

机器学习导论

习题四

141220132, 银琦, 141220132@smail.nju.edu.cn

2017年5月16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”，由 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 Networks²，本题目为此讲义的Exercise-5，已获得吴建鑫老师授权使用。

Solution.

(a) 传统的训练模型在训练很大的数据集时会产生很大的误差，并且无法使用饱和神经元模型，在ReLU之前用来替换的训练模型会过拟合，而ReLU可以大大加快训练速度，快速的学习对大型的模型产生较好的效果。

对于GPUs，文章中实现了GPU的交叉并行化，把卷积模板或者神经元平均分配到两个GPU上，

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>

但是只在固定的层次进行数据传输。

(b) 因为这样实现了图像数据的增强，可以有效的防止过拟合。

(c) “dropout”用于组合各种模型。

它通过50%的机率把每个隐藏神经元的输出变为0，这样，神经元相当于被抛弃了，对于前项传播就没有任何影响，也不会参与反向传播，这些结构共享权重，减少了复杂的神经网络组合。

每一次输入，神经网络都会建立不同的结构，并且网络收敛的迭代次数翻倍了。

(d) 在AlexNet中有6千万个参数。

因为数据库中数据总量有限，所以数据集大小不仅要保证能够训练出神经网络，还有有一定数量的验证集和测试集进行评估，所以120万的训练集数据量较为合适。

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： $\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^d$;

(ii) [10pts] 高斯核： $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$, 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) (i) 多项式核函数有效性证明：证明 $\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^d$, 即对任意的向量 $\mathbf{x}_i, \mathbf{x}_j$, $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ 可以被写成 $\phi^T(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ 的形式。

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^d$$

$$= \mathbf{x}_i^T \mathbf{x}_j \mathbf{x}_i^T \mathbf{x}_j \mathbf{x}_i^T \mathbf{x}_j \cdots \mathbf{x}_i^T \mathbf{x}_j$$

因为 $\mathbf{x}_j \mathbf{x}_i^T$ 为常数，设为A，所以原式可化为：

$$\begin{aligned} \kappa(\mathbf{x}_i, \mathbf{x}_j) &= \mathbf{x}_i^T A A A \cdots A \mathbf{x}_j \\ &= A^{d-2} \mathbf{x}_i^T \mathbf{x}_j \\ &= A^{\frac{d-2}{2}} \mathbf{x}_i^T \mathbf{x}_j A^{\frac{d-2}{2}} \\ &= \phi^T(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \end{aligned}$$

得证。

(ii) 高斯核函数有效性证明：证明 $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$, 其中 $\sigma > 0$ 是核函数，即对任意的向量 $\mathbf{x}_i, \mathbf{x}_j$, $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ 可以被写成 $\phi^T(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ 的形式。

因为

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \mathbf{x}_i^T \mathbf{x}_i + \mathbf{x}_j^T \mathbf{x}_j - 2\mathbf{x}_i^T \mathbf{x}_j$$

所以

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\mathbf{x}_i^T \mathbf{x}_i}{2\sigma^2}\right) \cdot \exp\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\sigma^2}\right) \cdot \exp\left(-\frac{\mathbf{x}_j^T \mathbf{x}_j}{2\sigma^2}\right)$$

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

$$\exp\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\sigma^2}\right) = \sum_{n=0}^{+\infty} \frac{1}{n!} \cdot \frac{(\mathbf{x}_i^T \mathbf{x}_j)^n}{\sigma^{2n}} = \sum_{n=0}^{+\infty} \frac{1}{\sqrt{n!}} \cdot \frac{(\mathbf{x}_i^T)^n}{\sigma^n} \cdot \frac{1}{\sqrt{n!}} \cdot \frac{(\mathbf{x}_j)^n}{\sigma^n} = \varphi^T(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$$

