贝叶斯决策论

翟婷婷

扬州大学
信息工程（人工智能）学院
zhtt@yzu.edu.cn

2023年春

课程目标

- 理解逆概率推理的过程，牢记贝叶斯公式。
- 掌握各种贝叶斯分类规则及其等价变换。
- 理解贝叶斯分类器是最小错误率的分类器。

推理的类型

- 推理可以分为正向推理和逆向推理。
- 正向推理：由事实推出该事实会导致的结果的过程。(原因推结果)

  例子：因为今天下雨，所以绝大部分人会打伞。

  因为今年的雨水少，所以粮食产量一定会减少。

- 逆向推理：由观察到的结果反推造成该结果的原因的过程。(结果反推原因)

  例子：这只小鸟飞不起来，它的翅膀一定受伤了。

  小明持续发烧且咳嗽，他一定感冒了。

推理的类型

- 推理可以分为确定性推理和概率推理。
- 确定性推理：推理过程如下：如果事件A发生了，则事件B一定会发生。

  例子：如果考试作弊，则该科成绩一定为零。

  如果小明故意杀人，则小明一定会坐牢。
- 概率推理：也称不确定推理，其推理过程如下：如果事件A发生了，则事件B以某一概率发生。
- 已知事件A发生后，事件B发生的概率记为$P(B|A)$，则$P(B|A)$为事件B发生的条件概率。

  例子：如果考前未复习，则该科成绩有 50%的可能性不及格。

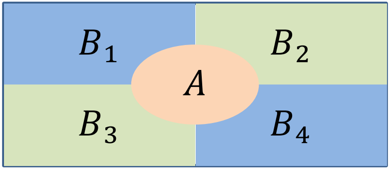
贝叶斯公式

- 贝叶斯决策是一种逆向的概率推理，其理论基础是贝叶斯公式。
- 贝叶斯公式是由托马斯·贝叶斯于 1763 年提出的。
- 贝叶斯公式：

  假设一个试验的样本空间为S，记$B_1, B_2 \cdots B_c$为S的一个划分，A为试验的事件，且$P(A) \neq 0$，则

$$P(B_i|A) = \frac{P(A,B_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{c} P(B_j)P(A|B_j)}$$

S



$$P(A) = \sum_i P(A \cap B_i)$$

贝叶斯公式

- 把A的发生作为结果，把$B_1, B_2 \cdots B_c$的发生看成导致A发生的c个可能的原因，则$P(B_i|A)$表示结果A的发生是由某个原因B_i造成的概率。
- 贝叶斯公式实现了逆概率推理的过程。
- 在贝叶斯公式，每个概率都有约定俗成的名称：

  $P(B_i)$称为先验概率：B_i发生的概率(边缘概率)，该概率与A是否发生无关。

  $P(A|B_i)$称为条件概率：B_i发生的情况下，A发生的概率。

  $P(B_i|A)$称为后验概率：A发生情况下，B_i发生的概率，该概率是根据先验概率和条件概率计算后得到。

贝叶斯公式

- 对于一个包含c个类别$\{\omega_1,...,\omega_c\}$的分类问题：
  - ✓ 记$P(\omega_i|\boldsymbol{x})$表示观察特征向量取值为$\boldsymbol{x}$时，$\boldsymbol{x}$属于$\omega_i$类的概率，称为后验概率；
  - ✓ 若特征向量$\boldsymbol{x}$的每维取值都是**连续的**,则贝叶斯公式为：

$$P(\omega_i|\boldsymbol{x})=\frac{P(\omega_i)p(\boldsymbol{x}|\omega_i)}{p(\boldsymbol{x})}=\frac{P(\omega_i)p(\boldsymbol{x}|\omega_i)}{\sum_{j=1}^{c}P(\omega_j)p(\boldsymbol{x}|\omega_j)}$$

