2.3 Generative Model for Classification

判别模型和生成模型

分类算法当中，我们想学习一个映射$y=f(\boldsymbol{x})$, 其通过输入特征$\boldsymbol{x}$，得到其分类标签y。

判别模型 Discriminative Models：在逻辑回归当中，我们学习的概率模型是一个条件概率$p(y|\boldsymbol{x})$，并直接对$p(y|\boldsymbol{x})$建模，即最小化$NLL(\boldsymbol{w})=-\log p(D|\boldsymbol{w})$得到最佳权重，从而得到一个判决边界；

生成模型 Generative Models：生成模型同样学习条件概率$p(y|\boldsymbol{x})$，但其通过贝叶斯函数，对联合分布$p(\boldsymbol{x},y)$进行建模：

$$p(y|\boldsymbol{x})=\frac{p(\boldsymbol{x},y)}{p(\boldsymbol{x})}=\frac{p(y)p(\boldsymbol{x}|y)}{p(\boldsymbol{x})}$$

判决条件：

$$\hat{y}=\arg\max_y p(y|\boldsymbol{x})=\arg\max_y p(y)p(\boldsymbol{x}|y)=\arg\max_y[\log p(y)+\log p(\boldsymbol{x}|y)]$$

学习$p(y=c),c\in\{1,2,\dots,C\}$是为了的得到类别的 size，其称为类别的先验概率 prior probabilities；

学习$p(\boldsymbol{x}|y=c),c\in\{1,2,\dots,C\}$是为了得到类别的特征：

1. 我们常假设特征$\boldsymbol{x}$是服从某种分布的，例如高斯分布；换言之，特征$\boldsymbol{x}$是由高斯分布生成 generated 的；
2. 学习$p(\boldsymbol{x}|y=c)$的过程是决定生成模型参数的过程，常用方法是最大似然估计 MLE；
3. $p(\boldsymbol{x}|y=c),c\in\{1,2,\dots,C\}$常被称作类别条件密度 class conditional densities

多变量高斯分布 Multivariate Gaussian Distributions

对于连续变量，最常用的联合分布是多变量高斯分布$N(\mu,\Sigma)$:

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp[-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}]$$

其中D代表维度；$\boldsymbol{x}$代表数据，其为有D个随机变量的向量；$\boldsymbol{\mu}$代表数据均值的向量；$\boldsymbol{\Sigma}$代表协方差矩阵，$|\boldsymbol{\Sigma}|$是其行列式。

高斯判决分析 Gaussian Discriminant Analysis/GDA 是一种假设特征$\boldsymbol{x}$服从多变量高斯分布的生成模型：

$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta})=N(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$$

其中$\boldsymbol{\theta}=\{\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c|c=1,2,\dots,C\}$. 令$\pi_c=p(y=c),\boldsymbol{\pi}=(\pi_1,\pi_2,\dots,\pi_C)^T$, 则 GDA 的

所有参数可以表示为$(\boldsymbol{\pi}, \boldsymbol{\theta})$.
接下来要得到参数$(\boldsymbol{\pi}, \boldsymbol{\theta})$的最佳估计，同样采用最大似然估计方法，其对数似然函数为：

$$\log p(D|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta})$$

$$= \sum_{i=1}^{N} [\log p(y_i|\boldsymbol{\theta}) + \log p(\boldsymbol{x}_i|y_i, \boldsymbol{\theta})$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{I}(y_i = c) \, [\log p(y = c|\boldsymbol{\theta}) + \log p(\boldsymbol{x}_i|y = c, \boldsymbol{\theta})$$

$$= \sum_{c=1}^{C} \sum_{i:y_i=c} \log \pi_c + \sum_{c=1}^{C} \sum_{i:y_i=c} \log N(\boldsymbol{x}_i|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

$$= \sum_{c=1}^{C} n_c \log \pi_c + \sum_{c=1}^{C} \sum_{i:y_i=c} \log N(\boldsymbol{x}_i|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$

