线性分类

对于分类任务，线性回归模型就无能为力了，但是我们可以在线性模型的函数进行后再加入一层激活函数，这个函数是非线性的，激活函数的反函数叫做链接函数。我们有两种线性分类的方式：

1. 硬分类，我们直接需要输出观测对应的分类。这类模型的代表为：
    1. 线性判别分析（Fisher 判别）
    2. 感知机
2. 软分类，产生不同类别的概率，这类算法根据概率方法的不同分为两种
    1. 生成式（根据贝叶斯定理先计算参数后验，再进行推断）：高斯判别分析（GDA）和朴素贝叶斯等为代表
        1. GDA
        2. Naive Bayes
    2. 判别式（直接对条件概率进行建模）：Logistic 回归

两分类-硬分类-感知机算法

我们选取激活函数为：

$$sign(a) = \begin{cases} +1, a \geq 0 \\ -1, a < 0 \end{cases} \tag{1}$$

这样就可以将线性回归的结果映射到两分类的结果上了。

定义损失函数为错误分类的数目，比较直观的方式是使用指示函数，但是指示函数不可导，因此可以定义：

$$L(w) = \sum_{x_i \in \mathcal{D}_{wrong}} -y_i w^T x_i \tag{2}$$

其中，$\mathcal{D}_{wrong}$是错误分类集合，实际在每一次训练的时候，我们采用梯度下降的算法。损失函数对 w 的偏导为：

$$\frac{\partial}{\partial w} L(w) = \sum_{x_i \in \mathcal{D}_{wrong}} -y_i x_i \tag{3}$$

但是如果样本非常多的情况下，计算复杂度较高，但是，实际上我们并不需要绝对的损失函数下降的方向，我们只需要损失函数的期望值下降，但是计算期望需要知道真实的概率分布，我们实际只能根据训练数据抽样来估算这个概率分布（经验风险）：

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\hat{p}}[\nabla_w L(w)]] = \mathbb{E}_{\mathcal{D}}[\frac{1}{N} \sum_{i=1}^{N} \nabla_w L(w)] \tag{4}$$

我们知道， N 越大，样本近似真实分布越准确，但是对于一个标准差为 σ 的数据，可以确定的标准差仅和 $\sqrt{N}$ 成反比，而计算速度却和 N 成正比。因此可以每次使用较少样本，则在数学期望的意义上损失降低的同时，有可以提高计算速度，如果每次只使用一个错误样本，我们有下面的更新策略（根据泰勒公式，在负方向）：

$$w^{t+1} \leftarrow w^t + \lambda y_i x_i \tag{5}$$

是可以收敛的，同时使用单个观测更新也可以在一定程度上增加不确定度，从而减轻陷入局部最小的可能。在更大规模的数据上，常用的是小批量随机梯度下降法。

两分类-硬分类-线性判别分析 LDA

在 LDA 中，我们的基本想法是选定一个方向，将试验样本顺着这个方向投影，投影后的数据需要满足两个条件，从而可以更好地分类：

1. 相同类内部的试验样本距离接近。
2. 不同类别之间的距离较大。

首先是投影，我们假定原来的数据是向量 x，那么顺着 w 方向的投影就是标量：

$$z = w^T \cdot x (= |w| \cdot |x| \cos\theta) \tag{6}$$

对第一点，相同类内部的样本更为接近，我们假设属于两类的试验样本数量分别是 N_1 和 N_2，那么我们采用方差矩阵来表征每一个类内的总体分布，这里我们使用了协方差的定义，用 S 表示原数据的协方差：

$$\begin{aligned}
C_1 : Var_z[C_1] &= \frac{1}{N_1}\sum_{i=1}^{N_1}(z_i - \overline{z_{c1}})(z_i - \overline{z_{c1}})^T \\
&= \frac{1}{N_1}\sum_{i=1}^{N_1}(w^T x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^T x_j)(w^T x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^T x_j)^T \\
&= w^T \frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - \overline{x_{c1}})(x_i - \overline{x_{c1}})^T w \\
&= w^T S_1 w \qquad (7) \\
C_2 : Var_z[C_2] &= \frac{1}{N_2}\sum_{i=1}^{N_2}(z_i - \overline{z_{c2}})(z_i - \overline{z_{c2}})^T \\
&= w^T S_2 w \qquad (8)
\end{aligned}$$

