机器学习导论
习题四

学号：141130077
作者姓名：邱梓豪
邮箱：2957606241@qq.com

2017 年 5 月 17 日

1 [20pts] Reading Materials on CNN

卷积神经网络(Convolution Neural Network,简称CNN)是一类具有特殊结构的神经网络，在深度学习的发展中具有里程碑式的意义。其中，Hinton于2012年提出的AlexNet可以说是深度神经网络在计算机视觉问题上一次重大的突破。

关于AlexNet的具体技术细节总结在经典文章“ImageNet Classification with Deep Convolutional Neural Networks”, by Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton in NIPS'12，目前已逾万次引用。在这篇文章中，它提出使用ReLU作为激活函数，并创新性地使用GPU对运算进行加速。请仔细阅读该论文，并回答下列问题(请用1-2句话简要回答每个小问题，中英文均可)。

(a) **[5pts]** Describe your understanding of how ReLU helps its success? And, how do the GPUs help out?

(b) **[5pts]** Using the average of predictions from several networks help reduce the error rates. Why?

(c) **[5pts]** Where is the dropout technique applied? How does it help? And what is the cost of using dropout?

(d) **[5pts]** How many parameters are there in AlexNet? Why the dataset size(1.2 million) is important for the success of AlexNet?

关于CNN，推荐阅读一份非常优秀的学习材料，由南京大学计算机系吴建鑫教授[1]所编写的讲义Introduction to Convolutional Neural Networks[2]，本题目为此讲义的Exercise-5，已获得吴建鑫老师授权使用。

[1]吴建鑫教授主页链接为cs.nju.edu.cn/wujx

[2]由此链接可访问讲义https://cs.nju.edu.cn/wujx/paper/CNN.pdf

Solution. 此处用于写解答(中英文均可)
（a）我对ReLU的理解如下：首先ReLU易于求导，所以可以提升训练速度；其次ReLU作为激活函数可以消除“梯度消失”的现象，也可以使得训练的效果更好。
GPU在训练CNN使起了很大的作用，因为GPU 相比 CPU 可以提供更多数量的核，而且当前的GPU适合跨GPU并行化，并可以读写其他GPU的内存，这使得CNN的神经元放在不同GPU上，从而加快训练速度。
（b）论文中提到使用使用若干网络的预测结果做平均可以降低错误率，但转而又说这样的时间代价太高，故没有采用。我觉得这种方法可以从直观上进行理解，因为每个网络的预测结果可以看做是在真实结果的基础上加上了一个随机的偏差，所以利用平均可以减弱这种随机偏差的影响，从而提高正确率。
（c）dropout在这里是用以防止过拟合的，具体的方法是在模型训练时随机（概率为0.5）将隐层神经元的输出设为0。从直观上来理解，这相当于减少了每个神经元的迭代次数，因而有利于减弱过拟合。
（d）AlexNet有6000万的参数，65万个神经元。很大的数据集能够防止严重的过拟合，这也是AlexNet成功的原因之一。

2 [20pts] Kernel Functions

(1) 试通过定义证明以下函数都是一个合法的核函数：

(i) **[5pts]** 多项式核: $\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^d$;

(ii) **[10pts]** 高斯核：$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2})$, 其中$\sigma > 0$.

(2) **[5pts]** 试证明$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{1+e^{-\mathbf{x}_i^T\mathbf{x}_j}}$不是合法的核函数。

Proof. 此处用于写证明(中英文均可)
（1）先证明多项式核是一个合法的核函数。由核函数的定义$\kappa(\mathbf{x}_i, \mathbf{x}_j)$=$\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$可知，如果能找到这样的映射函数$\phi(\mathbf{x})$，那么就相当于证明了这个核函数是合法的。
对于多项式核而言，可对d从1开始归纳：
d=1时，很明显有$\phi(\mathbf{x_i}) = \mathbf{x_i}$，此时$\kappa(\mathbf{x}_i, \mathbf{x}_j)$=$\mathbf{x}_i^T\mathbf{x}_j$=$\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$
d=2时，此时有$\phi(\mathbf{x_i}) = \mathbf{x_{i}}_\alpha\mathbf{x_{i}}_\beta, 1 \le \alpha, \beta \le n$，因此$\kappa(\mathbf{x}_i, \mathbf{x}_j)$=$(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^2$=$\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$
所以，多项式核的映射函数为$\phi(\mathbf{x_i}) = \mathbf{x_{i}}_\alpha\mathbf{x_{i}}_\beta\cdots\mathbf{x_{i}}_\gamma, 1 \le \alpha, \beta\cdots\gamma \le n$，所以多项式核是合法的核函数。
下面再证高斯核是一个合法的核函数，即对任意向量$\mathbf{x}$，$\mathbf{y}$，$\kappa(\mathbf{x}, \mathbf{y})$可以表示成$\phi(\mathbf{x}^T\phi(\mathbf{y}$的形式。
因为：

$$\|\mathbf{x}-\mathbf{y}\|^2 = \mathbf{x}^T\mathbf{x} + \mathbf{y}^T\mathbf{y} - 2\mathbf{x}^T\mathbf{y} \tag{2.1}$$

