高斯判别分析(GDA)和朴素贝叶斯(NB)

前言

我们在前边学习了一般线性回归和相关推论与指数族分布和相关推论，今天我们得换换口味了，再按照这条线推下去各位迟早得疯掉（~~我也得疯掉）实际上已经疯了~~）...

所以今天就来整点别的——高斯判别分析(GDA)和朴素贝叶斯(NB)（以下可能均简称）。两者都是不同于逻辑回归实现分类问题的解决方法。它们优点、缺点和用法不一。

这里需要解释的是，以往我们学习的那些回归都属于**判别学习模型（线性回归、逻辑回归、softmax回归...）**，而今天的这两个都是属于**生成学习模型**，详细两者的区别在下文有详细解释。

先来简单说一下GDA和NB的一些优缺点（混个眼熟，下面还会讲）：

1. GDA
    a. 优点：
        i. 通过协方差矩阵捕获特征间依赖关系，需要较强相关性。
        ii. 概率估计精确，输出条件概率 $P(Y \mid X)$ 可靠性高。
        iii. 低维数据效率高，分类性能优于部分判别模型。
    b. 缺点：
        i. 高斯分布假设强，所以非高斯数据性能会显著下降。
        ii. 如果计算复杂度高，协方差矩阵求逆复杂度为 $O(d2)$，在高维数据不可行。
        iii. 各类别协方差差异大时模型偏差严重。
2. NB
    a. 优点：
        i. 计算效率极高，其训练和预测复杂度 $O(d)$，适用于高维数据（如文本）。
        ii. 小样本鲁棒性强，不易过拟合，数据稀疏时仍有效。
        iii. 支持多数据类型，可扩展为高斯NB、多项NB等变体。
    b. 缺点：
        i. 条件独立假设不成立时会导致性能劣化，特征强相关时概率估计失真。
        ii. 若概率输出未校准，需额外校准才能获得可靠置信度。
        iii. 过度依赖假设（如用高斯NB处理计数数据），会导致偏差。

关于判别学习模型和生成学习模型

上边我们提到了**判别学习模型和生成学习模型**，我们需要做出一些详细解释：

1. 判别模型是直接从输入特征到输出类别进行建模，试图找到数据特征和类别标签之间的最佳决策边界。
2. 而生成模型则试图学习输入数据的分布，并生成或重建新的数据实例。这些模型可以捕捉特征之间的潜在关系，并对特征进行建模。**这种关系在下述的高斯判别分析中，体现为两个（或者多个）位置不同（均值）但大小形状（协方差矩阵）可能相同的多元高斯分布，其中每一个分布都是一个类别。**

或者我们可以不引用专业解释，直接使用大白话介绍给各位：**判别学习模型侧重于x和y的映射关系，最终拟合出那条线，比如线性回归的那条特征值所在线，逻辑回归的那条分界线。但是生成学习模型则是通过数据特征，判断它更贴合哪一部分，在分类问题中尤为经典（因为下述这两个就是分类的算法（**

判别学习模型更关注“如何区分”（通过分界线或映射关系），而生成学习模型更关注“数据是如何生成的”（通过建模每个类别的分布），这就是二者的区别

再深入细说生成学习模型，它形式化的表示就是是否使用了贝叶斯公式（没错，就是高中的那个贝叶斯）：

$$\max P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$
$$\sim \max P(X|Y)P(Y)$$

生成学习模型是通过贝叶斯模型最大后验概率 $P(X|Y)P(Y)$（联合概率分布），其中 $P(X|Y)$ 可以看作是从标签d生成数据，$P(Y)$ 则是标签的先验概率。*（说白了就是用“已观测到的果”去推断“最可能的因”，高中知识，不懂的可以重修贝叶斯了）*