于是

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

其中

$$\phi^T(\mathbf{x}_i) = \exp\left(-\frac{\mathbf{x}_i^T \mathbf{x}_i}{2\sigma^2}\right) \cdot \varphi^T(\mathbf{x}_i)$$

$$\phi(\mathbf{x}_j) = \varphi(\mathbf{x}_j) \cdot \exp\left(-\frac{\mathbf{x}_j^T \mathbf{x}_j}{2\sigma^2}\right)$$

得证。

参考资料<http://blog.csdn.net/u010551462/article/details/41748807>

- (2) 如果 $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{1+e^{-\mathbf{x}_i^T \mathbf{x}_j}}$ 是合法的核函数，设该核函数的核矩阵为K，则K总是半正定的，即对于任意非零向量z，都有 $z^T K z \geq 0$ 。

$$z^T K z = \sum_i \sum_j z_i \frac{1}{1+e^{-\mathbf{x}_i^T \mathbf{x}_j}} z_j$$

因为 $0 < \frac{1}{1+e^{-\mathbf{x}_i^T \mathbf{x}_j}} < 1$ ，所以取z使得 $z_s \cdot z_t < 0$ 并且z的其余值均为0，使得 $z_s \frac{1}{1+e^{-\mathbf{x}_s^T \mathbf{x}_t}} z_t < 0$ ，从而 $\sum_i \sum_j z_i \frac{1}{1+e^{-\mathbf{x}_i^T \mathbf{x}_j}} z_j < 0$ ，与假设矛盾，所以该K不是半正定的，即 $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{1+e^{-\mathbf{x}_i^T \mathbf{x}_j}}$ 不是合法的核函数。

□

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))，

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

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)由题意, 对正例的惩罚系数为C, 对负例的惩罚系数增加为 $k \cdot C$, 于是得到的优化问题如下:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\mathbb{I}(x = x^+) + k\mathbb{I}(x = x^-))\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \dots, m. \end{aligned}$$

(2)通过拉格朗日乘子法可得到(1)中式子的拉格朗日函数:

$$\begin{aligned} L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = & \frac{1}{2} \|\boldsymbol{\omega}\|^2 + C \sum_{i=1}^m (\mathbb{I}(x = x^+) + k\mathbb{I}(x = x^-))\xi_i \\ & + \sum_{i=1}^m \alpha_i(1 - \xi_i - y_i(\boldsymbol{\omega}^\top \mathbf{x}_i + b)) - \sum_{i=1}^m \mu_i \xi_i \end{aligned}$$

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

令 $L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$ 对 $\boldsymbol{\omega}, b, \xi_i$ 的偏导为零可得

$$\begin{aligned} \boldsymbol{\omega} &= \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \\ 0 &= \sum_{i=1}^m \alpha_i y_i \\ C(\mathbb{I}(x = x^+) + k\mathbb{I}(x = x^-)) &= \alpha_i + \mu_i \end{aligned}$$

将上面三个式子代入拉格朗日函数可得到对偶问题如下:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \\ \text{s.t.} \quad \sum_{i=1}^m \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C(\mathbb{I}(x = x^+) + k\mathbb{I}(x = x^-)), i = 1, 2, \dots, m \end{aligned}$$

对应的KKT条件:

$$\left\{ \begin{array}{l} \alpha_i \geq 0, \mu_i \geq 0 \\ y_i f(\mathbf{x}_i) - 1 + \xi_i \geq 0 \\ \alpha_i(y_i f(\mathbf{x}_i) - 1 + \xi_i) = 0 \\ \xi_i \geq 0, \mu_i \xi_i = 0 \end{array} \right.$$

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LIBSVM³。

- (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.

- (3) 核函数很巧妙的将低维空间中较难分开的数据映射到了高维空间，通过超平面更加容易的分类，在库函数中进行调用SVM库函数时参数很重要，合适的参数会使得分类的错误率减小。
- (2) 红色蓝色圆点为原数据，黄色星花为支持向量。

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

