lec10.1 Probabilistic Classifiers

Probabilistic Classifiers

介绍

普通分类器与概率分类器（Ordinary vs. Probabilistic）

Ordinary Classifier:

$$c = f(X)$$

其中，$c \in \{c_1, \ldots, c_k\}$

Probabilistic Classifier:

$$p = P\left(c_i | X\right)$$

其中，$P\left(c_i | X\right)$ 是条件概率，$p \in \{p_1, \ldots, p_k\}$，且 $\sum_{i=1}^{k} p_i = 1$

判别模型与生成模型（Discriminative vs. Generative）

Discriminative:
假设条件分布 $P(C|X)$ 具有某些参数 $\theta = (\theta_1, \ldots, \theta_k)$ 的特定形式 $P_\theta(C|X)$。使用训练集找到这些参数 $(\theta_1, \ldots, \theta_k)$，使得 P_θ 最佳

Generative:
假设数据来自某些参数 $\theta = (\theta_1, \ldots, \theta_k)$ 的特定分布 $P_\theta(X, C)$。使用训练集找到这些参数 $(\theta_1, \ldots, \theta_k)$，使得 P_θ 最佳。最后使用 P_θ 来分类新实例

值得一提的是，当我们说条件分布具有某些参数的特定形式时（P_θ），我们指的是条件分布可以用数学公式表达，并且这个公式包括了一些参数，这些参数可以调整分布的形状和特性

生成模型

数据生成分布

假设:
数据来自于某个未知的分布 P，涵盖对象-类别对 $(X, c) \in \mathscr{X} \times \mathscr{C}$。换句话说，这个分类问题由 P 描述

已知 P 的情况:
对于任何 $(X, c) \in \mathscr{X} \times \mathscr{C}$，我们可以计算 $P(X, c)$：具有特征向量 X 的对象属于 c 的概率

然后我们就可以计算贝叶斯最优分类器

Bayes Optimal Classifier

$$f^*(X) = \underset{c \in \mathscr{C}}{\operatorname{argmax}}\, P(X, c)$$

贝叶斯最优分类器是一种理论上的最佳的可能分类器，它通过利用先验知识和证据来做出最有可能的类别

判断。这意味着如果 g 是另一个分类器，那么对于任何 $X \in \mathscr{X}$，f^* 在 X 上出错的概率小于 g 在 X 上出错的概率

贝叶斯最优分类器的工作基于：

- 概率模型：它假设所有关于类别和特征的概率分布都是已知的。这包括
  - 先验概率（各类别的概率） $P(C)$
  - 条件概率（给定类别时特征的概率） $P(X|C)$
- 贝叶斯定理：它计算后验概率 $P(C|X)$，即在给定特征 X 的情况下每个类别 C 的概率

  $$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

  其中，$P(X)$ 是归一化常数，可以通过对所有可能的 C 求和来计算
- 决策规则：分类器选择具有最高后验概率 $P(C|X)$ 的类别 C

  $$f^*(X) = \underset{c \in \mathscr{C}}{\operatorname{argmax}}\, P(C = c\,|X)$$

  它选择使得给定观测数据 X 属于类别 C 的概率最大的类别

实践

在实际中，显然我们不知道数据真正的生成分布 P。我们需要使用训练集学习到一个分布 $\hat{P}$，它最好与 P 非常相似，这之后我们使用 $\hat{P}$ 做分类

构建 $\hat{P}$ 的一种方法：

1. 假设：假设分布 P 属于某个参数化分布族
2. 参数估计：估计这个参数化分布族中的参数，这些参数定义了训练数据最可能的数据生成分布
3. 关键假设：训练数据是从 P 中独立同分布抽取的，即训练实例 (X, c) 是独立同分布的

参数估计

参数估计的方法很多，她只介绍了最大似然估计（Maximum Likelihood Estimation，MLE）

MLE 单参数

Example 1

假设模型：
假设有一枚硬币偏向于某一面，想要模拟这种偏置（也可以看作是二元分类问题中的一类）。观测到的数据是 THHH，其中 H 表示正面，T 表示反面

模型假设：

- 所有抛硬币实验都使用同一枚硬币，并且每次抛硬币都是独立的（独立同分布 $i.i.d.$ 假设）
- 硬币出现正面的固定概率为 β，反面为 $1-\beta$。这里假设数据生成分布属于伯努利分布族，参数化为正面概率 $\beta \in [0, 1]$