我想我解释的应该足够通俗易懂了，接下来我们就试着探索一下这两个模型

高斯判别分析（GDA）

　　高斯判别分析（Gaussian Discriminant Analysis, GDA），不要被它的名字迷惑了，虽然带了判别二字，实际上是生成学习模型，我们在最后会解释为什么（**问题一：为什么高斯判别分析是生成学习模型**）。它是基于**联合概率分布**来推导的。具体来说，假设数据的联合概率分布 $p(x,y)$、可以分解为条件概率（在谁的前提条件下成立） $p(x|y)$ 和先验概率 $p(y)$ 的乘积，所以我们就得到了：

$$p(x,y) = p(x|y) \cdot p(y)$$ （其实就是高中的概率乘法公式

什么是联合概率分布？名字很高大上，实际上就是两个或多个随机变量（或随机事件）同时发生的概率分布。或者你可以这么理解：两个骰子同时掷出，他们的样本空间是 6*6 的样本元素组成，这就是联合概率分布，由两个或者多个骰子的笛卡尔乘积组成（问题二：联合概率分布和贝叶斯公式的关系，为什么我要放在一起说）

现在我们知道了大前提，那么先看看推导吧：

DeepSeek Prompt：你现在是一个人工智能大学教授，你要给你的本科学生讲述高斯判别分析，需要使用Latex给出高斯判别分析GDA的详细推导流程

AI：

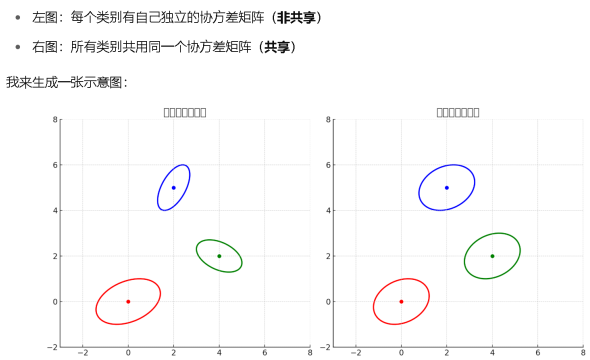
以下是关于高斯判别分析（Gaussian Discriminant Analysis, GDA）的详细推导流程，采用LaTeX格式呈现。作为人工智能教授，我将从生成模型的角度系统推导GDA的参数估计和决策边界。

Me：我们的GDA可以是二元分类，也可以是多元，如果是二元其中的协方差矩阵可以不共享，只不过计算量会变大，而且最终决策边界可能不是线性的。这里我们以二元分类作为推导。

高斯判别分析（GDA）推导

问题设定：

- 二分类任务：标签 $y \in 0,1$
- 输入特征： $\mathbf{x} \in \mathbb{R}^d$
- **生成模型假设：**
  a. 类先验服从伯努利分布：
     i. $y \sim \text{Bernoulli}(\phi) \implies p(y) = \phi^y (1-\phi)^{1-y}$
  b. 类条件概率服从多元高斯分布，且协方差矩阵共享：
     i. 什么是协方差矩阵？
     ii. 你会在多元高斯分布概率密度推导找到答案（其实就是矩阵 $\boldsymbol{\Sigma}$ ）
        1. 在分类模型如**LDA（线性判别分析）中，假设各类别的高斯分布共享协方差矩阵，**可以简化计算，并使得分类边界是**线性的**。
        2. 如果不共享，就类似于**QDA（二次判别分析）**，边界是曲线，计算更复杂。
        3. 详细的样子可以看看cs229的第六个视频（不用详细看，他没最大似然部分的推导（



     iii. $p(\mathbf{x} \mid y=k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad k \in 0,1$

也就是说：

1. 伯努利： $y \sim \text{Bernoulli}(\phi) \implies p(y) = \phi^y (1-\phi)^{1-y}$
2. 高斯： $p(\mathbf{x} \mid y=k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad k \in 0,1$
3. 由联合概率分布，得：

$$p(\mathbf{x}, y) = \begin{cases} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}1, \boldsymbol{\Sigma}) \cdot \phi & \text{if } y = 1 \\ \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}0, \boldsymbol{\Sigma}) \cdot (1 - \phi) & \text{if } y = 0 \end{cases}$$

参数估计（最大化联合似然估计）

给定训练集 $(\mathbf{x}^{(i)}, y^{(i)})_{i=1}^{m}$ ，似然函数为：

$$L(\phi, \boldsymbol{\mu}0, \boldsymbol{\mu}1, \boldsymbol{\Sigma}) = \prod i = 1^m p(\mathbf{x}^{(i)}, y^{(i)})$$

