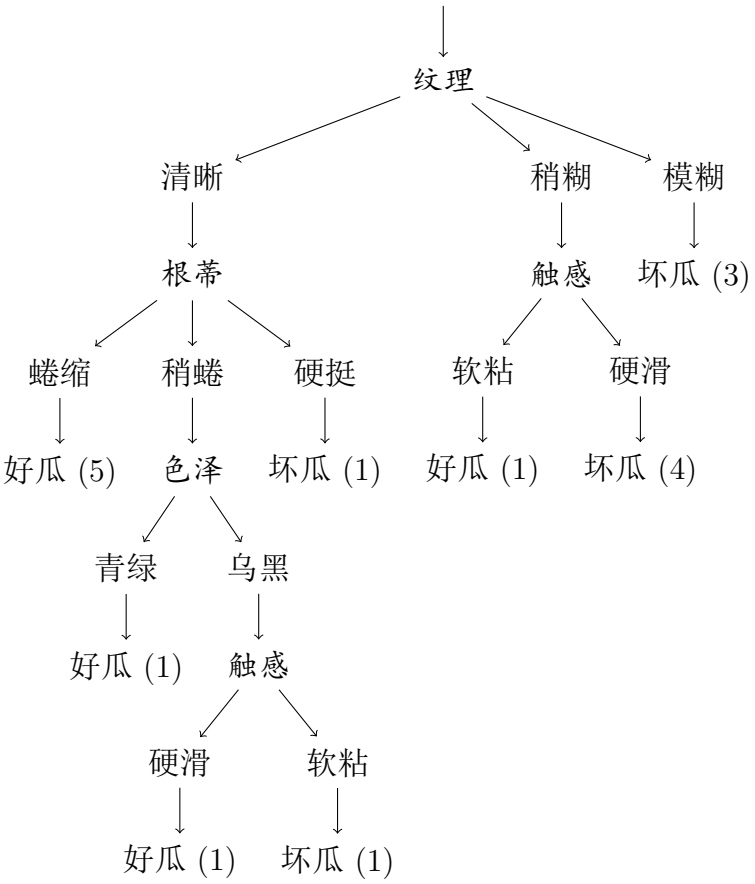
触感

$$
\begin{aligned}
H(D) &= -(\frac{8}{17}log\frac{8}{17} + \frac{9}{17}log\frac{9}{17}) = 0.998 \\
H(D_1) &= -(\frac{6}{8}log\frac{6}{8} + \frac{2}{8}log\frac{2}{8}) = 0.811 \\
H(D_2) &= -(\frac{6}{9}log\frac{6}{9} + \frac{3}{9}log\frac{3}{9}) = 0.918 \\
IG(D) &= H(D) - (\frac{8}{17}H(D_1) + \frac{9}{17}H(D_2)) = 0.130
\end{aligned}
$$

(2) 使用 ID3 算法建立决策树

由 (1) 得第一次选择时增益最大的属性为纹理，之后的每次选择与 (1) 同理，计算每个属性的信息增益并选择信息增益最大的属性，若有多个最大增益则选择排序靠前的

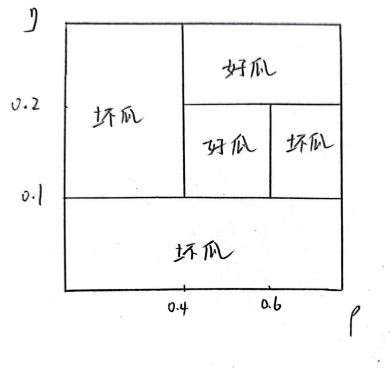
建立的决策树如下



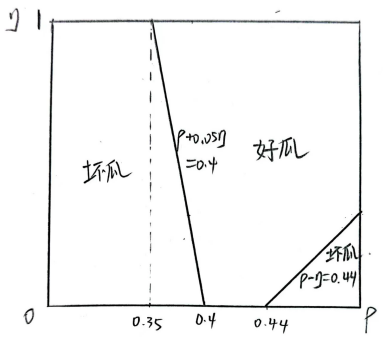
决策树中斜体字表示属性，叶节点后数字表示该叶节点代表的样本数量，总和为 17

(3) 两棵决策树的决策面

决策树 1



决策树 2



4 第四题

(1)

首先计算 ℓ 对推理结果向量 $\vec{o}$ 逐元素的导数

$$\frac{\partial \ell}{\partial o_k} = \frac{\partial(\sum_{j=1}^{K}(-y_i log \hat{y}_j))}{\partial o_k}$$

由链式法则得到

$$\begin{aligned}
&\frac{\partial(\sum_{j=1}^{K}(-y_j log \hat{y}_j))}{\partial o_k} \\
=&\frac{\partial(\sum_{j=1}^{K}(-y_j log \hat{y}_j))}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial o_k} \\
=&-\sum_{j=1}^{K} \frac{y_j}{\hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial o_k}
\end{aligned}$$