所以有：

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2})\exp(-\frac{\mathbf{y}^T\mathbf{y}}{2\sigma^2})\exp(\frac{\mathbf{x}^T\mathbf{y}}{\sigma^2}) \tag{2.2}$$

对$\exp(\frac{\mathbf{x}^T\mathbf{y}}{\sigma^2})$泰勒展开可得：

$$\begin{aligned}\exp(\frac{\mathbf{x}^T\mathbf{y}}{\sigma^2}) &= \sum_{n=0}^{+\infty}\frac{(\mathbf{x}^T\mathbf{y})^T}{n!\sigma^{2n}} \\ &= \sum_{n=0}^{+\infty}\frac{1}{\sqrt{n!}}\frac{(\mathbf{x}^T)^n}{\sigma^n}\frac{1}{\sqrt{n!}}\frac{(\mathbf{y}^T)^n}{\sigma^n} \\ &= \varphi(x)^T\varphi(y)\end{aligned}$$

其中：

$$\varphi(x) = \sum_{n=0}^{+\infty}\frac{1}{\sqrt{n!}}\frac{(\mathbf{x}^T)^n}{\sigma^n} \tag{2.3}$$

所以：

$$\begin{aligned}\kappa(\mathbf{x},\mathbf{y}) &= \exp(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2})\exp(-\frac{\mathbf{y}^T\mathbf{y}}{2\sigma^2})\exp(\frac{\mathbf{x}^T\mathbf{y}}{\sigma^2}) \\ &= \exp(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2})\varphi(x)^T\varphi(y)\exp(-\frac{\mathbf{y}^T\mathbf{y}}{2\sigma^2}) \\ &= \phi(\mathbf{x}^T\phi(\mathbf{y}\end{aligned}$$

其中：

$$\phi(\mathbf{x}) = \exp(-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2})\varphi(x) \tag{2.4}$$

由于找到了高斯核所对应的映射函数，所以高斯核是合法的核函数。

（2）首先$\kappa(\mathbf{x}_i,\mathbf{x}_j) = \frac{1}{1+e^{-\mathbf{x}_i^T\mathbf{x}_j}}$是对称的，所以要证明它不是合法的核函数，就是证明其对应的核矩阵非半正定。我这里通过一个反例来说明这点：设$\mathbf{x}_i = [1,0]$，$\mathbf{x}_j = [1,0]$，则这时核矩阵如下：

$$\mathbf{M} = \begin{bmatrix}\frac{1}{1+e^{-1}} & 1\\ 1 & 1\end{bmatrix} = \begin{bmatrix}\frac{e}{1+e} & 1\\ 1 & 1\end{bmatrix} \tag{2.5}$$

再求$\mathbf{M}$的特征值：

$$|\lambda\mathbf{E}-\mathbf{M}| = (\lambda - \frac{e}{e+1})(\lambda-1) - 1 = \lambda^2 - \frac{2e+1}{e+1}\lambda - \frac{1}{e+1} = 0 \tag{2.6}$$

观察上面的二次函数可知，该二次函数对应的二次曲线开口向上，且与y轴的截距小于0，所以必然有一个根小于0，即有小于0的特征值。因为原核矩阵对称，所以有小于0的特征值就意味着该核矩阵非半正定。所以该核函数不是合法的核函数。

□

3 [25pts] SVM with Weighted Penalty

考虑标准的SVM优化问题如下(即课本公式(6.35))，

$$\begin{aligned}\min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1-\xi_i \\ & \xi_i \geq 0, i = 1,2,\cdots,m.\end{aligned} \tag{3.1}$$

注意到，在(3.1)中，对于正例和负例，其在目标函数中分类错误的“惩罚”是相同的。在实际场景中，很多时候正例和负例错分的“惩罚”代价是不同的，比如考虑癌症诊断，将一个确实患有癌症的人误分类为健康人，以及将健康人误分类为患有癌症，产生的错误影响以及代价不应该认为是等同的。