$p(\boldsymbol{x})$为特征向量取值为$\boldsymbol{x}$时的概率密度; $P(\omega_i)$为ω_i类实例出现概率, 称为先验概率；$p(\boldsymbol{x}|\omega_i)$为$\omega_i$类中特征向量取值为$\boldsymbol{x}$的概率密度，称为类条件概率密度。

  - ✓ 若特征向量$\boldsymbol{x}$的每维取值都是**离散的**,则贝叶斯公式为：

$$P(\omega_i|\boldsymbol{x})=\frac{P(\omega_i)P(\boldsymbol{x}|\omega_i)}{P(\boldsymbol{x})}=\frac{P(\omega_i)P(\boldsymbol{x}|\omega_i)}{\sum_{j=1}^{c}P(\omega_j)P(\boldsymbol{x}|\omega_j)}$$

贝叶斯分类的原理

- 贝叶斯分类规则：

把$\boldsymbol{x}$分为ω_{i^*}类，其中$i^* = \underset{i\in\{1,\ldots c\}}{\operatorname{argmax}} P(\omega_i|\boldsymbol{x})$

即是，将$\boldsymbol{x}$分到后验概率最大的那个类别中。

等价于：

把$\boldsymbol{x}$分为ω_{i^*}类，其中$i^* = \underset{i\in\{1,\ldots c\}}{\operatorname{argmax}} P(\omega_i) * p(\boldsymbol{x}|\omega_i)$

即是，将$\boldsymbol{x}$分到先验概率乘以类条件概率密度最大的那个类别中。

可以说，先验概率是分类的基础，后验概率是在获得更多信息后，对先验概率进行修正后得到的。

贝叶斯分类例子

- 问题：假如一个社会中男女比例相等。在某条路上随机遇到一个人，请问这个人是男性还是女性？

- 先验信息：$P(\text{男性}) = 0.5$，$P(\text{女性}) = 0.5$
- 在只有先验信息的情况，这个人是男性和女性的概率分别是0.5。
- 增加一点信息：
  根据遇到的这个人的背影，得知其是留长头发的，并且男性留长发的概率$P(\text{长发}|\text{男性}) = 0.05$，女性留长发的概率为$P(\text{长发}|\text{女性}) = 0.7$，请问这个人是男性还是女性？

贝叶斯分类例子

- 根据贝叶斯公式:

$$P(男性|长发) = \frac{P(男性)*P(长发|男性)}{P(男性)*P(长发|男性)+P(女性)*P(长发|女性)} = \frac{0.5*0.05}{0.5*0.05+0.5*0.7} = 0.0667$$

$$P(女性|长发) = \frac{P(女性)*P(长发|女性)}{P(男性)*P(长发|男性)+P(女性)*P(长发|女性)} = \frac{0.5*0.7}{0.5*0.05+0.5*0.7} = 0.9333$$

- $P(男性|长发) < P(女性|长发)$，判别此人是女性。

贝叶斯分类例子

- 问题：已知某理工科大学中男生与女生的比例为3:1，则在该理工科大学校园内上随机遇到一个人，请问这个人是男生还是女生？

- 先验信息：$P(\text{男生}) = 0.75$，$P(\text{女生}) = 0.25$
- 在只有先验信息的情况下，这个人是男生和女生的概率分别是0.75和0.25。因此，判别该人是男生。
- 增加的信息与前一个例子中的相同：
  根据遇到的这个人的背影，得知其是留长头发的，并且男性留长发的概率$P(\text{长发}|\text{男性}) = 0.05$，女性留长发的概率为$P(\text{长发}|\text{女性}) = 0.7$，请问这个人是男性还是女性？

贝叶斯分类例子

- 根据贝叶斯公式:

$$P(男性|长发) = \frac{P(男性) * P(长发|男性)}{P(男性) * P(长发|男性) + P(女性) * P(长发|女性)} = \frac{0.75 * 0.05}{0.75 * 0.05 + 0.25 * 0.7} = 0.1765$$

$$P(女性|长发) = \frac{P(女性) * P(长发|女性)}{P(男性) * P(长发|男性) + P(女性) * P(长发|女性)} = \frac{0.25 * 0.7}{0.75 * 0.05 + 0.25 * 0.7} = 0.8235$$

