第 4 章：应用（2）

模式识别问题

授课教师：曹语

课程主页：https://yucaoyc.github.io/math3806

从预测问题出发

体检报告里常见很多指标：年龄、血压、血糖、BMI 等。

问题：我们能否用这些变量来判断某个人更可能属于“患病”还是“健康”这一类？

从预测问题出发

体检报告里常见很多指标：年龄、血压、血糖、BMI 等。

问题：我们能否用这些变量来判断某个人更可能属于“患病”还是“健康”这一类？

答案：可以。我们会把这些指标记成特征向量 $\boldsymbol{x}$，再通过一个模型来预测类别。

从优化的角度看，真正的问题是：

如何选择模型参数，使错误少、边界稳，并且预测结果容易解释？

分类问题的抽象

今天这类问题会被抽象成一个分类问题：

- 输入：一个特征向量 $\boldsymbol{x}$
- 输出：类别 0 或类别 1
- 目标：设计一个函数，尽量把两类样本分开

本节课目标

学完本节课后，你应该能够：

- 解释线性分类器与分离超平面的关系
- 理解最大间隔与软间隔支持向量机的优化形式
- 写出 Logistic 回归的概率模型、似然函数与凸优化形式
- 区分模型训练、参数选择与性能评估的基本作用

课前回顾：凸集与分离

问题：回忆第 2 章：给定两个凸集集合，什么时候可用超平面把它们分开?

课前回顾：凸集与分离

问题：回忆第 2 章：给定两个凸集集合，什么时候可用超平面把它们分开?

答案：超平面分离定理保证存在一个超平面 $\{\boldsymbol{x} : \boldsymbol{a}^\top \boldsymbol{x} + b = 0\}$，使两个不相交凸集分别落在半空间 $\{\boldsymbol{x} : \boldsymbol{a}^\top \boldsymbol{x} + b \geq 0\}$ 和 $\{\boldsymbol{x} : \boldsymbol{a}^\top \boldsymbol{x} + b \leq 0\}$ 中。

模式识别/分类问题

给定数据 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$ 和 $\{\boldsymbol{x}_{m+1}, \cdots, \boldsymbol{x}_M\}$，我们希望找到函数 f 使得

$$\begin{aligned} f(\boldsymbol{x}_i) &> 0, \qquad i = 1, \cdots, m \\ f(\boldsymbol{x}_i) &< 0, \qquad i = m+1, \cdots, M. \end{aligned} \tag{1}$$

应用： 视觉识别、医疗诊断、违约风险预测、缺陷检测、虫害识别等

目标： 如何找到这样的 f?

- 线性函数
- 支持向量分类器
- Logistic 回归

目录

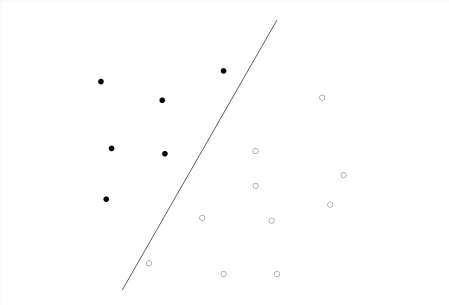
测试：线性分类器

选择 1：线性 $f(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} + b$

如果我们能找到这样的线性函数完美实现公式(1)的要求，则该平面被称为严格分离超平面。

线性可分示意

该直线严格分离两组数据，因此可作为分类器



从严格分离到标准间隔

若公式(1)的问题有可行解，则所有样本离分界面都有正的函数间隔。

进而可以把约束右端统一写成 1 和 -1。

令

$$\delta = \min\left\{\min_{1\leq i\leq m} f(\boldsymbol{x}_i),\ \min_{m+1\leq i\leq M} -f(\boldsymbol{x}_i)\right\} > 0.$$

把原来的 $(\boldsymbol{a}, b)$ 缩放为 $(\boldsymbol{a}/\delta, b/\delta)$，分类符号不变，并得到

$$\begin{aligned}
&\boldsymbol{a}^\top \boldsymbol{x}_i + b \geqslant 1, \quad i = 1, \cdots, m,\\
&\boldsymbol{a}^\top \boldsymbol{x}_i + b \leqslant -1, \quad i = m+1, \cdots, M.
\end{aligned}$$

提示：分类只看符号，整体缩放不改变分类结果，却可把最小函数间隔规范化为 1。

问题：由于这样的直线不是唯一的，那么我们该如何找到最优分离超平面呢?

