机器学习（本科生公选课）GEC6531

第7节 朴素贝叶斯 Naïve Bayes

计算机科学与技术学院

张瑞 教授

邮箱：ruizhang6@hust.edu.cn

签到 & 思考

■ 微助教签到（学校要求）

1. 加入课堂：微信扫码或者通过微助教公众号

课堂名称：GEC6531 机器学习（公选课）

课堂编号：OA628

1、扫码关注公众号：微助教服务号。

2、点击系统通知：“点击此处加入【GEC6531 机器学习（公选课）】课堂”，填写学生资料加入课堂。

*如未成功收到系统通知，请点击公众号下方“学生” - “全部(A)” - “加入课堂” --- “输入课堂编号”手动加入课堂

二维码有效期至：2024-11-16

2. 微信扫码签到

朴素贝叶斯例子视频

今天的目录

- **基本思想**
  - 简介
- **朴素贝叶斯**
  - 贝叶斯法则
  - 朴素贝叶斯假设
- **估计$P([x]_\alpha \mid y)$**
  - 情形1: 分类特征
  - 情形2: 多项特征
  - 情形3: 连续特征（高斯朴素贝叶斯）
- **朴素贝叶斯分类器**
  - 朴素贝叶斯分类器
  - 高斯朴素贝叶斯
- **实例与应用**
  - 利用朴素贝叶斯过滤垃圾邮件

今天的目录

- **基本思想**
  - 简介
- **朴素贝叶斯**
  - 贝叶斯法则
  - 朴素贝叶斯假设
- **估计$P([x]_\alpha \mid y)$**
  - 情形1: 分类特征
  - 情形2: 多项特征
  - 情形3: 连续特征（高斯朴素贝叶斯）
- **朴素贝叶斯分类器**
  - 朴素贝叶斯分类器
  - 高斯朴素贝叶斯
- **实例与应用**
  - 利用朴素贝叶斯过滤垃圾邮件

简介

基本思想

在机器学习中，朴素贝叶斯分类器是基于贝叶斯定理的、在强独立假设下的一类简单概率分类器。

- 训练数据：$D = \{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)\}, (\mathrm{x}_i, y_i)$ 从未知的 $P(X, Y)$ 分布中 i.i.d. 采样得到。
  由此可得：
$$P(D) = P((\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)) = \prod_{\alpha=1}^{n} P(\mathrm{x}_\alpha, y_\alpha).$$
- 若有足够的数据，可以估计 P(X,Y)，类似于上一讲的硬币例子，想象有一个巨大的骰子，每一面对应 (x, y) 的一种可能取值。通过计数的方式来估计某一面出现的概率。
- 估计 $P(X, Y)$:
$$\hat{P}(\mathrm{x}, y) = \frac{\sum_{i=1}^{n} I(\mathrm{x}_i = x \wedge y_i = y)}{n}.$$
$$I(\mathrm{x}_i = x \wedge y_i = y) = 1 \quad if \quad \mathrm{x}_i = x \qquad y_i = y,$$