- $P(男性|长发) < P(女性|长发)$，判别此人是女性。

贝叶斯分类的特点

① 要求先验概率已知：

- ✓ 频率派：先验概率可以由大量的重复实验中所获得的各类样本出现的频率来近似，其基础是“大数定律”。
- ✓ 贝叶斯派：许多事件的发生不具有可重复性，将概率解释为信念度，因此先验概率只能根据对置信度的主观判定来给出。

② 通过获得的新信息对先验概率进行修正：

- ✓ 没有获得任何信息的时候，要进行分类只能依据各类的先验概率，将样本划分到先验概率最大的一类中。
- ✓ 在获得了更多关于样本特征的信息后，可以依照贝叶斯公式对先验概率进行修正，得到后验概率，提高分类决策的准确性和置信度。

贝叶斯分类的错误率

- 贝叶斯分类规则：
  把$\boldsymbol{x}$分为ω_{i^*}类，其中$i^* = \underset{i\in\{1,\ldots c\}}{\operatorname{argmax}} P(\omega_i|\boldsymbol{x})$
- 贝叶斯分类的错误率是多少？

  记$P(error|\boldsymbol{x})$为观察到实例的特征向量取值为$\boldsymbol{x}$时，贝叶斯分类的错误率，则

  $$P(error|\boldsymbol{x}) = 1 - P(\omega_{i^*}|\boldsymbol{x}) = 1 - \max_{i\in\{1,\ldots c\}} P(\omega_i|\boldsymbol{x})$$

  记$P(error)$为贝叶斯分类的总错误率，则

  $$P(error) = \int_{\mathbb{R}^d} p(\boldsymbol{x})P(error|\boldsymbol{x})d\boldsymbol{x}$$

- 贝叶斯分类通过最小化$P(error|\boldsymbol{x})$来最小化总体错误率!

考虑不同的错分类代价

- 贝叶斯分类是最小化总体错误率的分类，但贝叶斯分类没有考虑错分类的不同后果/代价。
- 什么是错分类的代价?

  将普通感冒患者误诊为新冠肺炎患者，

  将新冠肺炎患者误诊断为普通感冒患者，

  这两种错分类的代价显然不一样，第二种错分类的代价远远大于第一种错分类的代价。
- 为了把分类决策所带来的后果考虑进去，因此需要最小风险的贝叶斯分类。

最小化风险的贝叶斯分类

- 决策α_i：表示将$\boldsymbol{x}$分为ω_i类。
- 损失$\lambda(\alpha_i|\omega_j)$：表示真实类别状态为$\omega_j$时，采取决策$\alpha_i$所导致的损失。
- 条件风险$R(\alpha_i|\boldsymbol{x})$：表示观察到实例的特征向量取值为$\boldsymbol{x}$后，采取决策$\alpha_i$所导致的期望损失，则

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{c} \underline{P(\omega_j|\boldsymbol{x})} \lambda(\alpha_i|\omega_j)$$

- 最小化风险的贝叶斯分类规则：

采用决策α_{i^*}，即将$\boldsymbol{x}$分为ω_{i^*}类，其中

$$i^* = \underset{i \in \{1,\ldots c\}}{\operatorname{argmin}} R(\alpha_i|\boldsymbol{x})$$

举例

- 假设某一年中，普通感冒的发病率为80%，而新冠肺炎的发病率为0.00001%，在得普通感冒的人中80%的人会出现发热、咳嗽等呼吸道感染的症状，而患新型冠状肺炎的人中99%的人也会出现相同的症状。假设只有普通感冒和冠状肺炎会导致这种发热、咳嗽等的呼吸道感染的症状。

- 现在有一个患者出现了发热、咳嗽等呼吸道感染的症状，请问是否应当将该患者按照新冠肺炎的疑似病例对待?