现在，我们希望对负例分类错误的样本(即false positive)施加$k > 0$倍于正例中被分错的样本的“惩罚”。对于此类场景下，

(1) **[10pts]** 请给出相应的SVM优化问题;

(2) **[15pts]** 请给出相应的对偶问题，要求详细的推导步骤，尤其是如KKT条件等。

Solution. 此处用于写解答(中英文均可)

(1)设对正例的惩罚系数为C_p，则对负例的惩罚系数为kC_p。设一共有m个样本，其中正例有n个，则得到的SVM优化问题如下：

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_p \sum_{i\in Pos} \xi_i + kC_p \sum_{i\in Neg} \xi_i = \frac{1}{2}\|\mathbf{w}\|^2 + C_p \sum_{i=1}^{n} \xi_i + kC_p \sum_{i=1}^{m-n} \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m. \end{aligned} \tag{3.2}$$

(2)可利用拉格朗日乘子法得到式(3.2)的拉格朗日函数：

$$L(\mathbf{w}, b, \alpha, \xi, \mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C_p \sum_{i=1}^{n} \xi_i + kC_p \sum_{i=1}^{m-n} \xi_i + \sum_{i=1}^{m} \alpha_i(1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)) - \sum_{i=1}^{m} \mu_i \xi_i \tag{3.3}$$

其中$\alpha_i \geq 0$，$\mu_i \geq 0$是拉格朗日乘子。

令$L(\mathbf{w}, b, \alpha, \xi, \mu)$对$\mathbf{w}$，b，$\xi_i$的偏导为0可得：

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \tag{3.4}$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i \tag{3.5}$$

$$nC_p + (m-n)kC_p = \sum_{i=1}^{m} (\alpha_i + \mu_i) \tag{3.6}$$

要满足KKT条件如下：

$$\begin{cases} \alpha_i \geq 0, \quad \mu_i \geq 0, \\ y_i(\mathbf{x}_i + b) - 1 + \xi_i \geq 0, \\ \alpha_i(y_i(\mathbf{x}_i + b) - 1 + \xi_i) = 0, \\ \xi_i \geq 0, \quad \mu_i \xi_i = 0 \end{cases} \tag{3.7}$$

将式(3.4)(3.5)(3.6)带入式(3.3)可得：

$$
\begin{aligned}
(3.3) &= \frac{1}{2}\|\mathbf{w}\|^2 + C_p\sum_{i=1}^{n}\xi_i + kC_p\sum_{i=1}^{m-n}\xi_i + \sum_{i=1}^{m}\alpha_i(1-\xi_i - y_i(\mathbf{w}^T\mathbf{x}_i+b)) - \sum_{i=1}^{m}\mu_i\xi_i \\
&= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_iy_j\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j + nC_p\xi_i + (m-n)kC_p\xi_i + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}\alpha_iy_i\mathbf{w}^T\mathbf{x}_i + 0 - \sum_{i=1}^{m}\mu_i\xi_i \\
&= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_iy_j\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j + nC_p\xi_i + (m-n)kC_p\xi_i + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}\sum_{j=1}^{m}y_iy_j\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{m}\mu_i\xi_i \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_iy_j\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j + nC_p\xi_i + (m-n)kC_p\xi_i + \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\alpha_i\xi_i - \sum_{i=1}^{m}\mu_i\xi_i \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_iy_j\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{m}\alpha_i
\end{aligned}
$$

(3.2)的对偶问题如下：

$$
\begin{aligned}
\max_{\alpha} \quad & -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_iy_j\alpha_i\alpha_j\mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{m}\alpha_i \\
\text{s.t.} \quad & \sum_{i=1}^{m}\alpha_iy_i = 0 \\
& 0 \le \alpha_i \\
& \sum_{i=1}^{m}\alpha_i \le nC_p + (m-n)kC_p, i = 1, 2, \cdots, m.
\end{aligned} \tag{3.8}
$$

对偶问题中第三个约束条件出现的原因是因为首先有：

$$
nC_p + (m-n)kC_p = \sum_{i=1}^{m}(\alpha_i + \mu_i) \tag{3.9}
$$

又因为$\mu_i \ge 0$，所以α_i要满足此约束条件。

4 [35pts] SVM in Practice - LIBSVM

支持向量机(Support Vector Machine，简称SVM)是在工程和科研都非常常用的分类学习算法。有非常成熟的软件包实现了不同形式SVM的高效求解，这里比较著名且常用的如LIBSVM[3]。