所以类内距离可以记为：

$$Var_z[C_1] + Var_z[C_2] = w^T(S_1 + S_2)w \tag{9}$$

对于第二点，我们可以用两类的均值表示这个距离：

$$\begin{aligned}
(\overline{z_{c1}} - \overline{z_{c2}})^2 &= (\frac{1}{N_1}\sum_{i=1}^{N_1} w^T x_i - \frac{1}{N_2}\sum_{i=1}^{N_2} w^T x_i)^2 \\
&= (w^T(\overline{x_{c1}} - \overline{x_{c2}}))^2 \\
&= w^T(\overline{x_{c1}} - \overline{x_{c2}})(\overline{x_{c1}} - \overline{x_{c2}})^T w \qquad (10)
\end{aligned}$$

综合这两点，由于协方差是一个矩阵，于是我们用将这两个值相除来得到我们的损失函数，并最大化这个值：

$$\begin{aligned}\hat{w} = \underset{w}{argmax}\, J(w) &= \underset{w}{argmax}\, \frac{(\overline{z_{c1}} - \overline{z_{c2}})^2}{Var_z[C_1] + Var_z[C_2]} \\ &= \underset{w}{argmax}\, \frac{w^T(\overline{x_{c1}} - \overline{x_{c2}})(\overline{x_{c1}} - \overline{x_{c2}})^T w}{w^T(S_1 + S_2)w} \\ &= \underset{w}{argmax}\, \frac{w^T S_b w}{w^T S_w w}\end{aligned} \tag{11}$$

这样，我们就把损失函数和原数据集以及参数结合起来了。下面对这个损失函数求偏导，注意我们其实对 w 的绝对值没有任何要求，只对方向有要求，因此只要一个方程就可以求解了：

$$\begin{aligned}&\frac{\partial}{\partial w} J(w) = 2S_b w(w^T S_w w)^{-1} - 2w^T S_b w(w^T S_w w)^{-2} S_w w = 0 \\ &\Longrightarrow S_b w(w^T S_w w) = (w^T S_b w) S_w w \\ &\Longrightarrow w \propto S_w^{-1} S_b w = S_w^{-1}(\overline{x_{c1}} - \overline{x_{c2}})(\overline{x_{c1}} - \overline{x_{c2}})^T w \propto S_w^{-1}(\overline{x_{c1}} - \overline{x_{c2}})\end{aligned} \tag{12}$$

于是 $S_w^{-1}(\overline{x_{c1}} - \overline{x_{c2}})$ 就是我们需要寻找的方向。最后可以归一化求得单位的 w 值。

两分类-软分类-概率判别模型-Logistic 回归

有时候我们只要得到一个类别的概率，那么我们需要一种能输出 $[0, 1]$ 区间的值的函数。考虑两分类模型，我们利用判别模型，希望对 $p(C|x)$ 建模，利用贝叶斯定理：

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)} \tag{13}$$

取 $a = \ln \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$，于是：

$$p(C_1|x) = \frac{1}{1 + \exp(-a)} \tag{14}$$

上面的式子叫 Logistic Sigmoid 函数，其参数表示了两类联合概率比值的对数。在判别式中，不关心这个参数的具体值，模型假设直接对 a 进行。

Logistic 回归的模型假设是：

$$a = w^T x \tag{15}$$

于是，通过寻找 w 的最佳值可以得到在这个模型假设下的最佳模型。概率判别模型常用最大似然估计的方式来确定参数。

对于一次观测，获得分类 y 的概率为（假定$C_1 = 1, C_2 = 0$）：

$$p(y|x) = p_1^y p_0^{1-y} \tag{16}$$

那么对于 N 次独立全同的观测 MLE为：

$$\hat{w} = \underset{w}{argmax}\, J(w) = \underset{w}{argmax} \sum_{i=1}^{N} (y_i \log p_1 + (1 - y_i) \log p_0) \tag{17}$$

注意到，这个表达式是交叉熵表达式的相反数乘 N，MLE 中的对数也保证了可以和指数函数相匹配，从而在大的区间汇总获取稳定的梯度。

对这个函数求导数，注意到：

$$p_1' = (\frac{1}{1+\exp(-a)})' = p_1(1-p_1) \tag{18}$$

则：

$$J'(w) = \sum_{i=1}^{N} y_i(1-p_1)x_i - p_1 x_i + y_i p_1 x_i = \sum_{i=1}^{N}(y_i - p_1)x_i \tag{19}$$

由于概率值的非线性，放在求和符号中时，这个式子无法直接求解。于是在实际训练的时候，和感知机类似，也可以使用不同大小的批量随机梯度上升（对于最小化就是梯度下降）来获得这个函数的极大值。