最小错误率贝叶斯分类举例

- $P(感冒) = 0.8$，$P(肺炎) = 0.0000001$
- $P(症状|感冒) = 0.8, P(症状|肺炎) = 0.99$

$$P(感冒|症状) = \frac{P(感冒)P(症状|感冒)}{P(感冒)P(症状|感冒) + P(肺炎)P(症状|肺炎)} = 0.999999845312524$$

$P(肺炎|症状) = 0.000000154687476$

$P(感冒|症状) > P(肺炎|症状)$，根据最小错误率的贝叶斯分类规则，应该将该患者按照普通感冒来对待，这样导致的错误率最低。

最小风险的贝叶斯分类举例

- 假设：将患新冠肺炎误判为患普通感冒所导致的损失是将普通感冒误判为患新冠肺炎所导致损失的10,000,000倍，此时有患者出现了发热、咳嗽等症状，请问该如何诊断？
- $\lambda(\text{肺炎}|\text{感冒}) = 1$，$\lambda(\text{感冒}|\text{肺炎}) = 10{,}000{,}000$

  $R(\text{肺炎}|\text{症状})$

  $= P(\text{肺炎}|\text{症状}) * \lambda(\text{肺炎}|\text{肺炎}) + P(\text{感冒}|\text{症状})$

  $* \lambda(\text{肺炎}|\text{感冒}) = 0.999999845312524$

  $R(\text{感冒}|\text{症状})$

  $= P(\text{肺炎}|\text{症状}) * \lambda(\text{感冒}|\text{肺炎}) + P(\text{感冒}|\text{症状})$

  $* \lambda(\text{感冒}|\text{感冒}) = 1.54687476$
- 因为$R(\text{肺炎}|\text{症状}) < R(\text{感冒}|\text{症状})$，根据最小风险的贝叶斯分类，应该将患者诊断为患新冠肺炎。

最小化错误率与最小化风险的关系

- 定义一个损失函数（0-1损失）：

$$\lambda(\alpha_i|\omega_j)=\begin{cases}0, & 如果 i=j\\ 1, & 如果 i\neq j\end{cases} \quad i,j\in\{1,\ 2,\ldots c\}$$

- 带入条件风险计算公式:

$$R(\alpha_i|\boldsymbol{x})=\sum_{j=1}^{c}P(\omega_j|\boldsymbol{x})\lambda(\alpha_i|\omega_j)$$

$$=\sum_{j,\,j\neq i}P(\omega_j|\boldsymbol{x})=\underline{1-P(\omega_i|\boldsymbol{x})}$$

$\boldsymbol{x}$被分为ω_i类的错误率

- 当采用0-1损失时，最小化风险等价于最小化错误率:

$$\operatorname*{argmin}_i R(\alpha_i|\boldsymbol{x})=\operatorname*{argmax}_i P(\omega_i|\boldsymbol{x})$$

一般分类器的错误率：二分类

- 对于一个二分类问题，包含类别$\{\omega_1, \omega_2\}$。
- 考虑任意一个二分类器，假设该分类器可能按照非最优的方式将特征空间划分为两个决策域：$\mathcal{R}_1$和$\mathcal{R}_2$，且采用的分类规则是：若$\boldsymbol{x} \in \mathcal{R}_1$，则判$\boldsymbol{x} \in \omega_1$；否则判$\boldsymbol{x} \in \omega_2$。
- 考虑该分类器的错误的概率$P(error)$如何表示？

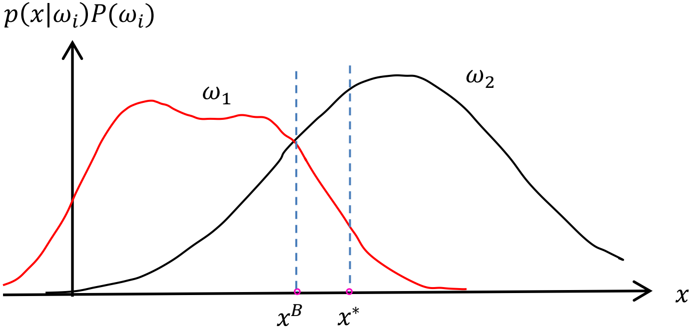
错分类的情况只有两种：

$\boldsymbol{x}$落在$\mathcal{R}_1$中，但$\boldsymbol{x}$的真实类别为ω_2； $\Rightarrow$ $P(\boldsymbol{x} \in \mathcal{R}_1, \omega_2)$