当 j = k 时

$$\begin{aligned}
&\frac{\partial \hat{y}_j}{\partial o_k} \\
=&\frac{\partial \hat{y}_k}{\partial o_k} \\
=&\frac{\partial(\frac{e^{o_k}}{e^{o_1}+...+e^{o_K}})}{\partial o_k} \\
=&\hat{y}_k \cdot (1-\hat{y}_k)
\end{aligned}$$

当 j != k 时

$$
\begin{aligned}
&\frac{\partial \hat{y_j}}{\partial o_k} \\
=&\frac{\partial(\frac{e^{o_j}}{e^{o_1}+\ldots+e^{o_K}})}{\partial o_k} \\
=&-\hat{y_j}\cdot\hat{y_k}
\end{aligned}
$$

所以原式

$$
\begin{aligned}
=&-\sum_{j=1}^{K}\frac{y_j}{\hat{y_j}}\cdot\frac{\partial \hat{y_j}}{\partial o_k} \\
=&-\frac{y_k}{\hat{y_k}}\cdot\frac{\partial \hat{y_k}}{\partial o_k}-\sum_{j!=k}^{K}\frac{y_j}{\hat{y_j}}\cdot\frac{\partial \hat{y_j}}{\partial o_k} \\
=&-y_k\cdot(1-\hat{y_k})+\sum_{j!=k}^{K}y_j\cdot\hat{y_k} \\
=&-y_k+\sum_{j=1}^{K}y_j\cdot\hat{y_k}
\end{aligned}
$$

因为 $y_k=1, y_j=0, \forall y\neq k$
所以 $\sum_{j=1}^{K}y_j=1$

$$
\frac{\partial \ell}{\partial o_k}=\hat{y_k}-y_k
$$

故

$$
\frac{\partial \ell}{\partial \vec{o}}=\begin{bmatrix}\frac{\partial \ell}{\partial o_1}\\ \frac{\partial \ell}{\partial o_2}\\ \vdots \\ \frac{\partial \ell}{\partial o_K}\end{bmatrix}=\begin{bmatrix}\hat{y_1}-y_1\\ \hat{y_2}-y_2\\ \vdots \\ \hat{y_K}-y_K\end{bmatrix}
$$

(2)

根据链式法则和矩阵求导术

$$
\begin{aligned}
&\frac{\partial \ell}{\partial W} \\
=&\frac{\partial \ell}{\partial \vec{o}} \cdot \frac{\partial \vec{o}}{\partial W} \\
=&\frac{\partial \ell}{\partial \vec{o}} \cdot \frac{\partial (W^T \vec{x})}{\partial W} \\
=&x \cdot (\frac{\partial \ell}{\partial \vec{o}})^T \\
=&\begin{bmatrix} x_1 \cdot (\hat{y_1} - y_1) & ... & x_1 \cdot (\hat{y_K} - y_K) \\ x_2 \cdot (\hat{y_1} - y_1) & ... & x_2 \cdot (\hat{y_K} - y_K) \\ \vdots & & \\ x_D \cdot (\hat{y_1} - y_1) & ... & x_D \cdot (\hat{y_K} - y_K) \end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \ell}{\partial \vec{x}} \\
=&\frac{\partial \ell}{\partial \vec{o}} \cdot \frac{\partial \vec{o}}{\partial \vec{x}} \\
=&\frac{\partial \ell}{\partial \vec{o}} \cdot \frac{\partial (W^T \vec{x})}{\partial \vec{x}} \\
=&W \cdot \frac{\partial \ell}{\partial \vec{o}} \\
=&\begin{bmatrix} w_{11} \cdot (\hat{y_1} - y_1) + ... + w_{1K} \cdot (\hat{y_K} - y_K) \\ w_{21} \cdot (\hat{y_1} - y_1) + ... + w_{2K} \cdot (\hat{y_K} - y_K) \\ \vdots \\ w_{D1} \cdot (\hat{y_1} - y_1) + ... + w_{DK} \cdot (\hat{y_K} - y_K) \end{bmatrix}
\end{aligned}
$$

(3)

首先求 $\frac{\partial \ell}{\partial \vec{z}}$

由链式法则及已知得

$$
\begin{aligned}
&\frac{\partial \ell}{\partial \vec{z}} \\
=&\frac{\partial \ell}{\partial \vec{o}} \cdot \frac{\partial \vec{o}}{\partial \vec{h}} \cdot \frac{\partial \vec{h}}{\partial \vec{z}} \\
=&W_2 \cdot \frac{\partial \ell}{\partial \vec{o}} \cdot \frac{\partial \sigma(z)}{\partial z} \\
=&W_2 \cdot \frac{\partial \ell}{\partial \vec{o}} \cdot (\sigma(z) \cdot (1 - \sigma(z)))
\end{aligned}
$$