最大似然估计：
给定观测数据 THHH，计算在 β 条件下观测到这些数据的概率

$$P_\beta(\mathrm{THHH}) = P_\beta(\mathrm{T}) \cdot P_\beta(\mathrm{H}) \cdot P_\beta(\mathrm{H}) \cdot P_\beta(\mathrm{H}) = (1-\beta)\beta\beta\beta = \beta^3 - \beta^4$$

求解使概率最大的 β：

对于 $P_\beta(\mathrm{THHH}) = \beta^3 - \beta^4$，进行如下操作

$$\frac{\partial}{\partial \beta}(\beta^3 - \beta^4) = 3\beta^2 - 4\beta^3$$
$$\Downarrow$$
$$3\beta^2 - 4\beta^3 = 0 \Rightarrow 3\beta^2 = 4\beta^3$$
$$\Rightarrow \beta = \frac{3}{4}$$

Example 2

在 example 1 的基础上，我们得到了 h 次正面与 t 次反面

$$\beta^h(1-\beta)^t$$

为了简化计算，我们进行对数似然

$$\log\left(\beta^h(1-\beta)^t\right) = h\log\beta + t\log(1-\beta)$$

我们需要计算 β 的最大似然估计值

$$l(\beta) = h\log\beta + t\log(1-\beta)$$
$$\frac{dl(\beta)}{d\beta} = \frac{h}{\beta} - \frac{t}{1-\beta}$$
$$\Downarrow$$
$$\text{let } \frac{h}{\beta} - \frac{t}{1-\beta} = 0$$
$$\Downarrow$$
$$\beta = \frac{h}{h+t}$$

$\frac{h}{h+t}$ 就是 β 的最大似然估计

MLE 多参数

Example 3

1. **假设模型**：我们要建模的是一个有 K 个面的骰子（或者在多分类问题中的一个类）。可以用参数 $\beta_1, \beta_2, \ldots, \beta_K$ 来建模，其中 β_i 是骰子出现第 i 面的概率。
   由于 β 是概率，我们还应假设：
   - 对于每个 $i = 1, \ldots, K$，$\beta_i \geq 0$
   - $\beta_1 + \beta_2 + \ldots + \beta_K = 1$
2. **概率计算**：x_1 表 示骰子投掷结果为 1 的次数，依此类推。则观测数据的概率是：

$$\beta_1^{x_1} \cdot \beta_2^{x_2} \cdot \ldots \cdot \beta_K^{x_K}$$

3. **对数似然**：对数概率（即对数似然）为：

$$\sum_{i=1}^{K} x_i \log \beta_i$$

4. **求解最大似然估计**：
   为了找到使对数似然最大的 β，我们需要找到最大化对数似然并且满足约束 $\beta_1 + \beta_2 + \ldots + \beta_K = 1$ 的 β。

使用拉格朗日乘数法（参考 lec02 最后部分），我们得到如下拉格朗日函数：

$$\mathscr{L}(\beta, \lambda) = \sum_{i=1}^{K} x_i \log \beta_i - \lambda \left(\sum_{i=1}^{K} \beta_i - 1 \right)$$

5. 计算导数并设为0:
   对于固定的 i，我们有：

$$\frac{\partial \mathscr{L}(\beta, \lambda)}{\partial \beta_i} = \frac{x_i}{\beta_i} - \lambda$$

   将导数设为0并解 β_i：

$$\frac{x_i}{\beta_i} - \lambda = 0 \implies \beta_i = \frac{x_i}{\lambda}$$

6. 求解拉格朗日乘数 λ：由约束条件 $\beta_1 + \beta_2 + \ldots + \beta_K = 1$（这相当于 $\frac{\partial \mathscr{L}(\beta,\lambda)}{\partial \lambda} = 0$），我们得到：

$$\frac{x_1 + x_2 + \ldots + x_K}{\lambda} = 1 \implies \lambda = \sum_{i=1}^{K} x_i$$

7. 最终解：因此，β_i 的最大似然估计是：

$$\beta_i = \frac{x_i}{\sum_{i=1}^{K} x_i}$$

   其中，$i = 1, \ldots, K$