为了最大化$\log p(D|\boldsymbol{\theta})$，我们分别最大化展开式的两项：
$\sum_{c=1}^{C} n_c \log \pi_c$，以及不同$c$时的$\sum_{c=1}^{C} \sum_{i:y_i=c} \log N(\boldsymbol{x}_i|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$
由吉布斯不等式 Gibbs' Inequality:

$$\sum_{c=1}^{C} n_c \log \pi_c \leq \sum_{c=1}^{C} n_c \log \frac{N_c}{N}$$

其中$N_c = \Sigma_{i=1}^{N} \mathbb{I}(y_i = c)$是类别$c$的个数/size, $N = \Sigma_C N_C$,
因此π_c的最大似然估计 MLE 是

$$\widehat{\pi_c} = \frac{N_c}{N}$$

可以得到，$\boldsymbol{\mu}_c$和$\boldsymbol{\Sigma}_c$的最大似然估计 MLE 是：

$$\widehat{\boldsymbol{\mu}_c} = \frac{1}{N_c} \Sigma_{i:y_i=c} \boldsymbol{x}_i$$

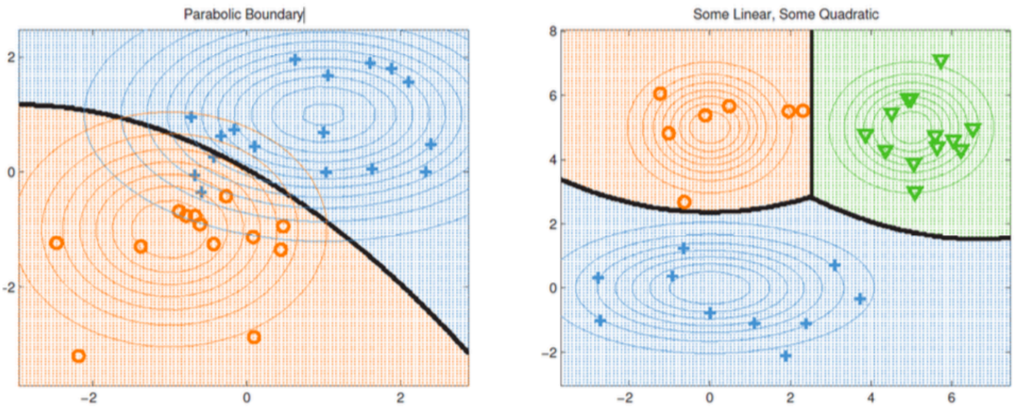
$$\widehat{\boldsymbol{\Sigma}_c} = \frac{1}{N_c} \Sigma_{i:y_i=c} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}_c})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}_c})^T$$

可以看到$\widehat{\boldsymbol{\mu}_c}$为样本均值，$\widehat{\boldsymbol{\Sigma}_c}$为样本协方差矩阵。
从而代入到之前的判决条件：

$$\hat{y} = \arg\max_c [\log p(y=c) + \log p(\boldsymbol{x}|y=c)]$$

$$= \arg\max_c [\log(\pi_c) - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_c|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)]$$

以上是一个二次函数 quadratic function，因此 GDA 也常被称作二次判决分析 Quadratic Discriminant Analysis，其判决边界常常是抛物线 parabolic，但也可以是线性的：