然后求 $\frac{\partial \ell}{\partial W_1}$

$$
\begin{aligned}
&\frac{\partial \ell}{\partial W_1} \\
=&\frac{\partial \ell}{\partial \vec{z}} \cdot \frac{\partial \vec{z}}{\partial W_1} \\
=&\vec{x} \cdot (W_2 \cdot \frac{\partial \ell}{\partial \vec{o}})^T \cdot (\sigma(z) \cdot (1 - \sigma(z)))
\end{aligned}
$$

(4)

经过调参，选择学习率 $= 0.2$，网络的训练集、验证集和测试集的准确率如下

```
Top-1 accuracy on the training set 0.9982222222222222
Top-1 accuracy on the validation set 0.957
Top-1 accuracy on the test set 0.9498
```

表 1: Accuracy on Training, Validation, and Test Sets

Dataset	Training Set	Validation Set	Test Set
	Accuracy (%)	Accuracy (%)	Accuracy (%)
mlp	99.82	95.7	94.98

损失函数的训练曲线如下

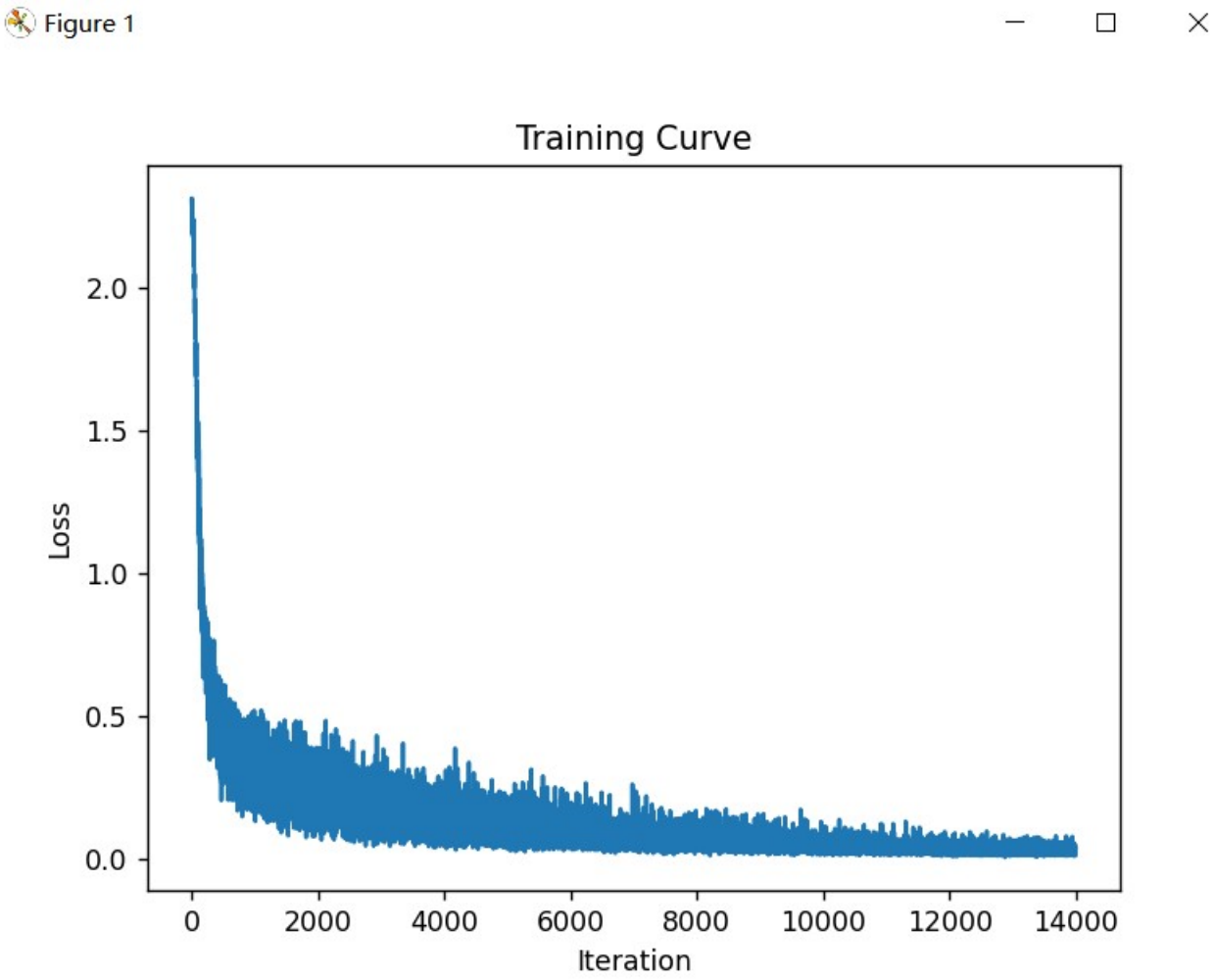


Figure 1
Training Curve
Loss
2.0
1.5
1.0
0.5
0.0
0
2000
4000
6000
8000
10000
12000
14000
Iteration