  否则为 0。

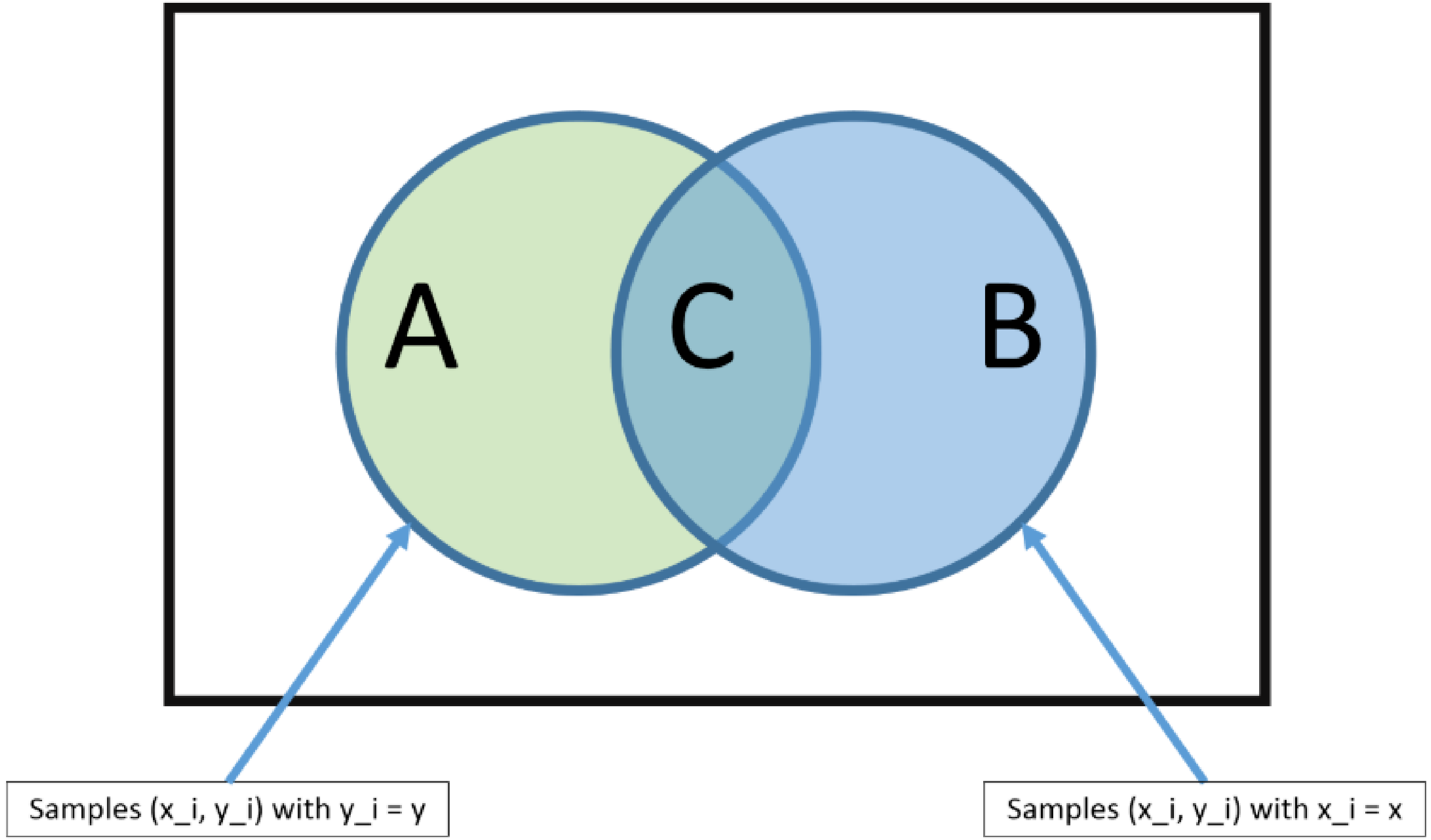
基本思想

- 如果关注的是从特征 x 中预测标签 y，则可以直接估计 $P(y|x)$，而不是 $P(x,y)$。
- 然后，可以使用贝叶斯最优分类器对特定的 $\hat{P}(y|x)$ 进行预测。

$$\hat{P}(y|\mathrm{x}) = \frac{\hat{P}(y,\mathrm{x})}{P(\mathrm{x})} = \frac{[\sum_{i=1}^{n} I(\mathrm{x}_i = \mathrm{x} \wedge y_i = y)]/n}{[\sum_{i=1}^{n} I(\mathrm{x}_i = \mathrm{x})]/n}.$$

$$= \frac{\sum_{i=1}^{n} I(\mathrm{x}_i = \mathrm{x} \wedge y_i = y)}{\sum_{i=1}^{n} I(\mathrm{x}_i = \mathrm{x})}$$

可视化



韦恩 (Venn) 图

用韦恩图说明 MLE 估计:

$$\hat{P}(y|\mathbf{x}) = \frac{|C|}{|B|}.$$

今天的目录

- **基本思想**
  - 简介
- **朴素贝叶斯**
  - 贝叶斯法则
  - 朴素贝叶斯假设
- **估计$P([x]_\alpha \mid y)$**
  - 情形1: 分类特征
  - 情形2: 多项特征
  - 情形3: 连续特征（高斯朴素贝叶斯）
- **朴素贝叶斯分类器**
  - 朴素贝叶斯分类器
  - 高斯朴素贝叶斯
- **实例与应用**
  - 利用朴素贝叶斯过滤垃圾邮件

贝叶斯法则

如果可以估计 $P(y)$ 和 $P(\mathrm{x} \mid y)$，则根据贝叶斯法则，$P(y|\mathrm{x}) = \frac{P(\mathrm{x}|y)P(y)}{P(\mathrm{x})}$.

估计 $P(y)$, $P(x|y)$

- 估计 $P(y)$ 很容易：例如，如果 y 采用离散的二进制值，估计 $P(y)$ 就会变成“抛硬币”。只需要计算观察到的每个结果的次数 (在本例中是每个类)：

$$P(y=c) = \frac{\sum_{i=1}^{n} I(y_i = c)}{n} = \hat{\pi}_c$$

- 然而对于一个特定的高维 x，估计 $P(x|y)$ 是**不容易**的。例如：垃圾邮件分类器。
- 我们做的另一个假设是：朴素贝叶斯假设。

朴素贝叶斯假设

朴素贝叶斯假设:

$$P(\mathbf{x}|y) = \prod_{\alpha=1}^{d} P(x_\alpha|y)，其中 x_\alpha = [\mathbf{x}]_\alpha 是第 \alpha 个特征的值。$$

即当标签给定时，特征值是相互**独立**的（每一维的特征值均条件独立于给定的 y）。
这是一个非常**大胆**的假设。

例如，垃圾邮件过滤经常使用朴素贝叶斯分类器的设定。输入数据对应电子邮件，标签是垃圾邮件或非垃圾邮件。朴素贝叶斯假设意味着电子邮件中的单词是条件独立的，前提是你知道一封电子邮件是不是垃圾邮件。显然这是不现实的。
垃圾邮件和非垃圾邮件的字词都不是独立随机采样的。然而，即便这违背了事实，基于条件独立假设所得到的分类器在实践中仍可以很好地工作。