最大间隔思想

我们希望找到 $\boldsymbol{a}, b$ 使得 $\|\boldsymbol{a}\|_2$ 足够小，即最小化

$$\begin{aligned} &\text{minimize} && \|\boldsymbol{a}\|_2 \\ &\text{subject to} && \boldsymbol{a}^\top \boldsymbol{x}_i + b \geq 1, \quad i = 1, \cdots, m \\ & && \boldsymbol{a}^\top \boldsymbol{x}_i + b \leq -1 \quad i = m+1, \cdots, M \end{aligned} \tag{2}$$

原因是因为这样的选择可以使得结果对于 $\boldsymbol{x}_i$ 的小扰动更加稳定。

比如数据 $\boldsymbol{x}_i$ 带有一点小偏差，上述选择可以使得 $\boldsymbol{a}^\top \boldsymbol{x}_i$ 应对偏差更加稳定。

最大间隔思想

我们希望找到 $\boldsymbol{a}, b$ 使得 $\|\boldsymbol{a}\|_2$ 足够小，即最小化

$$\begin{aligned} &\text{minimize} && \|\boldsymbol{a}\|_2 \\ &\text{subject to} && \boldsymbol{a}^\top \boldsymbol{x}_i + b \geq 1, \quad i = 1, \cdots, m \\ & && \boldsymbol{a}^\top \boldsymbol{x}_i + b \leq -1 \quad i = m+1, \cdots, M \end{aligned} \tag{2}$$

我们可以将问题转变为：

$$\begin{aligned} &\text{minimize} && \|\boldsymbol{a}\|_2 \\ &\text{subject to} && \left(\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}\right)^\top \boldsymbol{x}_i + \frac{b}{\|\boldsymbol{a}\|_2} \geq \frac{1}{\|\boldsymbol{a}\|_2}, \quad i = 1, \cdots, m \\ & && \left(\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}\right)^\top \boldsymbol{x}_i + \frac{b}{\|\boldsymbol{a}\|_2} \leq \frac{-1}{\|\boldsymbol{a}\|_2} \quad i = m+1, \cdots, M \end{aligned}$$

最大间隔思想

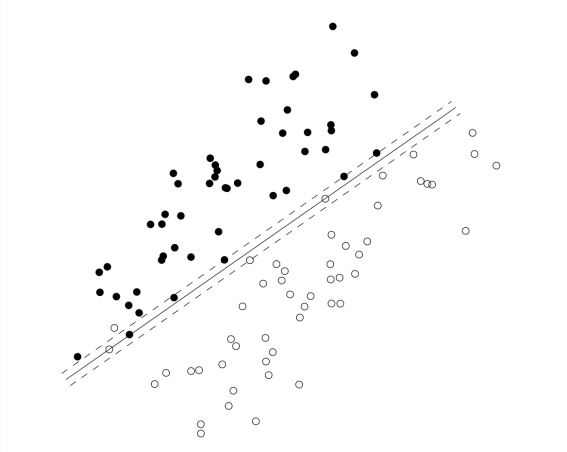
通过换元可得：

$$
\begin{aligned}
&\text{maximize} && t \\
&\text{subject to} && \boldsymbol{a}^\top \boldsymbol{x}_i + b \geq t, \quad i = 1, \cdots, m \\
& && \boldsymbol{a}^\top \boldsymbol{x}_i + b \leq -t \quad i = m+1, \cdots, M \\
& && \|\boldsymbol{a}\|_2 \leqslant 1
\end{aligned}
$$