分解：

$$L = \prod_{i=1}^{m} \underbrace{p(\mathbf{x}^{(i)} \mid y^{(i)})}_{\text{Gaussian高斯}} \cdot \underbrace{p(y^{(i)})}_{\text{Bernoulli伯努利}}$$

关于多元高斯分布推导：多元高斯分布概率密度推导

似然函数为：

$$L(\theta) = \prod_{i=1}^{m} p(\mathbf{x}^{(i)}, y^{(i)}; \theta) \quad (\theta = \phi, \boldsymbol{\mu}0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

代入拆解：

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} \left[\mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \cdot \phi\right]^{(y)^{(i)}} \cdot \left[\mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) \cdot (1 - \phi)\right]^{1-y^{(i)}} \\ &= \underbrace{\left[\prod_{i:y^{(i)}=1} \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right]}_{\text{Class 1 likelihood}} \cdot \underbrace{\left[\prod_{i:y^{(i)}=0} \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma})\right]}_{\text{Class 0 likelihood}} \cdot \underbrace{\phi^{m_1}(1 - \phi)^{m_0}}_{\text{Class prior}} \end{aligned}$$

因为 $m_1 = \sum_{i=1}^{m} y^{(i)}$ （类别1样本数），$m_0 = m - m_1$ （类别0样本数）

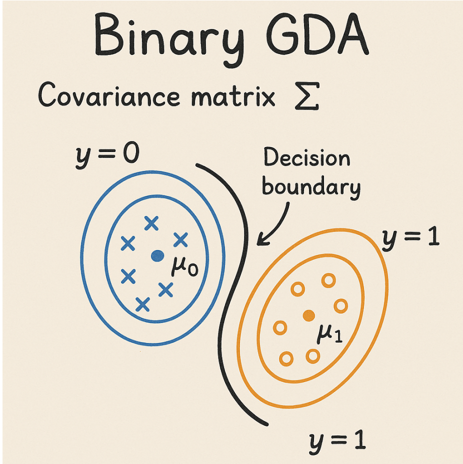
对数似然函数为：

$$\ell(\theta) = \sum_{i:y^{(i)}=1} \log \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \sum_{i:y^{(i)}=0} \log \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) + m_1 \log \phi + m_0 \log(1 - \phi)$$

之后我们需要对 ϕ、μ_0、μ_1、Σ 四个参数做估计，这四个参数的实际作用我相信大家还是云里雾里的，所以这里需要解释一下：

1. ϕ 是联合分布的时候，我们的伯努利分布的参数，也就是一个事件发生or不发生的概率
2. μ_0和μ_1 是针对于分类的两个高斯分布的中间点，相当于一元高斯的均值，只不过我们这里是多元，存在多个均值，代表了不同的类别

1. Σ 则是协方差矩阵，上述介绍了，这里再说一次。它代表了各个高斯分布的形状和大小，描述了每个类别的分布状态（这里两个协方差矩阵共用一个 Σ），如果不共享，那么两个类别的高斯分布最终决策边界线将呈现非线性关系。
    a. 害怕你们看不懂，所以主包动用了“黑魔法”，让AI给你们画了一张决策边界图。Prompt：帮我画一张详细的带有二元GDA和决策边界的描述图



至于可能各位会问：**（雷三：为什么对这四个值进行似然估计？生成学习模型的似然函数与判别学习模型的似然函数有什么区别？）**我会在最后统一解答。

1. **估计类先验参数 ϕ**

目标：最大化对数似然中与 ϕ 相关的部分

$$\ell_\phi = m_1 \log\phi + m_0 \log(1-\phi)$$

求导过程：

1. 对 ϕ 求导：

$$\frac{\partial \ell_\phi}{\partial \phi} = \frac{m_1}{\phi} - \frac{m_0}{1-\phi}$$

2. 令导数为零：

$$\frac{m_1}{\phi} = \frac{m_0}{1-\phi} \implies m_1(1-\phi) = m_0\phi$$

3. 解方程：

$$m_1 - m_1\phi = m_0\phi \implies m_1 = \phi(m_0+m_1) \implies \phi = \frac{m_1}{m}$$