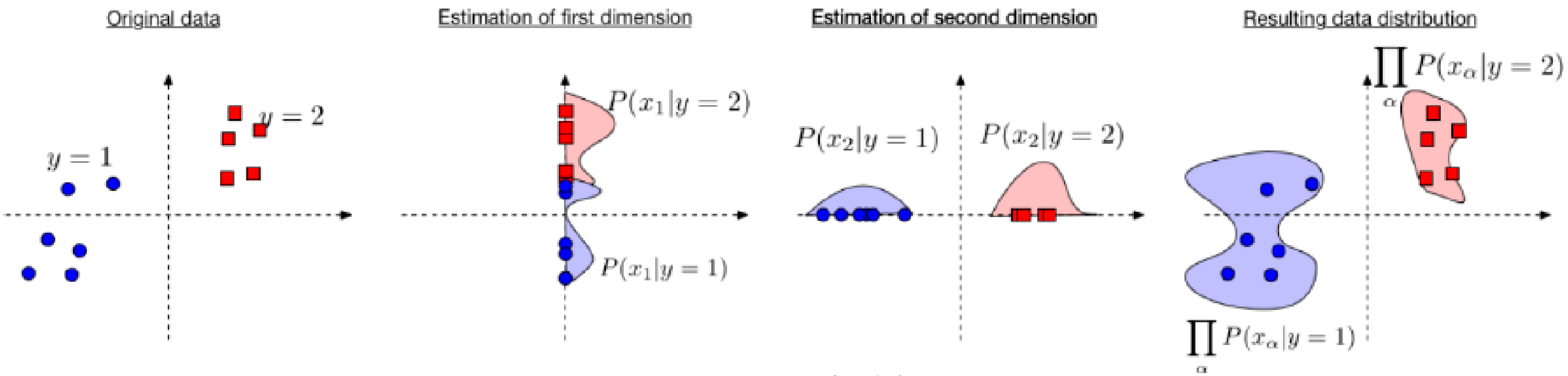


Illustration behind the Naive Bayes algorithm. We estimate $P(x_\alpha|y)$ independently in each dimension (middle two images) and then obtain an estimate of the full data distribution by assuming conditional independence $P(\mathbf{x}|y) = \prod_\alpha P(x_\alpha|y)$ (very right image).

估计P(x | y)

假设朴素贝叶斯假设成立，则贝叶斯分类器的定义如下：

贝叶斯分类器

根据朴素贝叶斯假设

$$
\begin{aligned}
h(\mathrm{x}) &= \operatorname*{argmax}_y P(y|\mathrm{x}) && (1)\\
&= \operatorname*{argmax}_y \frac{P(\mathrm{x}|y)P(y)}{P(\mathrm{x})} && (2)\\
&= \operatorname*{argmax}_y P(\mathrm{x}|y)P(y) \quad (P(\mathrm{x}) \text{ 与 } y \text{ 无关}) && (3)\\
&= \operatorname*{argmax}_y \prod_{\alpha=1}^{d} P(x_\alpha|y)P(y) \quad (\text{根据朴素贝叶斯假设}) && (4)\\
&= \operatorname*{argmax}_y \sum_{\alpha=1}^{d} \log(P(x_\alpha|y)) + \log(P(y)) \quad (\text{因为 log 是一个单调函数}) && (5)
\end{aligned}
$$

估计 $log(P(x_\alpha|y))$ 很容易，因为我们只需要考虑一维。而且估计 P(y) 不受假设的影响。

今天的目录

- **基本思想**
  - 简介
- **朴素贝叶斯**
  - 贝叶斯法则
  - 朴素贝叶斯假设
- **估计$P([x]_\alpha \mid y)$**
  - 情形1: 分类特征
  - 情形2: 多项特征
  - 情形3: 连续特征（高斯朴素贝叶斯）
- **朴素贝叶斯分类器**
  - 朴素贝叶斯分类器
  - 高斯朴素贝叶斯
- **实例与应用**
  - 利用朴素贝叶斯过滤垃圾邮件

情形1: 分类特征

特征:

$$[\mathrm{x}]_\alpha \in \{f_1, f_2, \cdots, f_{K_\alpha}\}.$$

每维特征 α 都属于 K_α 种类别之一。(二元特征的情况只是其特定情况，$K_\alpha = 2$)
例如，对医院的门诊数据，其中：
特征包括：性别 (男性/女性)、年龄（1 到 120)、血压、是否咳嗽（是/否)、体温（36, 36.5, ..., 40)、是否咽痛（不痛、有点痛、很痛)，
标签 y 为没有感冒、普通感冒、流感。

Illustration of categorical NB. For d dimensional data, there exist d independent dice for each class. Each feature has one die per class. We assume training samples were generated by rolling one die after another. The value in dimension i corresponds to the outcome that was rolled with the i^{th} die.

分类特征

建模 $P(x_\alpha \mid y)$: $[\theta_{jc}]_\alpha$ 表示标签为 c 时，特征 α 值为 j 的概率。
如 $c =$ 感冒的同学，其咳嗽特征的取值为 “咳嗽”。
约束：x_α 的取值必须为类别 $\{1 \ldots, K_\alpha\}$ 之一。
参数估计:

$$[\hat{\theta}_{jc}]_\alpha = \frac{\sum_{i=1}^{n} I(y_i = c) I(x_{i\alpha} = j) + \ell}{\sum_{i=1}^{n} I(y_i = c) + \ell K_\alpha}, \tag{6}$$

$$x_{i\alpha} = [\mathbf{x}_i]_\alpha$$

其中 ℓ 是一个平滑参数。
1） 若设 $\ell = 0$，则得到一个 MLE 估计器。
2） 若 $\ell \geq 0$，即 MAP。若设 $\ell = 1$，即拉普拉斯平滑。

简言之，即:

$$\frac{\text{标签为 } c \text{ 的样本中，第 } \alpha \text{ 维特征的取值为 } j \text{ 的数量}}{\text{标签为 } c \text{ 的样本数}}$$