和 SoftMax 的关系

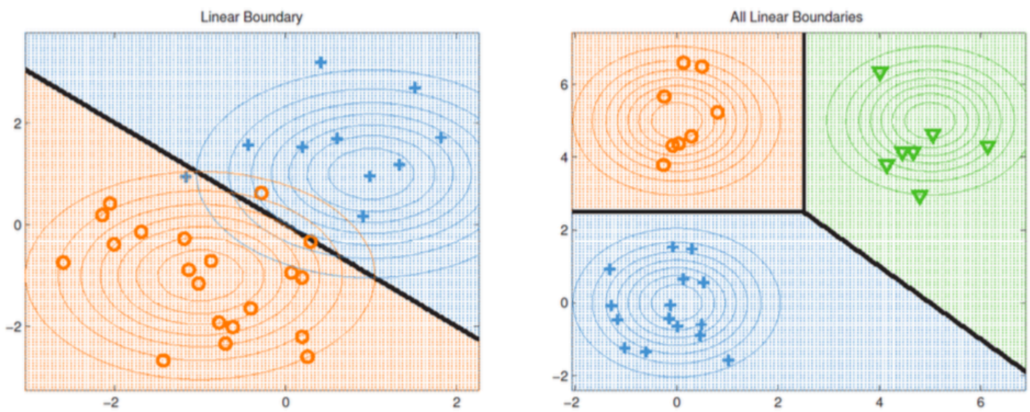
在 GDA 当中，每一个类别都有自己的协方差矩阵$\boldsymbol{\Sigma}_c$，当所有类别都有相同的协方差矩阵时，令

$$b_c = -\frac{1}{2}\boldsymbol{\mu}_c^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c \text{ and } \boldsymbol{w}_c^T = \boldsymbol{\mu}_c^T \boldsymbol{\Sigma}_c^{-1}$$

可以得到

$$p(y=c|\boldsymbol{x},\boldsymbol{\theta}) \approx \exp[\boldsymbol{w}_c^T \boldsymbol{x} + b_c]$$

GDA 退化为 SoftMax 模型；当总类别为 2 时，继续退化为逻辑回归模型；决策边界也由抛物线退化为直线，此时称为线性判决分析 Linear Discriminant Analysis：



判别模型 vs 生成模型

以高斯判决分析和逻辑回归为例：

1. 高斯判决分析做出的假设更强：假设高斯判决分析的所有类别拥有相同的协方差矩阵时，可以退化得到逻辑回归，反之不成立；
2. 参数估计上的差别：逻辑回归最大化条件对数似然 conditional log-likelihood

$$\Sigma_{\mathrm{i=1}}^{\mathrm{N}} \log p(y_i | \boldsymbol{x}_i, \boldsymbol{w})$$

   高斯判决分析最大化联合对数似然 joint log-likelihood

$$\Sigma_{\mathrm{i=1}}^{\mathrm{N}} \log p(y_i, \boldsymbol{x}_i | \boldsymbol{\theta})$$

3. 若高斯假设符合数据分布特征时（即假设特征$\boldsymbol{x}$服从多变量高斯分布），达到相同的训练水平时，高斯判决分析所需的数据量少于逻辑回归；
4. 若高斯假设不符合数据分布特征时，逻辑回归鲁棒性更好，对不正确的模型假设敏感度更低。

总结：

1. 参数估计在生成模型当中更易，在判别模型当中更难。原因在于生成模型常有最大似然估计的解析解，而判别模型需要通过梯度下降来得到数值解；
2. 生成模型更易处理 missing data：在训练过程中使用 Expectation Maximization/EM 算法，在测试过程当中采用边际化 marginalization 即可。而在判别模型当中没有处理 missing data 的标准方法。
3. 在生成模型当中，我们可以使用一些无标签数据来帮助训练，是为半监督学习 semi-supervised learning. 而在判决模型当中使用起来非常困难。
4. 在判决模型当中，我们可以做一些基本的函数扩展，例如之前在线性回归当中用$\phi(x)$替代x就得到了多项式回归；然而在生成模型当中不可以这样做。

一些常见的生成模型：

1. 朴素贝叶斯
2. 高斯判别分析
3. K 近邻算法
4. 混合高斯模型 Gaussian Mixture Model
5. 隐马尔科夫模型 Hidden Markov Model
6. 贝叶斯网络
7. Sigmoid 信念网络
8. 深度信念网络 Deep Belief Network
9. 马尔科夫随机场 Markov Random Fields

一些常见的判别模型：

1. 逻辑回归，包括二分类逻辑回归和 SoftMax
2. 线性回归
3. 神经网络 NN
4. 支持向量机 Support Vector Machine
5. 高斯过程 Gaussian Process
6. 条件随机场 Conditional Random Fields
7. CART Classification and Regression Tree

