机器学习导论
习题四

参考答案

2017 年 5 月 23 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。

Solution.

(a) (1.1) ReLU、Sigmoid、tanh都具有非线性的性质，可以用作神经网络的激活函数。(1.2)但是Sigmoid和tanh函数在输入值的绝对值较大时梯度接近于0，在误差逆传播(BackPropagation)时，可能使参数难以更新，导致训练困难。(1.3)另外，ReLU计算过程简单，使得训练和预测的

[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

时间开销减少。(2.1) 与使用CPU相比，GPU能够加速神经网络的训练和预测，尤其是卷积操作的运算。(2.2)通过共享内存进行通信降低了并行的开销，能够进一步加速。(2.总结)速度提升后，在相同的时间内，可以接收更多训练数据，进行更多迭代，尝试更多参数，从而提高在任务上的性能。

(b) 可以用集成学习(ensemble)中的一些观点来解释。由于在训练中引入了随机性，训练完成后的多个神经网络有所不同，取其平均输出能够减小分歧(见课本185页，误差–分歧分解)，可能减少泛化误差。(从偏差–方差分解的角度谈减小方差也可)

(c) (1)Dropout技术被用在前两个全连接层之后。(2.1)可以认为dropout技术减少了层与层之间神经元的依赖，提高了鲁棒性，不易过拟合。(2.2)另有观点认为dropout技术带来了隐式的ensemble，相当于多个神经网络的综合。(3)使用这一技术的代价是收敛速度变慢，训练时间增加了一倍。

(d) (1)AlexNet有6000万个参数。(2.1)视觉任务本身较为复杂，因此较多的训练数据可以涵盖更多的情形。(2.2)AlexNet模型参数众多，拟合能力强，训练数据不足很可能导致过拟合，因此需要较多的训练数据。

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.

(1) 下文证明核矩阵$\mathbf{K}$总是半正定的，从而根据书中定理6.1，完成证明。

(i) 首先证明线性核$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$对应的核矩阵$\mathbf{K}$总是半正定的。
对于任意数据$D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$，以及$\forall \mathbf{z} \in \mathcal{R}^m$，

$$\mathbf{z}^{\mathrm{T}}\mathbf{K}\mathbf{z} = \mathbf{z}^{\mathrm{T}}D^{\mathrm{T}}D\mathbf{z} = \|D\mathbf{z}\|^2 \geq 0$$

由书中6.26可知，若κ_1和κ_2为核函数，则核函数的直积也是核函数。因此由于线性核是核函数，多项式核也是核函数。

(ii) 高斯核可以写成$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i}{2\sigma^2})\exp(\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j}{\sigma^2})\exp(-\frac{\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j}{2\sigma^2})$。由书中6.27可知，若$\kappa_1$为核函数，则对于任意函数$g(\mathbf{x})$，$\kappa(\mathbf{x}, \mathbf{z}) = g(\mathbf{x})\kappa_1(\mathbf{x}, \mathbf{z})g(\mathbf{z})$也是核函数。因此我们只需证明$\kappa_1(\mathbf{x}_i, \mathbf{x}_j) = \exp(\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j}{\sigma^2})$是核函数即可。
由泰勒展式可知，

$$\exp(\frac{\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j}{\sigma^2}) = \sum_{d=0}^{\infty}\frac{(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^d}{d!\sigma^{2d}}$$

根据书中6.25，若κ_1和κ_2为核函数，则对于任意正数γ_1和γ_2，其线性组合$\gamma_1\kappa_1 + \gamma_2\kappa_2$也是核函数。由第一问中多项式核是合法的核函数，证毕。

请大家注意每一步推导中等式不等式的正确性。例如：$(\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^d \neq (\mathbf{x}_i^{\mathrm{T}})^d(\mathbf{x}_j)^d$。

(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。

反例：取$x_1 = 1$，$x_2 = 2$，则$|\mathbf{K}| = -0.0579 < 0$，因此$\mathbf{K}$不是半正定矩阵。 □

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. (1) 考虑所有正例样本的下标集合为$\mathcal{P}$以及负例样本的下标集合为$\mathcal{N}$, 根据题干中的要求，我们只需要对负例分类错误的样本施加$k > 0$倍于正例中被分错的样本的“惩罚”,因此可以得到下面的优化目标

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C(\sum_{i\in\mathcal{P}}^{m}\xi_i + k\cdot\sum_{i\in\mathcal{N}}^{m}\xi_i) \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m \end{aligned} \tag{3.2}$$

(2) 记α, μ 表示拉格朗日乘子，则

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \mu) = & \frac{1}{2}\|\mathbf{w}\|^2 + C(\sum_{i\in\mathcal{P}}^{m}\xi_i + k\cdot\sum_{i\in\mathcal{N}}^{m}\xi_i) \\ & + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i \end{aligned} \tag{3.3}$$

令$\nabla_{\mathbf{w}}L = \nabla_b L = \nabla_{\xi_i}L = 0$, 则有

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \\ 0 &= \sum_{i=1}^{m}\alpha_i y_i \\ C &= (\alpha_i + \mu_i)\cdot(\frac{1}{k}\mathbb{I}(i\in\mathcal{P}) + \mathbb{I}(i\in\mathcal{N})) \end{aligned} \tag{3.4}$$

其中，$\mathbb{I}(\cdot)$为示性函数(indicator function)，当$\cdot$为真时取值为1，否则为0.
于是可以得到对偶问题如下：

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right) \\ \text{s.t.} \quad & \sum_{i=1}^{m} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C \cdot (k\mathbb{I}(i \in \mathcal{P}) + \mathbb{I}(i \in \mathcal{N})) \end{aligned} \tag{3.5}$$

因此可以得到KKT条件如下：

$$\begin{cases} \alpha_i, \mu_i, \xi_i \geq 0 \\ \xi_i - 1 + y_i(\mathbf{w}^{\mathrm{T}} \mathbf{x}_i + b) \geq 0 \\ \alpha_i (1 - \xi_i - y_i(\mathbf{w}^{\mathrm{T}} \mathbf{x}_i + b)) = 0 \\ \mu_i \xi_i = 0 \end{cases}$$

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。

Solution.

(1) 本题旨在让大家学会查阅库文档，理解其中模型各项参数的含义，并使用这些内容来计算RBF核SVM的预测结果，以理解对偶形式下的预测过程。能通过测试样例即可。Python程序中调用decision function不能通过评测。

(2) 本题旨在让大家简要了解绘图工具的使用，增进对SVM各项参数的理解。请注意，在低维情形下的理解不一定适合高维实际问题，具体问题中这些参数如何发挥作用，还需要同学们在实践中学习。程序不能生成figure.pdf会被扣分。

(3) 能帮助大家理解SVM，本题的意义就达到了。

编程题的反馈和参考解答见页面`http://lamda.nju.edu.cn/ml2017/ml_faq.html`中的更新内容。

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