目录

线性不可分情形

现实问题中很难具有完美的分类器



从硬间隔到软间隔

在线性可分时，我们希望所有样本都满足标准间隔约束。这种完全不允许违反的要求称为硬间隔：

$$\begin{aligned}
&\boldsymbol{a}^{\top}\boldsymbol{x}_i + b \geq 1, \quad i = 1, \cdots, m, \\
&\boldsymbol{a}^{\top}\boldsymbol{x}_i + b \leq -1, \quad i = m+1, \cdots, M.
\end{aligned}$$

软间隔：松弛变量的含义

引入松弛变量（即一定的容错）$u_1, \cdots, u_M \geq 0$，把硬间隔约束放宽为**软间隔**：

$$\begin{aligned} \boldsymbol{a}^\top \boldsymbol{x}_i + b &\geq 1 - u_i, \quad i = 1, \cdots, m, \\ \boldsymbol{a}^\top \boldsymbol{x}_i + b &\leq -1 + u_i, \quad i = m+1, \cdots, M \end{aligned}$$

- $u_i = 0$：样本满足间隔要求
- $0 < u_i < 1$：分类仍正确，但离分界面太近
- $u_i > 1$：该样本可能被分错

软间隔：先惩罚违反量

当然，我们希望容错越小愈好，因此启发我们求解

$$
\begin{aligned}
\text{minimize} \quad & \mathbf{1}^\top \boldsymbol{u} \\
\text{subject to} \quad & \boldsymbol{a}^\top \boldsymbol{x}_i + b \geq 1 - u_i, \quad i = 1, \cdots, m, \\
& \boldsymbol{a}^\top \boldsymbol{x}_i + b \leq -1 + u_i, \quad i = m+1, \cdots, M \\
& \boldsymbol{u} \succeq 0
\end{aligned}
$$

软间隔支持向量机

软间隔支持向量分类器（Support Vector Machine, SVM）：

为了平衡增大间隔（见公式(2)）和减少松弛量（见上一页），我们考虑标准形式

$$
\begin{array}{ll}
\text{minimize} & \frac{1}{2}\|\boldsymbol{a}\|_2^2 + \gamma \mathbf{1}^\top \boldsymbol{u} \\
\text{subject to} & \boldsymbol{a}^\top \boldsymbol{x}_i + b \geq 1 - u_i, \quad i = 1, \cdots, m, \\
& \boldsymbol{a}^\top \boldsymbol{x}_i + b \leq -1 + u_i, \quad i = m+1, \cdots, M \\
& \boldsymbol{u} \succeq 0
\end{array}
$$

问题：γ 越大时，模型更重视增大间隔，还是减少违反量？

软间隔支持向量机

软间隔支持向量分类器（Support Vector Machine, SVM）：

为了平衡增大间隔（见公式(2)）和减少松弛量（见上一页），我们考虑标准形式

$$
\begin{array}{ll}
\text{minimize} & \frac{1}{2}\|\boldsymbol{a}\|_2^2 + \gamma \mathbf{1}^\top \boldsymbol{u} \\
\text{subject to} & \boldsymbol{a}^\top \boldsymbol{x}_i + b \geq 1 - u_i, \quad i = 1, \cdots, m, \\
& \boldsymbol{a}^\top \boldsymbol{x}_i + b \leq -1 + u_i, \quad i = m+1, \cdots, M \\
& \boldsymbol{u} \succeq 0
\end{array}
$$

