libSVM 多分类&概率估计

多分类

- 实现方法为"one-against-one"
- 对于k分类问题，训练$\frac{k(k-1)}{2}$个二分类器投票
- 平票时，直接选取排序下第一个类别

概率估计

原本的SVM对于一个输入样本x,在二分类下只能输出一个“判别值”$f(x)$，而在多分类中基于多个子分类器投票最终只能输出一个类别值。

$$f(x) = h(x) + b$$
$$h(x) = \sum_i y_i \alpha_i k(x_i, x)$$

然而在一些应用中，我们希望能得到一个概率值$P(y_i = k|x_i)$而非直接的分类结果。因此我们的问题也即为如果能基于训练好的SVM估计出:

$$p_i = P(y = i|x),\ i = 1, \ldots, k$$

类比多分类时"one-against-one"将k分类任务拆分为多个二分类任务，我们也可以将这一概率的估计拆分为多个二分类的概率估计：

$$r_{ij} \approx P(y = i|y = i \text{ or } j, x)$$

二分类下概率估计

对于这一任务，Platt在2000年提出了一种方法，即将SVM输出值套一层带参数的Sigmoid函数,从而将模型输出映射到$[0, 1]$，之后只需用最大似然法估计其中的参数即可：

$$P(y = 1|f) = \frac{1}{1 + exp(Af + B)}$$

其中f即为SVM的输出值$f(x)$. 同样将数据集中原本在$\{-1, 1\}$上的类别标记y做如下变换映射到$\{0, 1\}$上：

$$t_i = \frac{y_i + 1}{2}$$

因此，估计r_{ij}的问题也就变为了在数据集(f_i, t_i)上优化如下对数似然损失函数的问题：

$$\min_{A,B} -\sum_i [t_i log(p_i) + (1 - t_i) log(1 - p_i)]$$
$$p_i = \frac{1}{1 + exp(Af_i + B)}$$

然而有两个问题需要我们考虑：数据集如何选取、是否会过拟合。若直接选取训练集上的x_i送入SVM模型得到$f_i = f(x_i)$，则其是有偏的，因为SVM在优化过程中会明显地约束样本输出值$|f(x_i)| \geq 1$,即使是soft-margin改进下的SVM，也会对在边缘附近的样本输出值有明显影响，因此尤其是在非线性核的情况下，有一定部分的样本都是支持向量，直接在SVM训练集上的样本，$f(x_i)$的分布受到了模型优化中的影响，是不能直接拿来做为f_i的。因此一个简单的策略为k-折交叉验证，也即将训练集等分为k份，轮流取

一份数据不参与训练而作为f_i输出，最后取其并集为Sigmoid分布估计的训练集f_i。libsvm中采用了5折交叉检验。

第二个问题是在正负例不均衡时会很容易出现过拟合的情况,例如正例只有很少几个的情况时。对于这个问题，这里采用了一种模型无关、直接在采样上处理的方法，也即对正负样本标记t_i做确定的微小扰动$\epsilon_{\pm}$:

$$t_+ = 1 - \epsilon_+ = \frac{N_+ + 1}{N_+ + 2}$$
$$t_- = 0 + \epsilon_- = \frac{1}{N_- + 2}$$

其直观解释是对于一个样本x_i,考虑有可能有很少量相同的样本出现完全相反的标记，因此其平均标记为$t_+ = 1 - \epsilon_+ < 1$,从而增加模型的泛化能力。注意到原本$t_i \in \{0,1\}$是离散的，而这样修改后变为连续的，但这并不影响之前模型，因为之前的对数似然损失其实也可以视为是p_i和t_i上的KL散度。

最终优化问题为:

$$\min_{z=(A,B)} F(z) = -\sum_{i=1}^{l} \left(t_i \log(p_i) + (1-t_i)\log(1-p_i)\right),$$

$$\text{for } p_i = P_{A,B}(f_i), \text{ and } t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} & \text{if } y_i = +1 \\ \frac{1}{N_- + 2} & \text{if } y_i = -1 \end{cases}, \quad i = 1, \ldots, l.$$

显然，这是一个无约束凸优化的问题，可以使用任意的方法来优化。对于这个问题，libsvm采用了Lin在2007年发表的方法，也即使用带线搜索的牛顿法来优化，其梯度及Hessian阵为:

$$\nabla F(z) = \begin{bmatrix} \sum_{i=1}^{l} f_i(t_i - p_i) \\ \sum_{i=1}^{l}(t_i - p_i) \end{bmatrix},$$

$$H(z) = \begin{bmatrix} \sum_{i=1}^{l} f_i^2 p_i(1-p_i) & \sum_{i=1}^{l} f_i p_i(1-p_i) \\ \sum_{i=1}^{l} f_i p_i(1-p_i) & \sum_{i=1}^{l} p_i(1-p_i) \end{bmatrix}.$$