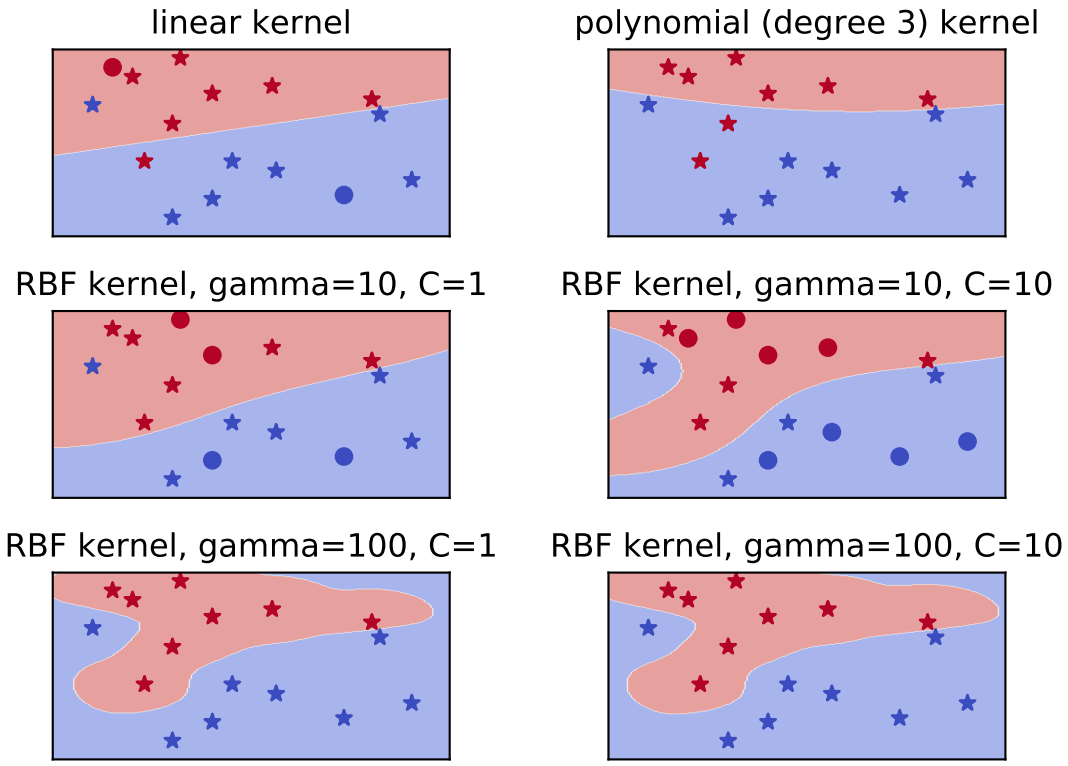
(1) **[20pts]** 调用库进行SVM的训练，但是用你自己编写的预测函数作出预测。

(2) **[10pts]** 借助我们提供的可视化代码，简要了解绘图工具的使用，通过可视化增进对SVM各项参数的理解。详细编程题指南请参见链接：`http://lamda.nju.edu.cn/ml2017/PS4/ML4_programming.html`.

(3) **[5pts]** 在完成上述实践任务之后，你对SVM及核函数技巧有什么新的认识吗？请简要谈谈。

[3]LIBSVM主页课参见链接：`https://www.csie.ntu.edu.tw/~cjlin/libsvm/`

Solution. 此处用于写解答(中英文均可)
对第二小题中的利用SVM分类的结果的可视化如下：



在上图中，五角星代表的点就是支持向量，圆圈代表的点是其他点。
从图中可以看出，对于较为复杂，难以线性划分的情况，还是使用RBF作为核函数效果较好。
在利用SVM进行训练的时候可以指定gamma和C两个参数。C是对不满足约束的样本的惩罚项，所以从图中可以看出，C越大，满足约束的点也就越多，支持向量也就越少。同时可以观察到gamma越大，支持向量越少，sklearn上对gamma的描述是它是核函数的系数，所以gamma与高斯核函数中的σ成反比，若σ趋于0，则拉格朗日乘子向量的所有分量都大于0，即全部样本点都是支持向量。所以gamma越大，支持向量越少。
在实践之后，我有以下收获：
SVM中利用核函数完成向高维空间的映射是关键。高斯核因为本质上对应的是无穷维空间，所以在某种程度上来说高斯核的表现能力最强，上面的可视化结果也说明了这点。参数gamma和C的设置会影响到模型的泛化性能的好坏，合理设置这两个参数能够缓解SVM的过拟合现象。