预测

实际上，分类特征模型将一个特殊的“骰子”与每个特征和标签相关联。
我们假设的生成模型首先通过选择标签来生成数据（如“感冒”）。该标签带有一组“骰子”(每一维特征对应一个“骰子”)。
生成器选择每个骰子，投掷它并用投掷的结果填充特征值。因此，如果有 C 个可能的标签和 d 个维度，我们从数据中估计 dC 个“骰子”每一面的概率。
每个数据点只投掷 d 个骰子 (每个特征对应一个)。(对于某个标签) 第 α 个骰子有 K_α 个“面”。
当然，数据在现实中并不是这样生成的——但这是我们所做的模型假设。
然后，我们从数据中学习这些模型。
在测试期间，看看哪个模型更有可能输出某测试样本 x。

预测

$$h(\mathrm{x}) = \operatorname*{argmax}_y \prod_{\alpha=1}^{d} P(x_\alpha|y)P(y)$$

$$\operatorname*{argmax}_y \; P(y=c \mid \mathrm{x}) \propto \operatorname*{argmax}_y \; \hat{\pi}_c \prod_{\alpha=1}^{d} [\hat{\theta}_{jc}]_\alpha$$

$$\hat{\pi}_c = P(y=c) = \frac{\sum_{i=1}^{n} I(y_i=c)}{n}$$

情形2: 多项特征

多项特征

如果特征值不是枚举型类别 (例如性别)，但可以计算，则需要使用不同的模型。例如，在文本文档分类中，特征值 $x_\alpha = j$ 意味着在这个特定的文档 x 中，第 α^{th} 个单词出现了 j 次。

以垃圾邮件过滤为例。假设第 α^{th} 个单词如（捐赠 donation 或钱 money）的值与是否是垃圾邮件有关。如果 $x_\alpha = 10$ 意味着这封电子邮件很可能是垃圾邮件 (因为第 α 个单词在其中出现了 10 次)。则另一封带有 $x'_\alpha = 20$ 的电子邮件则更有可能是垃圾邮件 (因为这个词出现的频率是两倍)。

特征:

$$x_\alpha \in \{0, 1, 2, \ldots, m\} \text{ 且} m = \sum_{\alpha=1}^{d} x_\alpha \tag{7}$$

每个 α 维特征表示一个计数，m 是序列的长度。例如，长度为 m 的电子邮件中特定单词（捐赠,donation）的计数，d 是词汇表的大小。

估计P(x | y)

使用多项分布:

$$P(\mathrm{x} \mid m, y=c) = \frac{m!}{x_1! \cdot x_2! \cdot \cdots \cdot x_d!} \prod_{\alpha=1}^{d} (\theta_{\alpha c})^{x_\alpha}$$

其中 $\theta_{\alpha c}$ 是选择第 α 个单词的概率, $\sum_{\alpha=1}^{d} \theta_{\alpha c} = 1$。
因此，可以使用它来生成一个垃圾邮件，即类 $y = \mathrm{spam}$ 的一个电子邮件 x，通过 $P(\mathrm{x} \mid y = \mathrm{spam})$ 来生成.

参数估计:

$$\hat{\theta}_{\alpha c} = \frac{\sum_{i=1}^{n} I(y_i = c) x_{i\alpha} + \ell}{\sum_{i=1}^{n} I(y_i = c) m_i + \ell \cdot d} \tag{8}$$

其中 $m_i = \sum_{\beta=1}^{d} x_{i\beta}$ 表示文档 i 中的总词数。
分子对特征 x_α（如 donation）的所有计数求和，分母对所有数据的所有特征的计数求和。
即:

$$\frac{\text{所有垃圾邮件中第 } \alpha \text{ 个单词出现的总次数}}{\text{所有垃圾邮件中所有单词出现的总次数}}$$

预测:

$$\operatorname*{argmax}_{c} P(y = c \mid \mathrm{x}) \propto \operatorname*{argmax}_{c} \hat{\pi}_c \prod_{\alpha=1}^{d} \hat{\theta}_{\alpha c}^{x_\alpha}$$

情形3: 连续的特征（高斯朴素贝叶斯）

特征:（如身高）

$$x_\alpha \in \mathbb{R} \quad \text{(每个特征对应一个实数值)} \tag{9}$$

模型 $P(x_\alpha \mid y)$ 使用高斯分布:（如 $y =$ 男生，其身高的分布）

$$P(x_\alpha \mid y = c) = \mathcal{N}\left(\mu_{\alpha c}, \sigma_{\alpha c}^2\right) = \frac{1}{\sqrt{2\pi}\sigma_{\alpha c}} e^{-\frac{1}{2}\left(\frac{x_\alpha - \mu_{\alpha c}}{\sigma_{\alpha c}}\right)^2} \tag{10}$$

参数估计:

$$\mu_{\alpha c} = ?$$

$$\sigma_{\alpha c} = ?$$

情形3: 连续的特征（高斯朴素贝叶斯）

特征:（如身高）

$$x_\alpha \in \mathbb{R} \text{（每个特征对应一个实数值）} \tag{11}$$

模型 $P(x_\alpha \mid y)$ 使用高斯分布:（如 $y=$ 男生，其身高的分布）

$$P(x_\alpha \mid y = c) = \mathcal{N}\left(\mu_{\alpha c}, \sigma^2_{\alpha c}\right) = \frac{1}{\sqrt{2\pi}\sigma_{\alpha c}} e^{-\frac{1}{2}\left(\frac{x_\alpha - \mu_{\alpha c}}{\sigma_{\alpha c}}\right)^2} \tag{12}$$