4. **结论：**

$$\hat{\phi} = \frac{\text{类别1样本数}}{\text{总样本数}}$$

2. **估计均值向量 $\boldsymbol{\mu}0, \boldsymbol{\mu}1$**

目标：最大化高斯密度部分的对数似然

对类别 (k)，其对数似然为：

$$\ell_{\boldsymbol{\mu}_k} = \sum_{i:y^{(i)}=k} \log \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

代入多元高斯密度函数（上文有多元高斯分布推导）：

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

取对数后（忽略常数项）：

$$\log\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+C$$

求导过程（使用矩阵微分法则）：

1. 定义二次型：$Q=(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$
2. 对 $\boldsymbol{\mu}$ 的梯度：

$$\frac{\partial Q}{\partial\boldsymbol{\mu}}=-2\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\quad(\text{参考：}\frac{\partial\mathbf{a}^\top\mathbf{B}\mathbf{a}}{\partial\mathbf{a}}=2\mathbf{B}\mathbf{a})$$

3. 对 $\ell\boldsymbol{\mu}_k$ 求导：

$$\frac{\partial\ell\boldsymbol{\mu}_k}{\partial\boldsymbol{\mu}_k}=\sum_{i:y^{(i)}=k}\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k)=0$$

4. 解得：

$$\sum_{i:y^{(i)}=k}\mathbf{x}^{(i)}=\boldsymbol{\mu}_k\cdot m_k\implies\hat{\boldsymbol{\mu}}_k=\frac{1}{m_k}\sum_{i:y^{(i)}=k}\mathbf{x}^{(i)}$$

3. **估计共享协方差矩阵 $\boldsymbol{\Sigma}$**

对数似然中与 $\boldsymbol{\Sigma}$ 相关的部分：

$$\ell\boldsymbol{\Sigma}=-\frac{m}{2}\log|\boldsymbol{\Sigma}|-\frac{1}{2}\sum_{i=1}^{m}(\mathbf{x}^{(i)}-\boldsymbol{\mu}_{y^{(i)}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu}_{y^{(i)}})$$

定义类内散度矩阵：

$$\mathbf{S}_k=\sum_{i:y^{(i)}=k}(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k)(\mathbf{x}^{(i)}-\boldsymbol{\mu}_k)^\top,\quad\mathbf{S}=\mathbf{S}_0+\mathbf{S}_1$$

代入并利用迹的性质：

$$\ell\boldsymbol{\Sigma}=C-\frac{m}{2}\log|\boldsymbol{\Sigma}|-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)$$

求导并令导数为零（使用矩阵微分）：

$$\frac{\partial\ell}{\partial\boldsymbol{\Sigma}}=-\frac{m}{2}\boldsymbol{\Sigma}^{-\top}+\frac{1}{2}\boldsymbol{\Sigma}^{-\top}\mathbf{S}\boldsymbol{\Sigma}^{-\top}=0$$

解得：

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{m}\mathbf{S}=\frac{1}{m}\sum k=0^1\sum i:y^{(i)}=k(\mathbf{x}^{(i)}-\hat{\boldsymbol{\mu}}k)(\mathbf{x}^{(i)}-\hat{\boldsymbol{\mu}}_k)^\top$$

决策边界推导

简单理解为楚汉河界（落在哪个范围，就是谁的

预测新样本 $\mathbf{x}$ 的类别：

$$p(y=1\mid\mathbf{x})=\frac{p(\mathbf{x}\mid y=1)p(y=1)}{p(\mathbf{x})}$$

判别函数为对数几率：

$$\log\frac{p(y=1\mid\mathbf{x})}{p(y=0\mid\mathbf{x})}=\log\frac{p(\mathbf{x}\mid y=1)}{p(\mathbf{x}\mid y=0)}+\log\frac{\phi}{1-\phi}$$

代入高斯密度函数：

$$\log\frac{\mathcal{N}(\mathbf{x}\mid\boldsymbol{\mu}1,\boldsymbol{\Sigma})}{\mathcal{N}(\mathbf{x}\mid\boldsymbol{\mu}0,\boldsymbol{\Sigma})}=-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}0)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}0)$$