两分类-软分类-概率生成模型-高斯判别分析 GDA

生成模型中，我们对联合概率分布进行建模，然后采用 MAP 来获得参数的最佳值。两分类的情况，我们采用的假设：

1. $y \sim Bernoulli(\phi)$
2. $x|y=1 \sim \mathcal{N}(\mu_1, \Sigma)$
3. $x|y=0 \sim \mathcal{N}(\mu_0, \Sigma)$

那么独立全同的数据集最大后验概率可以表示为：

$$\underset{\phi,\mu_0,\mu_1,\Sigma}{argmax} \log p(X|Y)p(Y) = \underset{\phi,\mu_0,\mu_1,\Sigma}{argmax} \sum_{i=1}^{N}(\log p(x_i|y_i) + \log p(y_i))$$

$$= \underset{\phi,\mu_0,\mu_1,\Sigma}{argmax} \sum_{i=1}^{N}((1-y_i)\log \mathcal{N}(\mu_0,\Sigma) + y_i \log \mathcal{N}(\mu_1,\Sigma) + y_i \log\phi + (1-y_i)\log(1-\phi)) \tag{20}$$

- 首先对 ϕ 进行求解，将式子对 ϕ 求偏导：

$$\sum_{i=1}^{N} \frac{y_i}{\phi} + \frac{y_i - 1}{1-\phi} = 0$$

$$\Longrightarrow \phi = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{N_1}{N} \tag{21}$$

- 然后求解 μ_1：

$$\begin{aligned}\hat{\mu_1} &= \underset{\mu_1}{argmax} \sum_{i=1}^{N} y_i \log \mathcal{N}(\mu_1, \Sigma) \\ &= \underset{\mu_1}{argmin} \sum_{i=1}^{N} y_i (x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1) \end{aligned} \tag{22}$$

由于：

$$\sum_{i=1}^{N} y_i (x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1) = \sum_{i=1}^{N} y_i x_i^T \Sigma^{-1} x_i - 2 y_i \mu_1^T \Sigma^{-1} x_i + y_i \mu_1^T \Sigma^{-1} \mu_1 \tag{23}$$

求微分左边乘以 Σ 可以得到：

$$\begin{aligned} \sum_{i=1}^{N} -2 y_i \Sigma^{-1} x_i + 2 y_i \Sigma^{-1} \mu_1 &= 0 \\ \Longrightarrow \mu_1 = \frac{\sum\limits_{i=1}^{N} y_i x_i}{\sum\limits_{i=1}^{N} y_i} &= \frac{\sum\limits_{i=1}^{N} y_i x_i}{N_1} \end{aligned} \tag{24}$$

- 求解 μ_0，由于正反例是对称的，所以：

$$\mu_0 = \frac{\sum\limits_{i=1}^{N} (1 - y_i) x_i}{N_0} \tag{25}$$

- 最为困难的是求解 Σ，我们的模型假设对正反例采用相同的协方差矩阵，当然从上面的求解中我们可以看到，即使采用不同的矩阵也不会影响之前的三个参数。首先我们有：

$$\begin{aligned} \sum_{i=1}^{N} \log \mathcal{N}(\mu, \Sigma) &= \sum_{i=1}^{N} \log \left(\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \right) + \left(-\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right) \\ &= Const - \frac{1}{2} N \log |\Sigma| - \frac{1}{2} Trace((x_i - \mu)^T \Sigma^{-1} (x_i - \mu)) \\ &= Const - \frac{1}{2} N \log |\Sigma| - \frac{1}{2} Trace((x_i - \mu)(x_i - \mu)^T \Sigma^{-1}) \\ &= Const - \frac{1}{2} N \log |\Sigma| - \frac{1}{2} N Trace(S \Sigma^{-1}) \end{aligned} \tag{26}$$

在这个表达式中，我们在标量上加入迹从而可以交换矩阵的顺序，对于包含绝对值和迹的表达式的导数，我们有：

$$\frac{\partial}{\partial A}(|A|) = |A|A^{-1} \tag{27}$$

$$\frac{\partial}{\partial A}Trace(AB) = B^T \tag{28}$$

因此：

$$\begin{aligned} &[\sum_{i=1}^{N}((1-y_i)\log\mathcal{N}(\mu_0,\Sigma) + y_i\log\mathcal{N}(\mu_1,\Sigma)]' \\ &= Const - \frac{1}{2}N\log|\Sigma| - \frac{1}{2}N_1 Trace(S_1\Sigma^{-1}) - \frac{1}{2}N_2 Trace(S_2\Sigma^{-1}) \end{aligned} \tag{29}$$