参数估计:

$$\mu_{\alpha c} = \frac{1}{n_c}\sum_{i=1}^{n} I(y_i = c)x_{i\alpha}; \quad n_c = \sum_{i=1}^{n} I(y_i = c)$$

$$\sigma^2_{\alpha c} = \frac{1}{n_c}\sum_{i=1}^{n} I(y_i = c)(x_{i\alpha} - \mu_{\alpha c})^2;$$

情形3: 连续的特征（高斯朴素贝叶斯）

特征:（如身高）

$$x_\alpha \in \mathbb{R} \quad \text{(每个特征对应一个实数值)} \tag{13}$$

模型 $P(x_\alpha \mid y)$ 使用高斯分布:（如 $y=$ 男性，其身高的分布）

$$P(x_\alpha \mid y=c) = \mathcal{N}\left(\mu_{\alpha c}, \sigma_{\alpha c}^2\right) = \frac{1}{\sqrt{2\pi}\sigma_{\alpha c}} e^{-\frac{1}{2}\left(\frac{x_\alpha - \mu_{\alpha c}}{\sigma_{\alpha c}}\right)^2} \tag{14}$$

注意，上面指定的模型是基于我们对数据的假设——每个特征 α 来自基于类条件的高斯分布。

总的分布:

$$P(\mathbf{x}|y) \sim \mathcal{N}(\mu_y, \Sigma_y)$$

其中 Σ_y 是一个协方差对角矩阵

$$[\Sigma_y]_{\alpha,\alpha} = \sigma_{\alpha,y}^2$$

参数估计

参数估计:

我们独立地估计每个维度和类的分布参数。
高斯分布只有两个参数，均值和方差。
均值 $\mu_{\alpha,y}$ 是由标签为 y 的所有样本的维数 α 的平均特征值来估计的。
(平方) 标准差就是这个估计值的方差。

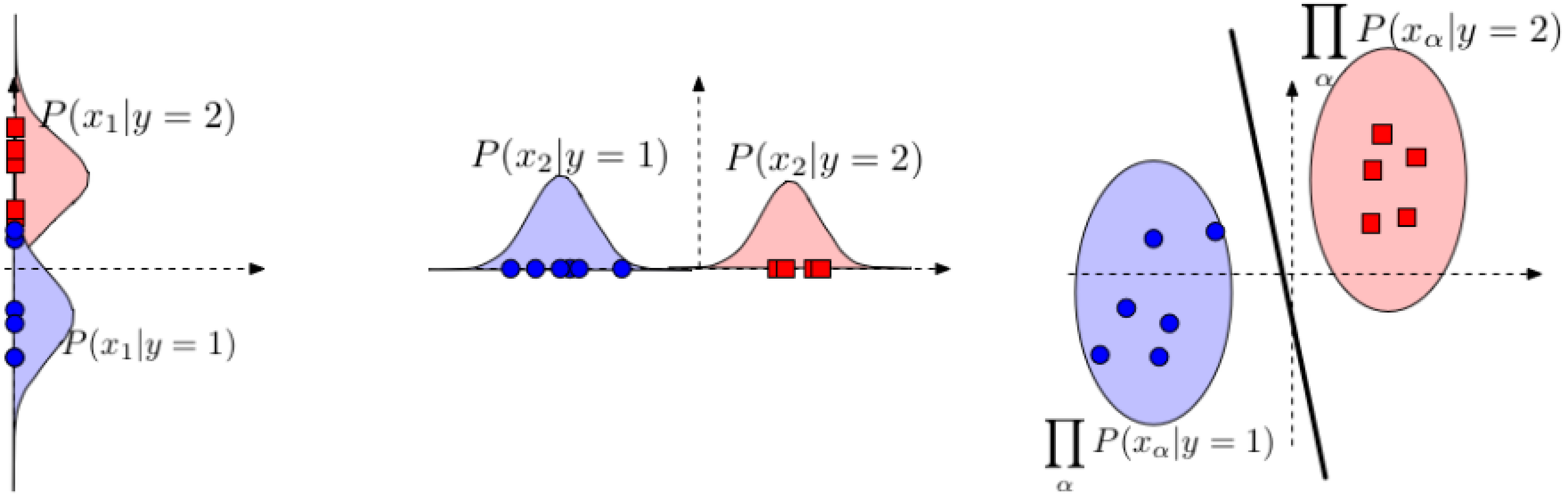
$$\mu_{\alpha c} \leftarrow \frac{1}{n_c}\sum_{i=1}^{n} I(y_i = c)x_{i\alpha} \qquad \text{其中 } n_c = \sum_{i=1}^{n} I(y_i = c) \tag{15}$$

$$\sigma^2_{\alpha c} \leftarrow \frac{1}{n_c}\sum_{i=1}^{n} I(y_i = c)(x_{i\alpha} - \mu_{\alpha c})^2 \tag{16}$$

今天的目录

- **基本思想**
  - 简介
- **朴素贝叶斯**
  - 贝叶斯法则
  - 朴素贝叶斯假设
- **估计$P([x]_\alpha \mid y)$**
  - 情形1: 分类特征
  - 情形2: 多项特征
  - 情形3: 连续特征（高斯朴素贝叶斯）
- **朴素贝叶斯分类器**
  - 朴素贝叶斯分类器
  - 高斯朴素贝叶斯
- **实例与应用**
  - 利用朴素贝叶斯过滤垃圾邮件

朴素贝叶斯是一个线性分类器



Naive Bayes leads to a linear decision boundary in many common cases. Illustrated here is the case where $P(x_\alpha|y)$ is Gaussian and where $\sigma_{\alpha,c}$ is identical for all c (but can differ across dimensions α). The boundary of the ellipsoids indicate regions of equal probabilities $P(\mathbf{x}|y)$. The red decision line indicates the decision boundary where $P(y=1|\mathbf{x}) = P(y=2|\mathbf{x})$.

