Author：Liu Jian

Time：2020-10-15

Least squares for classification

> 相关章节：4.1.3 Least squares for classification

Least squares for classification

思路是对采样得到的数据 (x,t)，使用模型 $\mathrm{y}(\mathrm{x})$ 拟合 t，由此将分类问题转化为回归问题。对于回归问题，一个很自然的想法就是使用 least squares。

注意，和一般回归问题的拟合目标 t 天生就存在不同，由分类问题转化而来的回归问题的 target t 并不是本身就存在的，而是由人为指定的，这一将类别 $\mathcal{C}_k$，$k=1,2,\cdots,K$ 转化为 K 个离散数值的过程，就是 embedding。 比如，最简单的 target $\mathrm{t}=k$ 表示第 k 个类别；再比如，使用 one-hot 编码 (1-of-K coding scheme) target $\mathrm{t}=I_k$ 表示第 k 个类别，其中 I_k 表示第 k 行为 1 其余行均为 0 的 K 维列向量，等等。可以看到，target t 的取值范围是指定的 K 个离散值，不妨记为 $\mathcal{T}_k$，$k=1,2,\cdots,K$。

从概率论的角度，least squares 实际上就是在使用模型 $\mathrm{t}=\mathrm{y}(\mathrm{x})+\boldsymbol{\varepsilon}$，$\boldsymbol{\varepsilon}\sim\mathcal{N}(0,\Sigma)$，即概率分布 $p_{model}(\mathrm{t}|\mathrm{x})=\mathcal{N}(\mathrm{y}(\mathrm{x}),\Sigma)$ 去拟合目标概率 $p_{object}(\mathrm{t}|\mathrm{x})$：

$\mathrm{t}\|\mathrm{x}$	$\mathcal{T}_1$	$\mathcal{T}_2$	$\cdots$	$\mathcal{T}_K$
$p_{object}(\mathrm{t}\|\mathrm{x})$	$N_1(\mathrm{x})/N(\mathrm{x})$	$N_2(\mathrm{x})/N(\mathrm{x})$	$\cdots$	$N_K(\mathrm{x})/N(\mathrm{x})$

其中 $N(\mathrm{x})$ 表示样本中 x 出现的总次数，$N_k(\mathrm{x})$ 表示其中类别为 $\mathcal{C}_k$ 的个数，$\sum_{k=1}^K N_k(\mathrm{x})=N(\mathrm{x})$。记 $p_k(\mathrm{x})=N_k(\mathrm{x})/N(\mathrm{x})$，此外，为了简便起见，我们有时会简记 $p_k(\mathrm{x})$ 为 p_k。可以看到，当样本集中某个 x 只出现一次，或其类别不存在随机性时，$\mathrm{t}|\mathrm{x}$ 的分布律只在其类别处为 1，其余类别处为 0；此时若将 $\mathcal{T}_k$ 取为 one-hot 编码，则 $\mathrm{t}|\mathrm{x}$ 的值恰好给出了其分布律。

我们指出：

1. 对任意点 x，$p_{model}(\mathrm{t}|\mathrm{x})$ 是连续的高斯分布，而 $p_{object}(\mathrm{t}|\mathrm{x})$ 是只有 K 个不同状态的 multinoulli distribution (也称范畴分布，categorical distribution)，两个分布的形式相差太多，使用 p_{model} 去近似 p_{object} 效果显然不会很好；
2. 条件高斯分布$p_{model}(\mathrm{t}|\mathrm{x})$ 的均值由 $\mathrm{y}(\mathrm{x})$ 给出，由 least squares 的知识可知，若不限制 $\mathrm{y}(\mathrm{x})$ 函数形式 (泛函问题)，其最优解为 $\mathbb{E}_{p_{object}(\mathrm{t}|\mathrm{x})}[\mathrm{t}]$，即

$$\mathrm{y}_{opt}(\mathrm{x})=\mathbb{E}_{p_{object}(\mathrm{t}|\mathrm{x})}[\mathrm{t}]=\sum_{k=1}^K p_k(\mathrm{x})\mathcal{T}_k$$

进一步，当取 $\mathcal{T}_k$ 为 one-hot 编码时：

$$\mathrm{y}_{opt}(\mathrm{x})=\sum_{k=1}^K p_k(\mathrm{x})I_k=\begin{bmatrix}p_1(\mathrm{x})\\p_2(\mathrm{x})\\\vdots\\p_K(\mathrm{x})\end{bmatrix}$$

可见，$\mathrm{y}_{opt}(\mathrm{x})$ 恰好给出了 t 的后验分布律 $p_{object}(\mathrm{t}|\mathrm{x})$。我们强调，对于分布 $p_{model}(\mathrm{t}|\mathrm{x})$ 而言，$\mathrm{y}(\mathrm{x})$ 是参数，即 K 维均值向量；而就 K 维向量 $\mathrm{y}(\mathrm{x})$ 各元素的值而言，$\mathrm{y}(\mathrm{x})$ 在近似一个分布律 $\mathrm{y}_{opt}(\mathrm{x})$。由此，我们会想到，既然 $\mathrm{y}(\mathrm{x})$ 在近似一个分布律 $\mathrm{y}_{opt}(\mathrm{x})$，那么 $\mathrm{y}(\mathrm{x})$ 的各元素是否可被解释为概率呢？答案是不能。

由书中所述，若 $\mathcal{T}_k$ 满足线性约束条件：

$$\mathrm{a}^T\mathcal{T}_k + b = 0,\ \ k = 1, 2, \cdots, K$$

则对任意的 x，least squares 的解 $\mathrm{y}^*(\mathrm{x})$ 满足同样的线性约束：

$$\mathrm{a}^T\mathrm{y}^*(\mathrm{x}) + b = 0$$

因此，当 target t 使用 one-hot 编码，即 $\mathcal{T}_k = I_k$ 时，对任意的 x，$\mathrm{y}^*(\mathrm{x})$ 中的各元素之和为 1 。但需要指出的是，我们并不能保证各元素落在区间 $[0, 1]$ 内，因为 $\mathrm{y}(\mathrm{x})$ 并不能取任意函数，我们会限制其形式为线性模型：

$$\mathrm{y}(\mathrm{x}) = \begin{bmatrix} y_1(\mathrm{x}) \\ y_2(\mathrm{x}) \\ \vdots \\ y_K(\mathrm{x}) \end{bmatrix}$$

$$y_k(\mathrm{x}) = \mathrm{w}_k^T\mathrm{x} + w_{k0},\ \ k = 1, 2, \cdots, K$$

因此，我们求得的 $\mathrm{y}(\mathrm{x})$ 虽然在近似一个分布律 $\mathrm{y}_{opt}(\mathrm{x})$，且各元素之和为 1，但各元素并不保证一定落在区间 $[0, 1]$ 内，因此并不具有概率的含义。

综上，考虑到高斯分布和 multinoulli distribution 相去甚远，且高斯分布的均值向量 $\mathrm{y}(\mathrm{x})$ 取线性模型的局限性，不难想到，上述方法所得模型的分类效果并不会很好。