离散数据的生成模型

以上我们考虑的数据集$D=\{\boldsymbol{x}_i,y_i\}_{i=1}^N$，其中$y_i\in\{1,2,\dots,C\}$且$\boldsymbol{x}_i\in\mathbb{R}^D$为连续值特征 continuous-value features. 高斯判决分析当中我们假设

$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta})=N(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$$

每个类别的参数个数为$D+\frac{D(D+1)}{2}$，我们可以通过独立假设来减少参数个数：

$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta})=\prod_{j=1}^{D}p(x_j|y=c,\boldsymbol{\theta}_{jc})$$

进一步假设x_j服从高斯分布，则有

$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta})=\prod_{j=1}^{D}N(x_j|\mu_{jc},\sigma_{jc}^2)$$

这等同于协方差矩阵为对角阵的高斯判决分析，参数个数从$D+\frac{D(D+1)}{2}$减少为$2D$.

接下去我们考虑离散数据集$D=\{\boldsymbol{x}_i,y_i\}_{i=1}^N$，其中$y_i\in\{1,2,\dots,C\}$且$\boldsymbol{x}_i\in\{1,2,\dots,K\}^D$. 对于每个类别，其联合分布概率为

$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta})=p(x_1,x_2,\dots,x_D|y=c,\boldsymbol{\theta})$$

当所有的特征均为二进制时，参数个数为2^D-1，参数个数过多不利于计算。因此我们通过独立假设来减少参数个数。

朴素贝叶斯模型：

朴素贝叶斯当中我们假设所有特征条件独立 conditionally independent，从而

$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta})=\prod_{j=1}^{D}p(x_j|y=c,\boldsymbol{\theta}_{jc})$$

对于 binary feature，我们可以用伯努利分布代替：

$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta})=\prod_{j=1}^{D}Ber(x_j|\mu_{jc})$$

其中μ_{jc}代表特征x_j出现在类别c当中的概率。
在每个特征取值 >2 个时，我们可以采用分类分布 categorical distribution 代替:

$$p(\boldsymbol{x}|y=c,\boldsymbol{\theta})=\prod_{j=1}^{D}Cat(x_j|\boldsymbol{\mu}_{jc})$$

其中$\boldsymbol{\mu}_{jc}=\left(\mu_{jc1,}\,\mu_{jc2},\dots,\mu_{jcK}\right)^T,\mu_{jck}$代表特征$x_j$取值$k$出现在类别$c$当中的概率.
朴素贝叶斯的参数估计：仍然采用 MLE，其对数似然函数为

$$\log p(D|\boldsymbol{\theta})=\sum_{i=1}^{N}\log p(\boldsymbol{x}_i,y_i|\boldsymbol{\theta})$$

$$=\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)\log\pi_c+\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{i:y_i=c}\log p(x_{ij}|\boldsymbol{\mu}_{jc})$$

为了最大化似然函数，我们分别最大化其展开式的两项：
$\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i=c)\log\pi_c$以及$\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{i:y_i=c}\log p(x_{ij}|\boldsymbol{\mu}_{jc})$
根据吉布斯不等式，可以得到：

$$\widehat{\pi_c}=\frac{N_c}{N}$$

其中$N_c=\Sigma_{i=1}^{N}\mathbb{I}(y_i=c)$是类别$c$的个数/size，$N=\Sigma_C N_C$
从而有

$$\widehat{\mu_{jck}}=\frac{N_{jck}}{N_c}$$

其中N_{jck}代表类别c当中$x_{ij}=k$的个数。
拉普拉斯平滑 Laplace Smoothing：实际运用以上公式时可能会有$N_c=0$的情况出现，此时采用平滑参数 smoothing parameter $\alpha>0$来避免该问题：

$$\widehat{\mu_{jck}}=\frac{N_{jck}+\alpha}{N_c+K\alpha}$$