1. 多项特征

假设 $y_i \in \{-1, +1\}$，特征满足多项分布。则可以得出:

$$h(\mathbf{x}) = \underset{y}{\operatorname{argmax}} \; P(y) \prod_{\alpha-1}^{d} P(x_\alpha \mid y) = \operatorname{sign}(\mathbf{w}^\top \mathbf{x} + b)$$

$$\mathbf{w}^\top \mathbf{x} + b > 0 \Longleftrightarrow h(\mathbf{x}) = +1.$$

如前所述，定义:

$$P(x_\alpha | y = +1) \propto \theta_{\alpha+}^{x_\alpha}; P(Y = +1) = \pi_+.$$

$$[\mathbf{w}]_\alpha = \log(\theta_{\alpha+}) - \log(\theta_{\alpha-}) \quad (17)$$

$$b = \log(\pi_+) - \log(\pi_-) \quad (18)$$

如果使用上面的方法进行分类，可以计算 $\mathbf{w}^\top \mathbf{x} + b$

$$\mathbf{w}^\top \mathbf{x} + b > 0 \iff \sum_{\alpha=1}^{d} [\mathbf{x}]_\alpha \overbrace{(\log(\theta_{\alpha+}) - \log(\theta_{\alpha-}))}^{[\mathbf{w}]_\alpha} + \overbrace{\log(\pi_+) - \log(\pi_-)}^{b} > 0 \tag{19}$$

$$\iff \exp\left(\sum_{\alpha=1}^{d} [\mathbf{x}]_\alpha (\log(\theta_{\alpha+}) - \log(\theta_{\alpha-})) + \log(\pi_+) - \log(\pi_-)\right) > 1 \tag{20}$$

$$\iff \prod_{\alpha=1}^{d} \frac{\exp\left(\log \theta_{\alpha+}^{[\mathbf{x}]_\alpha} + \log(\pi_+)\right)}{\exp\left(\log \theta_{\alpha-}^{[\mathbf{x}]_\alpha} + \log(\pi_-)\right)} > 1 \tag{21}$$

$$\iff \prod_{\alpha=1}^{d} \frac{\theta_{\alpha+}^{[\mathbf{x}]_\alpha} \pi_+}{\theta_{\alpha-}^{[\mathbf{x}]_\alpha} \pi_-} > 1 \tag{22}$$

$$\iff \frac{\prod_{\alpha=1}^{d} P([\mathbf{x}]_\alpha | Y = +1)\pi_+}{\prod_{\alpha=1}^{d} P([\mathbf{x}]_\alpha | Y = -1)\pi_-} > 1 \tag{23}$$

$$\iff \frac{P(\mathbf{x}|Y = +1)\pi_+}{P(\mathbf{x}|Y = -1)\pi_-} > 1 \tag{24}$$