其中，S_1, S_2 分别为两个类数据内部的协方差矩阵，于是：

$$\begin{aligned} N\Sigma^{-1} - N_1 S_1^T\Sigma^{-2} - N_2 S_2^T\Sigma^{-2} &= 0 \\ \Longrightarrow \Sigma &= \frac{N_1 S_1 + N_2 S_2}{N} \end{aligned} \tag{30}$$

这里应用了类协方差矩阵的对称性。

于是我们就利用最大后验的方法求得了我们模型假设里面的所有参数，根据模型，可以得到联合分布，也就可以得到用于推断的条件分布了。

两分类-软分类-概率生成模型-朴素贝叶斯

上面的高斯判别分析的是对数据集的分布作出了高斯分布的假设，同时引入伯努利分布作为类先验，从而利用最大后验求得这些假设中的参数。

朴素贝叶斯队数据的属性之间的关系作出了假设，一般地，我们有需要得到 $p(x|y)$ 这个概率值，由于 x 有 p 个维度，因此需要对这么多的维度的联合概率进行采样，但是我们知道这么高维度的空间中采样需要的样本数量非常大才能获得较为准确的概率近似。

在一般的有向概率图模型中，对各个属性维度之间的条件独立关系作出了不同的假设，其中最为简单的一个假设就是在朴素贝叶斯模型描述中的条件独立性假设。

$$p(x|y) = \prod_{i=1}^{p} p(x_i|y) \tag{31}$$

即：

$$x_i \perp x_j|y, \forall\, i \neq j \tag{32}$$

于是利用贝叶斯定理，对于单次观测：

$$p(y|x)=\frac{p(x|y)p(y)}{p(x)}=\frac{\prod_{i=1}^{p}p(x_i|y)p(y)}{p(x)} \tag{33}$$

对于单个维度的条件概率以及类先验作出进一步的假设：

1. x_i 为连续变量：$p(x_i|y)=\mathcal{N}(\mu_i,\sigma_i^2)$
2. x_i 为离散变量：类别分布（Categorical）：$p(x_i=i|y)=\theta_i,\sum_{i=1}^{K}\theta_i=1$
3. $p(y)=\phi^y(1-\phi)^{1-y}$

对这些参数的估计，常用 MLE 的方法直接在数据集上估计，由于不需要知道各个维度之间的关系，因此，所需数据量大大减少了。估算完这些参数，再代入贝叶斯定理中得到类别的后验分布。

小结

分类任务分为两类，对于需要直接输出类别的任务，感知机算法中我们在线性模型的基础上加入符号函数作为激活函数，那么就能得到这个类别，但是符号函数不光滑，于是我们采用错误驱动的方式，引入 $\sum_{x_i\in\mathcal{D}_{wrong}}-y_iw^Tx_i$ 作为损失函数，然后最小化这个误差，采用批量随机梯度下降的方法来获取最佳的参数值。而在线性判别分析中，我们将线性模型看作是数据点在某一个方向的投影，采用类内小，类间大的思路来定义损失函数，其中类内小定义为两类数据的方差之和，类间大定义为两类数据中心点的间距，对损失函数求导得到参数的方向，这个方向就是 $S_w^{-1}(\overline{x}_{c1}-\overline{x}_{c2})$，其中 S_w 为原数据集两类的方差之和。

另一种任务是输出分类的概率，对于概率模型，我们有两种方案，第一种是判别模型，也就是直接对类别的条件概率建模，将线性模型套入 Logistic 函数中，我们就得到了 Logistic 回归模型，这里的概率解释是两类的联合概率比值的对数是线性的，我们定义的损失函数是交叉熵（等价于 MLE），对这个函数求导得到 $\frac{1}{N}\sum_{i=1}^{N}(y_i-p_1)x_i$，同样利用批量随机梯度（上升）的方法进行优化。第二种是生成模型，生成模型引入了类别的先验，在高斯判别分析中，我们对数据集的数据分布作出了假设，其中类先验是二项分布，而每一类的似然是高斯分布，对这个联合分布的对数似然进行最大化就得到了参数，$\frac{\sum_{i=1}^{N}y_ix_i}{N_1},\frac{\sum_{i=1}^{N}(1-y_i)x_i}{N_0},\frac{N_1S_1+N_2S_2}{N},\frac{N_1}{N}$。在朴素贝叶斯中，我们进一步对属性的各个维度之间的依赖关系作出假设，条件独立性假设大大减少了数据量的需求。