$\boldsymbol{x}$落在$\mathcal{R}_2$中，但$\boldsymbol{x}$的真实类别为ω_1； $\Rightarrow$ $P(\boldsymbol{x} \in \mathcal{R}_2, \omega_1)$

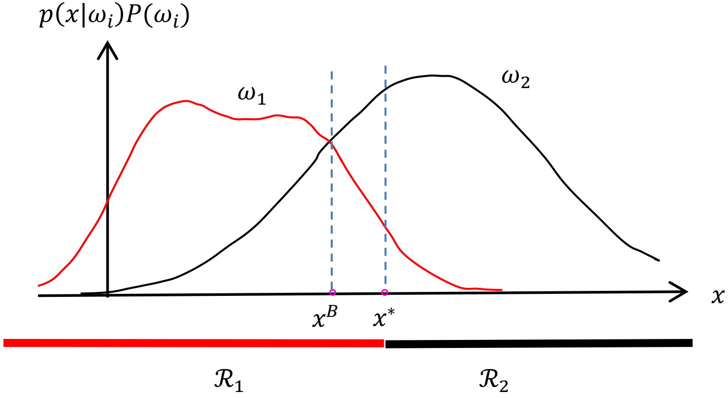
$$
\begin{aligned}
P(error) &= P(\boldsymbol{x} \in \mathcal{R}_2, \omega_1) + P(\boldsymbol{x} \in \mathcal{R}_1, \omega_2) \\
&= P(\boldsymbol{x} \in \mathcal{R}_2 | \omega_1) P(\omega_1) + P(\boldsymbol{x} \in \mathcal{R}_1 | \omega_2)\, P(\omega_2) \\
&= \int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1) P(\omega_1) d\boldsymbol{x} + \int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2) P(\omega_2) d\boldsymbol{x}
\end{aligned}
$$

