LDA和QDA的分类原理

0. 参考资料

CSDN [判别模型：logistic,GDA,QDA（一）]

https://esl.hohoweiya.xyz/04-Linear-Methods-for-Classification/4.3-Linear-Discriminant-Analysis/index.html

scikit-learn官网 https://scikit-learn.org/stable/modules/lda_qda.html#mathematical-formulation-of-the-lda-and-qda-classifiers

吴恩达**CS229**讲义中相关内容

《**ESL**》一书中对应章节

1. LDA

线性判别分析（Linear Discriminant Analysis, LDA）和二次判别分析（Quadratic Discriminant Analysis, QDA）都是分类的生成模型，它们的原理与贝叶斯法则有关。

对于分类问题而言，我们需要求得后验概率$P(y=k|X)$，其中$k=1,\cdots,K$代表类别。利用贝叶斯公式可以求得后验概率，即：

$$P(y=k|X)=\frac{P(X|y=k)P(y=k)}{P(X)}=\frac{P(X|y=k)P(y=k)}{\sum_{l=1}^{K}P(X|y=l)P(y=l)}$$

对于LDA和QDA，我们假设$P(X|y=k)$服从多元高斯分布，先验概率$P(y=k)=\pi_k$，这时又称为高斯判别分析（GDA）。

此时有：

$$\begin{aligned}P(X|y=k)&=f_k(x)\\&=\frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(X-\mu_k)^t\Sigma_k^{-1}(X-\mu_k)\right)\end{aligned}$$

其中n时特征维度。

假设所有类别都具有相同的协方差矩阵$\Sigma_k=\Sigma$，会导出LDA。在比较两个类别k和l时，考虑它们的对数比率即可，所以有

$$\begin{aligned}&\log\frac{\Pr(y=k|X=x)}{\Pr(y=\ell|X=x)}\\=&\log\frac{f_k(x)}{f_\ell(x)}+\log\frac{\pi_k}{\pi_\ell}\\=&\log\frac{\pi_k}{\pi_\ell}-\frac{1}{2}(\mu_k+\mu_\ell)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell)+x^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_\ell)\end{aligned}\tag{1.1}$$

这是个关于x的线性等式，协方差矩阵相等消除了二次项，这也是名称“线性判别分析”的由来。

根据式（1.1）可以看出线性判别函数为：

$$\delta_k(x)=x^T\mathbf{\Sigma}^{-1}\mu_k-\frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k+\log\pi_k$$

然后类别$k^*=\operatorname{argmax}_k\delta_k(x)$，这是判别规则的等价描述。

我们可以通过最大后验的方式从数据集中估计参数：

- $\hat{\pi}_k=N_k/N$，其中N_k是第k类观测值的个数；
- $\hat{\mu}_k=\sum_{i:y_i=k}x_i/N_k$；
- $\hat{\mathbf{\Sigma}}=\sum_{k=1}^K\sum_{i:y_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T/(N-K)$。

2. QDA

如果我们假设每一类别的协方差矩阵不同，那么式（1.1）中方便的抵消就不会发生，这时就是QDA的情况。

此时判别函数变为：

$$\begin{aligned}\delta_k(x)&=-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)-\frac{1}{2}\log|\Sigma_k|+\log\pi_k\\&=-\frac{1}{2}x^T\Sigma_k^{-1}x+x^T\mathbf{\Sigma}_k^{-1}\mu_k-\frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k-\frac{1}{2}\log|\Sigma_k|+\log\pi_k\end{aligned}$$

这时$\delta_k(x)$是关于x的二次函数。每个类别对k和l的判别边界由二次等式$\{x:\delta_k(x)=\delta_\ell(x)\}$来描述。

3. 与LR的联系

可以证明，LDA的后验概率能被表示成logistic回归的形式。那么应该如何选择LDA和LR呢？

注意到LDA假设数据服从多元高斯分布，而某些不服从高斯分布的数据也可以导出LR的形式。因此，LDA的假设要比LR更强。

当数据确实服从多元高斯分布时，使用LDA会带来更好的效果，并且对数据的利用率更高（即只需要少量数据就能学习的很好）。但当数据不服从高斯分布时，强行使用LDA效果不佳，这是LR的表现会更鲁棒，因此实际操作时，我们常常更偏向于使用LR。