$$\iff \frac{P(Y = +1|\mathbf{x})}{P(Y = -1|\mathbf{x})} > 1 \tag{25}$$

$$\iff P(Y = +1|\mathbf{x}) > P(Y = -1|\mathbf{x}) \tag{26}$$

$$\iff \operatorname*{argmax}_y P(Y = y|\mathbf{x}) = +1 \tag{27}$$

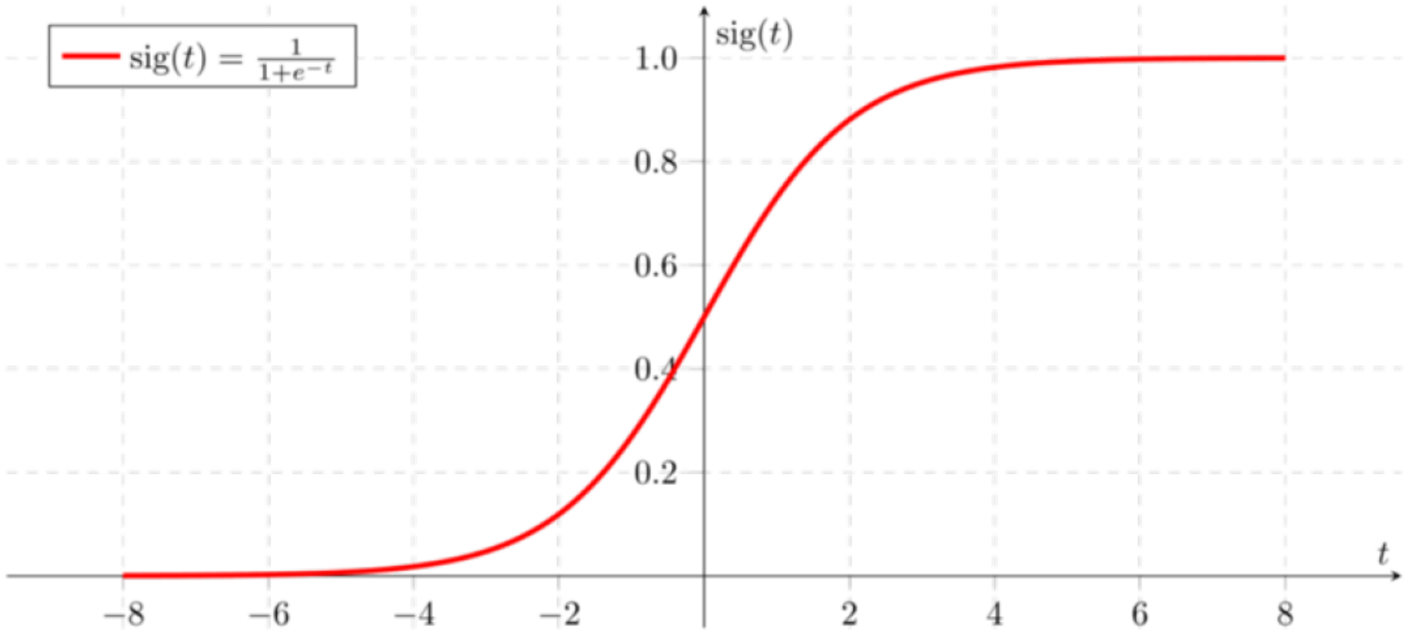
2. 连续特征

高斯朴素贝叶斯

在连续特征（高斯朴素贝叶斯）的情况下，我们可以得出：

$$P(y \mid \mathrm{x}) = \frac{1}{1 + e^{-y(\mathrm{w}^\top \mathrm{x} + b)}}$$

这个模型也被称为逻辑回归。



图：如果 t→∞，y(预测) 将变为 1，如果 t→-∞，y(预测) 将变为 0。

高斯朴素贝叶斯

$P(y|x)=\frac{1}{1+e^{-y(w^Tx+b)}}$ 恰是LR的MLE分布假设 $\underset{y}{argmin}\, p(y|x), y=\{0,1\}$

$$\theta = P(y=1|x) = \frac{P(y=1)p(x|y=1)}{p(y=1)p(x|y=1)+p(y=0)p(x|y=0)}$$

$$=\frac{1}{1+\exp(ln\frac{p(y=0)p(x|y=0)}{p(y=1)p(x|y=1)})} = \frac{1}{1+\exp(ln\frac{p(y=0)}{p(y=1)}+\sum_{i=1}^{n} ln\frac{p(x_i|y=0)}{p(x_i|y=1)})}$$

若属于高斯分布，则

$$\sum_{i=1}^{n} ln\frac{p(x_i|y=0)}{p(x_i|y=1)} = \sum_{i=1}^{n} ln\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(\frac{-(x_i-\mu_{i0})^2}{2\sigma_i^2})}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(\frac{-(x_i-\mu_{i1})^2}{2\sigma_i^2})} = -\sum_{i=1}^{n}(\frac{(x_i-\mu_{i0})^2}{2\sigma_i^2} - \frac{(x_i-\mu_{i1})^2}{2\sigma_i^2})$$

$$=-\sum_{i=1}^{n}(\frac{1}{2\sigma_i^2}(2(\mu_{i0}-\mu_{i1})x_i + (\mu_{i0}^2-\mu_{i1}^2))=-\sum_{i=1}^{n} w_i x_i + w_0$$

设p(y=1) =π，$p(y=0)=1-\pi$

$$\theta = P(y=1|x)=\frac{1}{1+\exp(ln\frac{\pi}{1-\pi}-(w^Tx+w_0))} = \frac{1}{1+\exp(-y(w^Tx+b))}$$

结果证明Naive Bayes在高斯分布时本质上是Logistic Regression

今天的目录

- **基本思想**
  - 简介
- **朴素贝叶斯**
  - 贝叶斯法则
  - 朴素贝叶斯假设
- **估计$P([x]_\alpha \mid y)$**
  - 情形1: 分类特征
  - 情形2: 多项特征
  - 情形3: 连续特征（高斯朴素贝叶斯）
- **朴素贝叶斯分类器**
  - 朴素贝叶斯分类器
  - 高斯朴素贝叶斯
- **实例与应用**
  - 利用朴素贝叶斯过滤垃圾邮件

实例与应用：利用朴素贝叶斯过滤垃圾邮件

核心算法: 朴素贝叶斯分类器训练函数

```
def trainNB0(trainMatrix, trainCategory):
    numTrainDocs = len(trainMatrix)
    numWords = len(trainMatrix[0])
    pAbusive = sum(trainCategory)/float(numTrainDocs)
    p0Num = ones(numWords); p1Num = ones(numWords)
    p0Denom = 2.0; p1Denom = 2.0
    for i in range(numTrainDocs):
        if trainCategory[i] == 1:
            p1Num += trainMatrix[i]
            p1Denom += sum(trainMatrix[i])
        else:
            p0Num += trainMatrix[i]
            p0Denom += sum(trainMatrix[i])

    p1Vect = log(p1Num / p1Denom)
    p0Vect = log(p0Num / p0Denom)
    .   return p0Vect, p1Vect, pAbusive
```

分类器

分类器

```
def classifyNB(vec2Classify, p0Vec, p1Vec, pClass1):

    p1=sum(vec2Classify*p1Vec)+log(pClass1)
    p0=sum(vec2Classify*p0Vec)+log(1.0-pClass1)
    if p1 > p0:
        return 1
    else:
        return 0
```

因为:

$$p(c_i|\mathbf{w}) = \frac{p(\mathbf{w}|c_i)p(c_i)}{p(\mathbf{w})}, w: word\ vector; c_i: label$$