一般分类器错误率vs贝叶斯分类器错误率

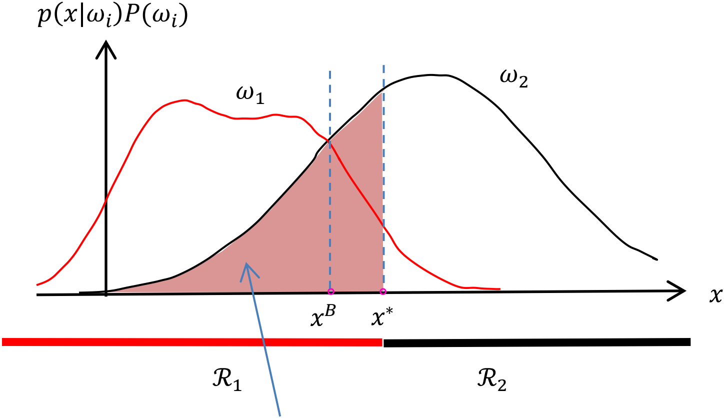


x^B是贝叶斯分类器的分类决策点，假设x^*是任意二分类器给出的决策点。

一般分类器错误率vs贝叶斯分类器错误率

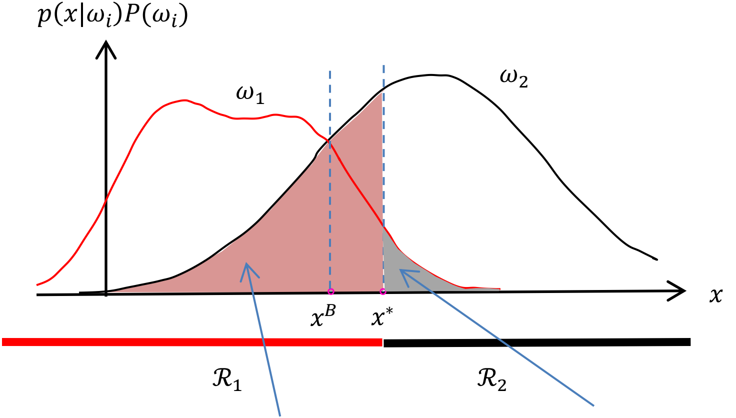


一般分类器错误率vs贝叶斯分类器错误率



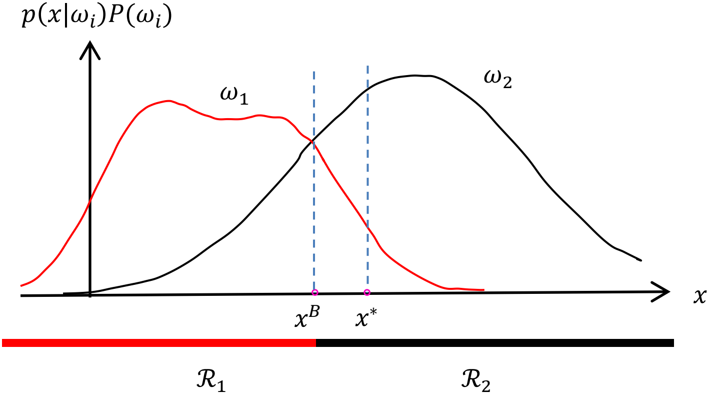
$$\int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2)P(\omega_2)d\boldsymbol{x}$$

一般分类器错误率vs贝叶斯分类器错误率

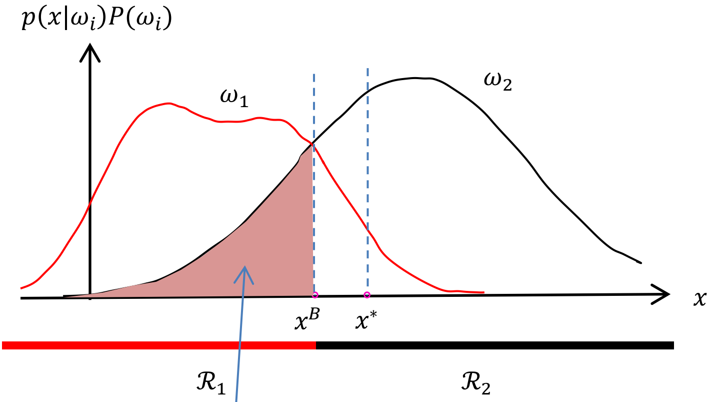


$$\int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2)P(\omega_2)d\boldsymbol{x} \qquad \int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1)P(\omega_1)d\boldsymbol{x}$$