问题：γ 越大时，模型更重视增大间隔，还是减少违反量？

答案：更重视减少违反量；γ 较小时，模型更重视让 $\|\boldsymbol{a}\|_2$ 变小，从而增大间隔。

目录

Logistic 回归

问题背景： 倘若某种疾病是否发病由变量 $\boldsymbol{x} \in \mathbb{R}^n$ 来决定，例如 $\boldsymbol{x}$ 包含体重，年龄，身高，血压等医学变量。对于一个个体，如果发病记为 1，健康记为 0。记标签随机变量为 $Y \in \{0, 1\}$，并假设它在给定特征 $\boldsymbol{x}$ 时满足

$$\mathrm{Prob}(Y = 1 \mid \boldsymbol{x}) = p(\boldsymbol{x}), \quad \mathrm{Prob}(Y = 0 \mid \boldsymbol{x}) = 1 - p(\boldsymbol{x})$$

其中 $p(\boldsymbol{x}) \in [0, 1]$。

Logistic 回归先计算线性风险分数 $s = \boldsymbol{a}^\top \boldsymbol{x} + b$，再用 Sigmoid 函数把它压缩到概率区间：

$$p(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{a}^\top \boldsymbol{x} + b\right)}{1 + \exp\left(\boldsymbol{a}^\top \boldsymbol{x} + b\right)}$$

它和直接使用 SVM 的区别是这里引入了概率的视角。

似然函数（1）

假设我们具有 M 个人的医疗数据，即

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_M \in \mathbb{R}^n, \qquad y_1, y_2, \cdots, y_M \in \{0, 1\}.$$

记

$$p_i := p(\boldsymbol{x}_i) = \frac{\exp\left(\boldsymbol{a}^\top \boldsymbol{x}_i + b\right)}{1 + \exp\left(\boldsymbol{a}^\top \boldsymbol{x}_i + b\right)}, \qquad i = 1, \ldots, M.$$

若假设样本条件独立，则似然函数为

$$L(\boldsymbol{a}, b) = \prod_{i=1}^{M} p_i^{y_i} (1 - p_i)^{1 - y_i}.$$

参与推导：一个样本的似然贡献

问题：当 $y_i = 1$ 时，上式中的哪一部分会保留下来?

问题：当 $y_i = 0$ 时，又是哪一部分?

参与推导：一个样本的似然贡献

问题：当 $y_i = 1$ 时，上式中的哪一部分会保留下来?

答案：$p_i^{y_i}(1-p_i)^{1-y_i}$ 会变成 p_i。

问题：当 $y_i = 0$ 时，又是哪一部分?

答案：这时该项会变成 $1 - p_i$。

似然函数（2）

对数似然为

$$\begin{aligned}\ell(\boldsymbol{a}, b) &= \log L(\boldsymbol{a}, b) \\ &= \sum_{i=1}^{M} \Big(y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big) \\ &= \sum_{i=1}^{M} y_i \Big(\boldsymbol{a}^\top \boldsymbol{x}_i + b \Big) - \sum_{i=1}^{M} \log \Big(1 + \exp\big(\boldsymbol{a}^\top \boldsymbol{x}_i + b \big) \Big).\end{aligned}$$

问题：该函数是凸函数、凹函数，还是都不是？

似然函数（2）

对数似然为

$$\begin{aligned}\ell(\boldsymbol{a}, b) &= \log L(\boldsymbol{a}, b) \\ &= \sum_{i=1}^{M} \Big(y_i \log p_i + (1 - y_i) \log(1 - p_i) \Big) \\ &= \sum_{i=1}^{M} y_i \Big(\boldsymbol{a}^\top \boldsymbol{x}_i + b \Big) - \sum_{i=1}^{M} \log \Big(1 + \exp\big(\boldsymbol{a}^\top \boldsymbol{x}_i + b \big) \Big).\end{aligned}$$

问题：该函数是凸函数、凹函数，还是都不是？

答案：它是关于 $\boldsymbol{a}, b$ 的凹函数，因此 $-\ell(\boldsymbol{a}, b)$ 是凸函数。

优化形式

由上一页可知，$\ell(\boldsymbol{a}, b)$ 是关于 $\boldsymbol{a}, b$ 的凹函数。

因此最大化对数似然等价于求解无约束凸优化问题

$$\min_{\boldsymbol{a}, b} -\ell(\boldsymbol{a}, b).$$

提示：比较梯度法与牛顿法在该问题上的效果，是一个很好的课程大作业题目。

预测规则

我们可以根据

$$p(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{a}^\top \boldsymbol{x} + b\right)}{1 + \exp\left(\boldsymbol{a}^\top \boldsymbol{x} + b\right)}$$