牛顿法流程为(为了避免Hessian阵奇异，其添加了一个乘以很小数的单位阵σI在Hessian阵上):

Algorithm 1 Newton's method with backtracking line search

Input: Initial point z_0, and parameter $\sigma \geq 0$ such that $H(z) + \sigma I$ is positive definite for all z

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Solve $(H_k + \sigma I)\delta_k = -\nabla F(z_k)$
3: Find α_k as the first element of the sequence $1, \frac{1}{2}, \frac{1}{4}, \ldots$ to satisfy

$$F(z_k + \alpha_k \delta_k) \leq F(z_k) + 0.0001 \cdot \alpha_k \left(\nabla F(z_k)^T \delta_k\right) \qquad (3)$$

4: Set $z_{k+1} = z_k + \alpha_k \delta_k$
5: **end for**

这样，我们就可以得到二分下$r_{ij} \approx P(y = i | y = i \text{ or } j, x)$的估计了。

多分类下概率估计

考虑已知所有两两类别概率r_{ij}下，我们希望求得多分类的概率估计$p_i = P(y=i|x)$. Wu在2004年的一篇文章列举出了多种方法，libsvm选用了其中提出的第二种方法,也即优化如下的问题：

$$\min_{p} \quad \frac{1}{2}\sum_{i=1}^{k}\sum_{j:j\neq i}(r_{ji}p_i - r_{ij}p_j)^2$$
$$\text{subject to} \quad p_i \geq 0, \forall i, \quad \sum_{i=1}^{k} p_i = 1.$$

这一优化目标的来源是理论上应满足如下式子：

$$P(y=i|y=i \text{ or } j, x)P(y=j|x) = P(y=j|y=i \text{ or } j, x)P(y=i|x)$$
$$r_{ji}p_i \approx r_{ij}p_j$$

对于上述优化问题，可以写成矩阵的形式如下：

$$\min_{p} \frac{1}{2}p^T Qp$$
$$\text{subject to } p_i \geq 0, \forall i, p^T\mathbf{1} = 1$$

其中：

$$Q_{ij} = \begin{cases} \sum_{s:s\neq i} r_{si}^2 & \text{if } i = j, \\ -r_{ji}r_{ij} & \text{if } i \neq j. \end{cases}$$

可以证明，去掉对$p_i \geq 0$的不等约束后优化问题不变，因此显然这是一个等式约束下的二次优化问题，其最优解需满足：

$$\begin{bmatrix} Q & \boldsymbol{e} \\ \boldsymbol{e}^T & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{p} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}$$

其中e是全为1的向量，$\mathbf{0}$是全为0的向量，b是拉格朗日乘子。

对于上式，除了可直接使用高斯消元法解以外，还可以通过一种迭代算法来解，libsvm正是采用了如下的方法：

对于上述等式，可以拆开写为：

$$Qp + eb = 0$$
$$\therefore p^TQp + p^Teb = 0$$
$$\therefore b = -p^TQp$$

而对于上述等式的第t行，有：

$$Q_tp + b = 0$$

带入$b=-p^TQp$:

$$Q_tp-p^TQp=Q_{tt}p_t+\sum_{j\neq t}Q_{tj}p_j-p^TQp=0$$

因此可以推导出迭代公式:

$$p_t\leftarrow\frac{1}{Q_{tt}}[-\sum_{j\neq t}Q_{tj}p_j+p^TQp]$$

迭代算法为:

Algorithm 3

1. Start with an initial $\boldsymbol{p}$ satisfying $p_i\geq 0,\forall i$ and $\sum_{i=1}^k p_i=1$.

2. Repeat $(t=1,\ldots,k,1,\ldots)$

$$p_t\leftarrow\frac{1}{Q_{tt}}[-\sum_{j:j\neq t}Q_{tj}p_j+\boldsymbol{p}^TQ\boldsymbol{p}] \tag{47}$$

normalize $\boldsymbol{p}$

until Eq. (45) is satisfied.

可以证明，上述算法可以确保收敛到全局最优解。

参考文献:

[1] J. C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Sch olkopf, and D. Schuurmans, editors, Advances in Large Margin Classiers, Cambridge, MA, 2000. MIT Press.

[2] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for support vector machines. Machine Learning, 68:267{276, 2007. URL http://www.csie.ntu.edu.tw/~cjlin/papers/plattprob.pdf.

[3] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classication by pairwise coupling. Journal of Machine Learning Research, 5:975{1005, 2004. URL http://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/svmprob.pdf.