一般分类器错误率vs贝叶斯分类器错误率

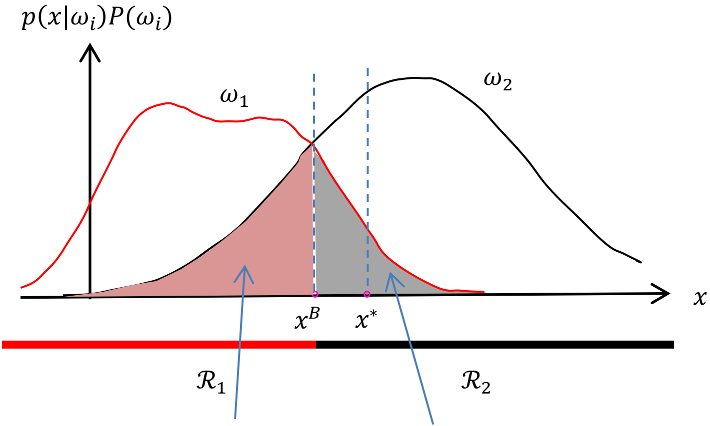


一般分类器错误率vs贝叶斯分类器错误率



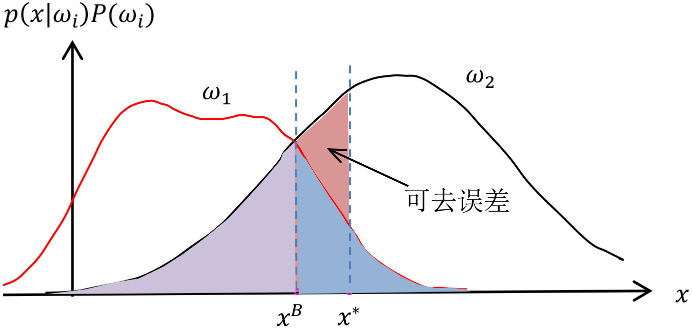
$$\int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2)P(\omega_2)d\boldsymbol{x}$$

一般分类器错误率vs贝叶斯分类器错误率



$$\int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2)P(\omega_2)d\boldsymbol{x} \qquad \int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1)P(\omega_1)d\boldsymbol{x}$$

一般分类器错误率vs贝叶斯分类器错误率



贝叶斯决策规则给出的决策点x_B能够取得最小的错误率，这个错误率是最优的错误率。

一般分类器的错误率：多分类

- 对于一个多分类问题，包含c个类别$\{\omega_1, \omega_2, \cdots \omega_c\}$。
- 考虑任意一个多分类器，假设该分类器可能按照一种非最优的方式将特征空间划分为c个决策域：$\mathcal{R}_i, i = 1, \cdots c$，且采用的分类规则是：若$\boldsymbol{x} \in \mathcal{R}_i$，则判$\boldsymbol{x} \in \omega_i$。
- 考虑该多分类器的错误的概率$P(error)$如何表示？

  由于错分类的情况太多，而正确分类的情况相对较少，先考虑正确分类的概率$P(correct)$。

  正确分类的情况只有c种：

  $\boldsymbol{x}$落在$\mathcal{R}_j$中，且$\boldsymbol{x}$的真实类别为ω_j；　$\Rightarrow \ P\left(\boldsymbol{x} \in \mathcal{R}_j, \omega_j\right)$

  $$P(error) = 1 - P(correct) = 1 - \sum_{j=1}^{c} P\left(\boldsymbol{x} \in \mathcal{R}_j, \omega_j\right)$$
  $$= 1 - \sum_{j=1}^{c} \int_{\mathcal{R}_j} p\left(\boldsymbol{x} | \omega_j\right) P(\omega_j) d\boldsymbol{x}$$

小结

- 贝叶斯推理是一种逆概率推理，其数学基础是贝叶斯公式。贝叶斯分类规则是将$\boldsymbol{x}$分到后验概率最大的那个类中：

即把$\boldsymbol{x}$分为ω_{i^*}类，其中$i^* = \operatorname*{argmax}_{i \in \{1,\dots c\}} P(\omega_i | \boldsymbol{x})$

重点

- 贝叶斯分类是最小化错误率的分类。
- 在实际问题中，不同的错分类决策会带来不同的影响和后果，最小化错误率的贝叶斯分类没有考虑这一点。为了把分类决策所带来的后果考虑进去，因此需要最小化风险的贝叶斯分类。

小结

- 最小化风险的贝叶斯分类规则：

  采用决策α_{i^*}，即将$\boldsymbol{x}$分为ω_{i^*}类，其中

$$i^* = \underset{i\in\{1,\ldots c\}}{\operatorname{argmin}} R(\alpha_i|\boldsymbol{x})$$

  条件风险计算公式：

重点

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{c} P(\omega_j|\boldsymbol{x})\lambda(\alpha_i|\omega_j)$$

- 在最小化风险的贝叶斯分类中，当误分类的损失采用0-1损失时，最小化风险等于最小化错误率。