来预测新数据 $\boldsymbol{x}$ 取到 1 的概率。若该概率大于 $1/2$，则可以预测事件 1 会发生；否则，预测事件 0 会发生。由于 Sigmoid 函数 $\sigma(z) = \frac{\exp(z)}{1+\exp(z)}$ 单调递增，这等价于判断 $\boldsymbol{a}^\top \boldsymbol{x} + b$ 的符号。因此，Logistic 回归可以作为分类器。

问题：为什么阈值 $1/2$ 对应的恰好是 $\boldsymbol{a}^\top \boldsymbol{x} + b = 0$?

预测规则

我们可以根据

$$p(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{a}^\top \boldsymbol{x} + b\right)}{1 + \exp\left(\boldsymbol{a}^\top \boldsymbol{x} + b\right)}$$

来预测新数据 $\boldsymbol{x}$ 取到 1 的概率。若该概率大于 $1/2$，则可以预测事件 1 会发生；否则，预测事件 0 会发生。由于 Sigmoid 函数 $\sigma(z) = \frac{\exp(z)}{1+\exp(z)}$ 单调递增，这等价于判断 $\boldsymbol{a}^\top \boldsymbol{x} + b$ 的符号。因此，Logistic 回归可以作为分类器。

问题：为什么阈值 $1/2$ 对应的恰好是 $\boldsymbol{a}^\top \boldsymbol{x} + b = 0$?

答案：因为 $\sigma(0) = 1/2$，且 Sigmoid 函数单调递增，所以 $p(\boldsymbol{x}) > 1/2$ 当且仅当 $\boldsymbol{a}^\top \boldsymbol{x} + b > 0$。

训练与测试

该问题给出了某个线性的分类器：

- 若 $\boldsymbol{a}^\top \boldsymbol{x} + b < 0$，我们认为它属于类别 0
- 若 $\boldsymbol{a}^\top \boldsymbol{x} + b > 0$，则我们认为它应该属于类别 1

如果该分类器可以完美分类，则 $\boldsymbol{a}^\top \boldsymbol{x} + b = 0$ 对应一个严格分离超平面。

在实际问题中，

- 训练集：用于求最优参数 $\boldsymbol{a}, b$
- 测试集：只用于最终评估模型性能；看完测试结果后不应再调模型

一般而言，测试集的数据量会比训练集稍微小一些。

目录

拓展：双层神经网络

以下内容作为拓展，帮助大家看到 Logistic 回归与神经网络之间的联系。

之前的 Logistic 回归预测器可以写成

$$p(\boldsymbol{x}) = \sigma(\boldsymbol{a}^\top \boldsymbol{x} + b), \qquad \sigma(z) := \frac{1}{1+e^{-z}}$$

该函数 σ 就是著名的**Sigmoid 函数**，它常被用于作为神经网络的激活函数。

如果我们使用 ℓ 个 Sigmoid 单元，并将其线性组合起来，可以得到一个更灵活的函数：

$$s(\boldsymbol{x}) = \sum_{i=1}^{\ell} w_i^{(2)} \sigma\left(\boldsymbol{a}_i^\top \boldsymbol{x} + b_i^{(1)}\right) + b^{(2)} \overset{\text{改写成}}{=} \left(\boldsymbol{W}^{(2)}\right)^\top \sigma\left(\boldsymbol{W}^{(1)} \boldsymbol{x} + \boldsymbol{b}^{(1)}\right) + b^{(2)}$$