展开二次型并合并同类项：

$$=\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}1-\boldsymbol{\mu}0)-\frac{1}{2}\boldsymbol{\mu}1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}1+\frac{1}{2}\boldsymbol{\mu}0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}0$$

决策边界为线性函数：

$$\underbrace{\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}1-\boldsymbol{\mu}0)}_{\text{线性项}}+\underbrace{\log\frac{\phi}{1-\phi}-\frac{1}{2}(\boldsymbol{\mu}1+\boldsymbol{\mu}0)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}1-\boldsymbol{\mu}0)}_{\text{常数项}}=0$$

问题

1. **为什么高斯判别分析是生成学习模型**

   Kimi+me：事实上，如果你知道了生成学习模型和判别学习模型的区别，知道高斯判别分析推导，那这个问题就不攻自破了，因为它建模的是“数据+标签的联合概率” $p(X,y)$ ，而不是只建模条件概率 $p(y\mid X)$ 。

2. **联合概率分布和贝叶斯公式的关系，为什么我要放在一起说**

   Kimi：联合概率分布是贝叶斯公式能够成立的“原材料”；贝叶斯公式只是把这张联合分布“切片”或“归一化”，得到条件分布。

   a. **先给出联合分布**
      设随机变量/参数 θ 与观测数据 D 的联合分布为

$$p(\theta,D)=p(\theta)p(D\mid\theta).$$

   这里 $p(\theta)$ 是边缘分布（先验）， $p(D\mid\theta)$ 是似然；二者一乘就得到联合分布。

   b. **用联合分布“切”出条件分布**

   从联合分布里提取条件分布的通用公式是

$$P(\theta|D)=\frac{P(\theta,D)}{P(D)}=\frac{p(\theta)p(D\mid\theta)}{\int p(\theta\prime)p(D\mid\theta\prime)d\theta\prime}$$

   这正是贝叶斯公式。

   - 分子 $p(\theta,D)$ 是联合分布在点 (θ,D) 上的值；

- 分母 $p(D)$ 是对 θ 的边缘化，用来归一化，使整个后验积分为 1。

所以联合分布是“全局地图”；贝叶斯公式是“查表+归一化”的工具，用来从这张地图中提取给定数据后的条件分布。

3. **为什么对这四个值进行似然估计？生成学习模型的似然函数与判别学习模型的似然函数有什么区别？**

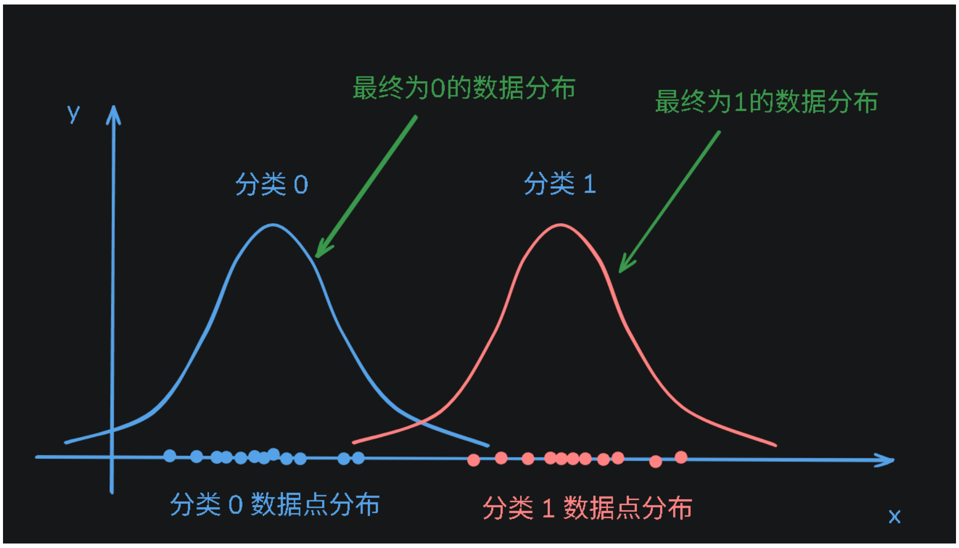
Kimi+me：

a. **结合上边问题一，我们可以知道生成学习模型是对 “数据+标签的联合概率”进行建模，而不只是条件概率，所以在以往的那几个回归中，我们都是似然条件概率公式。因此，GDA 通过似然估计这四个参数来最大化联合似然** $\prod_{i=1}^{n} P(x^{(i)}, y^{(i)})$ **。**