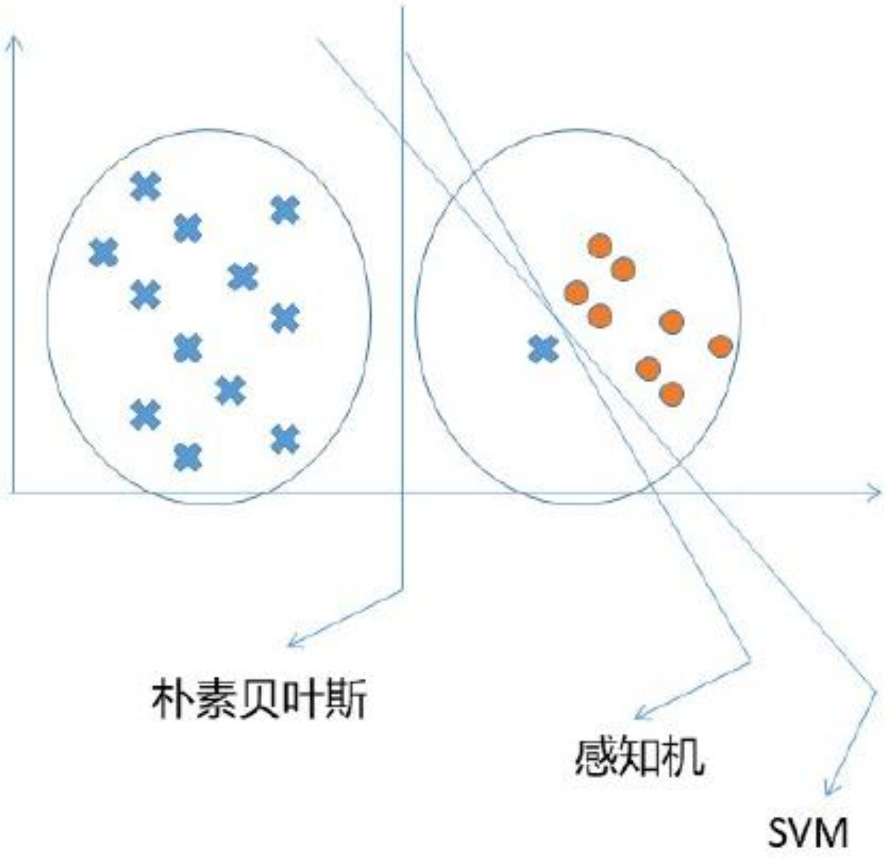
$$p(\mathbf{w}|c_i) = p(w_0, w_1, ..., w_N|c_i) = p(w_0|c_i)p(w_1|c_i)...p(w_N|c_i)$$

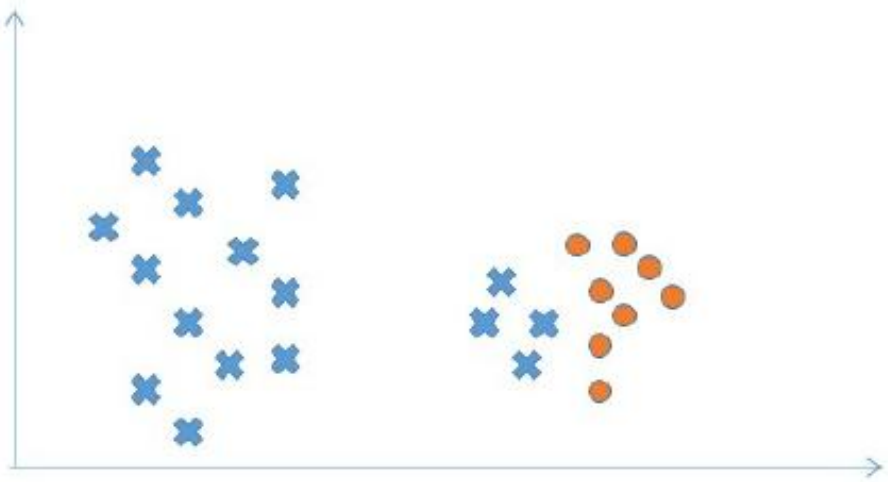
$$\begin{aligned} & log(p(\mathbf{w}|c_i)p(c_i)) \\ &= log(p(w_0|c_i)p(w_1|c_i)...p(w_N|c_i)p(c_i)) \\ &= log(p(w_0|c_i)) + log(p(w_1|c_i)) + ... + log(p(w_N|c_i)) + log(p(c_i)) \end{aligned}$$

朴素贝叶斯与感知机、SVM 之间的区别

思考题

能否根据朴素贝叶斯与感知机、SVM 之间的算法区别，画出右图的决策边界？





总结

贝叶斯公式:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

假设:

$$p(x_1, x_2, ..., x_n|y) = p(x_1|y)p(x_2|y)...p(x_n|y)$$

似然函数:

$$\prod_{i=1}^{n} p(x_i|y, \theta)$$

对数似然函数:

$$\sum_{i=1}^{n} log(p(x_i|y, \theta))$$

最大似然估计:

$$\underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} log(p(x_i|y, \theta))$$

分类器:

$$\underset{y}{\operatorname{argmax}}\, p(y) \prod_{i=1}^{n} p(x_i|y, \theta) = \underset{y}{\operatorname{argmax}}\, log(p(y)) + \sum_{i=1}^{n} log(p(x_i|y, \theta))$$