双层神经网络的参数

写成矩阵形式后，各个参数分别为

$$\boldsymbol{W}^{(2)} = \begin{bmatrix} w_1^{(2)} \\ w_2^{(2)} \\ \vdots \\ w_\ell^{(2)} \end{bmatrix} \in \mathbb{R}^{\ell}, \qquad \boldsymbol{W}^{(1)} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \boldsymbol{a}_2^\top \\ \vdots \\ \boldsymbol{a}_\ell^\top \end{bmatrix} \in \mathbb{R}^{\ell \times n},$$

$$\boldsymbol{b}^{(1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_\ell^{(1)} \end{bmatrix} \in \mathbb{R}^{\ell}, \qquad b^{(2)} \in \mathbb{R}$$

双层神经网络示意图

输入层　　隐藏层　　输出层

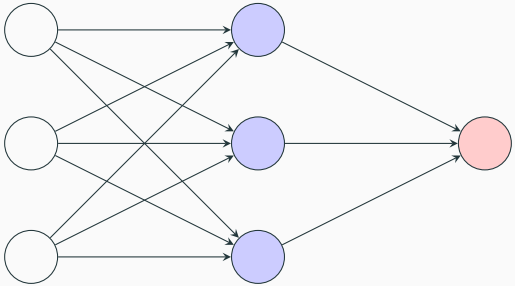


图 1: 前一页的两层神经网络，$n = \ell = 3$

深度神经网络

若进一步把这样的结构组装成复合函数，则得到了全连接的神经网络：

$$s(\boldsymbol{x}) = f_L \circ f_{L-1} \circ \cdots \circ f_2 \circ f_1(\boldsymbol{x})$$

其中

$$\begin{aligned} f_i(\boldsymbol{x}) &= \sigma\left(\boldsymbol{W}^{(i)}\boldsymbol{x} + \boldsymbol{b}^{(i)}\right) \qquad \text{if } i \leq L-1 \\ f_L(\boldsymbol{x}) &= \boldsymbol{W}^{(L)}\boldsymbol{x} + \boldsymbol{b}^{(L)} \end{aligned}$$

优化变量为 $\boldsymbol{W}^{(i)}$ 和 $\boldsymbol{b}^{(i)}$。

深度神经网络的特点

在二分类任务中，常把最终 $s(\boldsymbol{x})$ 再送入一个 Sigmoid 函数，得到预测概率。

- 该复合函数一般而言不再是凸函数，但具有更好的非线性拟合能力；
- 对于该非凸优化的算法常是基于梯度法等算法改造，并会在凸优化问题中先行验证有效性。

本节课测试

问题：请快速回答下面问题：

- 为什么在线性不可分时要引入松弛变量?
- 为什么 $p(\boldsymbol{x}) > 1/2$ 等价于 $\boldsymbol{a}^\top \boldsymbol{x} + b > 0$?
- 训练集和测试集的作用有什么区别?

本节课测试

问题：请快速回答下面问题：

- 为什么在线性不可分时要引入松弛变量?
- 为什么 $p(\boldsymbol{x}) > 1/2$ 等价于 $\boldsymbol{a}^\top \boldsymbol{x} + b > 0$?
- 训练集和测试集的作用有什么区别?

答案：松弛变量允许违反间隔约束；Sigmoid 单调且 $\sigma(0) = 1/2$；训练集求参数，测试集做最终评估。

目录

总结

- 线性分类器：看 $s = \boldsymbol{a}^\top \boldsymbol{x} + b$ 的符号，分界面是超平面
- 软间隔 SVM：用 $\frac{1}{2}\|\boldsymbol{a}\|_2^2 + \gamma \sum_i u_i$ 平衡间隔与容错
- Logistic 回归：用 $\sigma(s)$ 建模概率，最小化负对数似然
- 预测规则：阈值 $1/2$ 等价于判断 s 的符号
- 神经网络：作为拓展，可看作非线性单元的复合，通常对应非凸优化

阅读作业 & 参考资料：

- 课本第 7.1，8.6 章
- 其他部分可选择性浏览