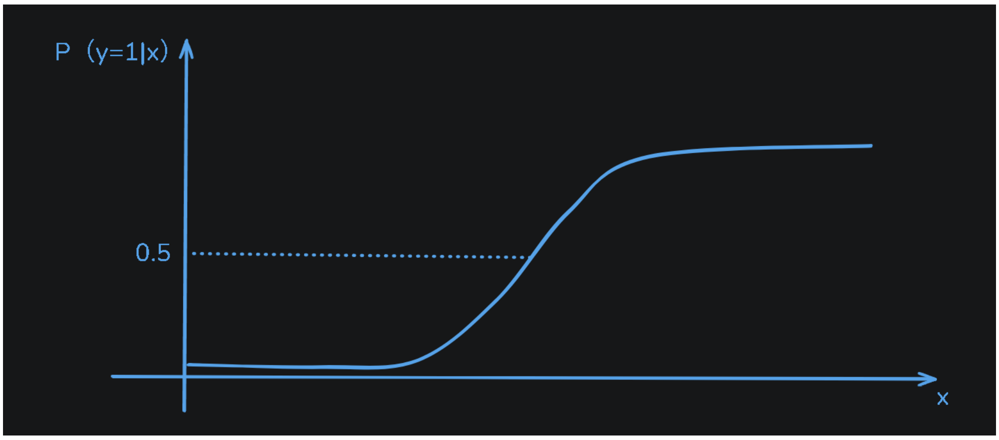
b. **生成学习模型使用的是建模联合分布** $P(x,y) = P(x \mid y)P(y)$ **，而判别学习模型直接建模条件分布** $P(y \mid x)$ **，忽略** $P(x)$

与逻辑回归和softmax回归

我们上述介绍的GDA相关内容，最终我们得出了决策边界，这个边界划分了分类的标准，线的那边是哪一类，另一边又是哪一类。事实上，我们可以将这种形式“拍成”平面，以二维平面的形式展开，它的形式大致为：



然而当我们将坐标轴改一下，它就变成了这样：



好熟悉...没错，这就是我们之前推导过的sigmoid函数，也就是S型函数。所以**GDA和逻辑回归本质机制都是使用sigmoid函数来计算** $P(y=1|x)$。**（尽管GDA实际上使用bayes rule计算而不是计算一个sigmoid函数。）**

那我们就该吐槽了：“本质相同，为什么还要整这么多费劲的拐弯抹角？”

实际上这事不是这么讲的，我们需要知道的是，**逻辑回归的实际使用场景，它需要的数据量通常是要比GDA大，因为我们无需前提假设模型是什么，而GDA不是，它假设了我们的数据符合高斯分布，这种假设的前提下，我们的数据只要不脱离高斯，那GDA就是要比逻辑回归更好，因为GDA不涉及什么迭代和求导，这是高效的关键，再加上无需考虑最终模型。这是逻辑回归没办法比的。**

但是话又说回来，我们往往没有办法强假设数据符合什么分布，我们都是在拆盲盒。段子：“你会带着你的疑惑去问医院负责人员：‘你们医院的这50个病例单是符合高斯分布的么？’这样做确实可以快速确认模型，高效的搭建GDA，但这样做很傻很蠢”🤗

至于SoftMax，和逻辑回归的异同类似，这里也就不多赘述了。

朴素贝叶斯（NB）

这个我就不讲了，很简单的一个算法模型，直接看这篇文章吧，我觉得他写的非常棒，例子很浅显易懂了：

https://zhuanlan.zhihu.com/p/26262151
zhuanlan.zhihu.com

GDA和朴素贝叶斯并不是最准确的算法，但它们的优势在于**训练速度很快，它们是非迭代算法，只是通过计数来学习参数生成学习模型进行预测**，如GDA只是计算均值和协方差，所以它们很有竞争力而且很容易实现。